Changing assessment practices resulting from the shift towards on-screen assessment in schools

Rose Clesham

Submitted to the University of Hertfordshire in partial fulfilment of the degree of Doctor of Education

March 2010

Abstract

This dissertation reports a study into the appropriateness of on-screen assessment materials compared to paper-based versions, and how any potential change in assessment modes might affect assessment practices in schools.

The research was centred around a controlled comparative trial of paper and on-screen assessments with 1000 school students. The appropriateness of the assessments was conceptualised in terms of exploring the comparative reliability, validity and scoring equivalence of these assessments in paper and on-screen modes.

Reliability was considered using quantitative analysis: calculating the performance and internal reliability of the assessments using classical test theory, Cronbach's alpha and Rasch latent trait modelling. Equivalence was also addressed empirically. Marking reliability was not quantified, however it is discussed.

Validity was considered through qualitative analysis, using questionnaire and interview data obtained from the students and teachers participating in the trial; the focus on the comparative authenticity and fitness for purpose in assessments in different modes.

The outcomes of the research can be summarised as follows: the assessment tests in both modes scored highly in terms of internal reliability, however they were not necessarily measuring the same constructs. The scores from different modes were not equivalent, with students performing better on paper. The on-screen versions were considered to have greater validity by students and teachers.

All items in the assessments that resulted in significant differences in performance were analysed and categorised in terms of item types. Consideration is then given to whether differences in performance are the result of construct irrelevant or relevant factors.

The recommendations from this research focus on three main areas; that in order for on-screen assessments to be used in schools and utilise their considerable potential, the equivalence issue needs to be removed, the construct irrelevant factors need to be clearly identified and minimised and the construct relevant factors need to be enhanced.

Finally a model of comparative modal dependability is offered, which can be used to contrast and compare the potential benefits and issues when changing assessment modes or item types are considered.

Acknowledgements

The work in this dissertation has been carried out over a period of four years and could not have been completed without the support, advice and patience of my supervisory team, work colleagues and in particular my family.

I am extremely grateful to my principal supervisor Dr Eddie Blass for her unfailing good advice, positivity and enthusiasm. Any problems seemed less after our discussions. Thanks also go to my second and third supervisors, Professor Trevor Barker and Dr Brendon Larvor respectively for their sound advice and stimulating discussions in tutorials.

Thanks also go to my colleagues Malcolm Hayes and Zeek Sweiry who helped me to analyse vast amounts of data, and discuss the outcomes.

This research would not have been possible without the large numbers of students who participated in trials and the informative and insightful comments given by these students and their teachers. They have all made an active contribution to understanding the effects of assessment mode on reliability, validity and equivalence issues.

Finally to my husband Paul and my children Ellen and Finn: thank you for your patience and ICT skills. I would not have been able to produce this dissertation without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims of the research

This dissertation reports a study into the appropriateness of on-screen science assessment materials compared to paper-based versions, and how any potential change in assessment modes might affect assessment practices. This study is underpinned by the following premise from Samuel Messick:

> *Every time a substantial change is made to an examination program…a thorough study of the impact of the change upon the students, sub groups of students and others who may be affected by the results of the testing program must be made prior to the implementation of the change.*
>
> *Messick (1989:42)*

This may seem an obvious statement to make with respect to changing assessment systems and practices, although it is not always implemented as I will describe in this chapter. This dissertation presents one interpretation of how ‘impact’ can be evidenced.

The interpretation I am going to use to evidence impact is to explore the comparative reliability and validity of science assessments presented in paper and on-screen modes, my research questions being:

- Does the mode of science assessments result in any differences in performance by students or in test reliability?
- How do paper and on-screen science assessments compare in terms of authenticity and fitness for purpose?
- Do science assessments in differing modes assess the same constructs?

Although this research study is carried out in the context of science, my interest in the comparability issues surrounding assessments presented in different modes is not limited to this particular subject area and the research outcomes will inform general subject as much as specific science issues.

The concepts of reliability, validity and constructs will be discussed in Chapter 2 and my research questions will be expanded on in terms of methodology and method in Chapters 5 and 6 respectively.

1.2 Location of the research

Before I proceed to describe how my research has been conceptualised and operationalised, I need to locate my research with reference to four interlinking themes; my professional and reflexive position, ideas about appropriate research methodology, ideas about the outcomes of different research approaches and finally ideas about changing assessment principles and practices.

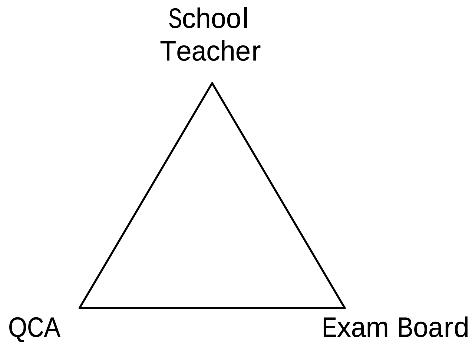
1.2.1 My professional and reflexive position

Reflexivity refers to the process by which our observations are dependent upon our prior experiences and the way in which these experiences inform our opinions or judgements (Blatchford & Blatchford, 1997). As such, there are strongly held views by many researchers that the role of reflexivity is a central concern to educational researchers and their work in terms of the questions they ask, their research approaches and philosophies and the manner in which their research outcomes are reported (see Blatchford & Blatchford, 1997; Halliday, 2002; Griffiths, 1997).

I agree with the position that we cannot separate our actions from the reflexive experiences that informed them, and my research questions have been informed by various aspects of my professional experience and practices, which I will go on to describe. I do not take the view however that these experiences should or do inform research outcomes. This is more dependent on the form and nature of the research, and its intended purpose.

The reasons I am interested in reliability and validity issues associated with assessment are idiosyncratic in part and the result of a background in a variety of educational sectors.

Figure 1: Three professional educational perpectives



Many years of my professional career were spent teaching and managing science departments in schools. In terms of curriculum and assessment my aims were that the students were given access to an engaging and stimulating science curriculum and that assessment reflected these aims. Although formative assessment was on-going and an integral part of the teaching and learning, it was summative assessment that presented the most challenges for many students. As a science educator, I took the view that the high stakes nature of assessment in England often resulted in a narrow interpretation of the science curriculum and that many students were disadvantaged by a highly codified and structured approach to the form, context and content of summative examinations.

While I assumed a level of marker and test reliability which may or may not have been evident, my concerns as a teacher were more related to validity, accessibility and fairness issues. Those were the aspects that influenced the curriculum we provided and the chances my students had to achieve measurable successes. My views then on validity were not expressed in the unified manner I describe in Chapter 2, however they reflected the concerns of the educational environment that I worked in.

My primary concerns working within Test Development (and in particular national curriculum tests) at The Qualifications and Curriculum Authority (QCA) were different. In this environment, there was a national governmental agenda about the maintenance of educational standards. To this end, there was far more emphasis on aspects of assessment reliability than validity. In Chapter 2, I describe how assessment systems often operate on variations of Goodhart's Law; that they emphasise the elements that can be quantified and measured, sometimes at the expense of important, unquantifiable and therefore less valued aspects. At QCA, my concerns about individual students were replaced by concerns about national standards and test quality assurance measures, no more or less worth than each other; however quite different in emphasis.

Working within the environment of an exam board has added yet another professional perspective on educational assessment. In some respects exam boards are the conduit between governmental regulation and the standards agenda and the provision of flexibility for schools in the form, context and content of appropriate assessments for their cohorts. Exam boards need to adhere to regulatory frameworks, and are accordingly regularly scrutinised and monitored. However, in post 14 qualifications, schools have the choice between three exam boards in England, and therefore each exam board tries to be as pro-active as possible in terms of positioning themselves as providers of reliable, valid, accessible and fair assessments.

In Chapter 2, I describe the concept of assessment dependability, the intersection between reliability and validity (Gipps, 1994). This essentially describes the relationship between schools and exam boards. Schools select, as much as they can, the most dependable assessments for their purposes.

As I have described, my interests and experience of assessment reliability and validity have been experienced through differing lenses of educational providers and users, and they have all informed my educational values and research aims.

Rokeach (1973) categorized four types of values that influence our conduct; these being moral, competency, personal and social. Simple definitions of these values would describe *moral* as being what is the ‘right’ thing to do, *competency* as the most effective way to go about doing something, *personal* as what an individual hopes to achieve for themselves and *social,* how an individual would wish society to operate.

Our attitude and behaviour in given situations are determined by the interplay of these four values (see Hammersley & Gomm, 1999; Harrison, 1999), however problems can arise in situations where values are in conflict with each other, and the individual has to select one value over another (Glen, 2000).

It is clear therefore, that my research cannot be value-neutral, as I have had personal interest and experience of educational assessment from various sectors and viewpoints, however I would hope that my research aims, in terms of interest in both reliability and validity issues encourage a convergence rather than a conflict of values.

1.2.2 Appropriate research methodology

The second issue to consider is the nature of my research questions, approaches and methodology. I have read with interest some polar positions and arguments from positivist and interpretivist research camps, the former believing in the power of empiricism, the latter in narrative and interpretation.

Hargreaves (1996) brought this dispute into sharp focus in his speech to the Teacher Training Agency, where he questioned the quality and value of much educational research.

Educational researchers, like other social scientists, are often engaged in bitter disputes among themselves about the philosophy and methodology of the social sciences (p2).

It is this gap between researchers and practitioners which betrays the fatal flaw in educational research (p3).

He used medical research as a useful analogy for the movement that educational research should take, where there is little distance between the medical research community and the practitioners (doctors). Evidence based medicine, Hargreaves argued, had more direct relevance to doctors and their patients, and it was in all of the stakeholders interests to keep up to date with developments. He argued that most educational research should be associated with addressing tangible rather than esoteric needs and that action research, in terms of researchers working alongside teachers in schools should be the favoured research approach.

Blake (1997) interpreted this 'relevance gap' as a reaction against perceived failures of educational research to improve practice over decades and in some cases, even making situations worse. In the late 1990's, there was political endorsement of Hargreaves' views by leading figures such as Michael Barber (1996) and Chris Woodhead (1998), which has continued ever since, to the effect that educational research had become 'laissez-faire' (Homan, 1990), irrelevant and a distraction (Woodhead, 1998), 'sloppy' (Tooley & Darby, 1998) and 'Ivory towered' (Blunkett, 2000).

Many researchers took these statements to be the death knell of critical, exploratory, intellectual research, in favour of a state controlled agenda of empirically driven research to inform evidence-based policy making and practice (Hodkinson,2004); even 30 years ago there was a fear that researchers would become ‘technicians’ rather than ‘intellectuals’ (Morris, 1972).

Hammerersly (2005) on the other hand has been a consistent defender of the development of knowledge through educational enquiry, his view conceptualising knowledge accumulation, not as a simple continuous wall-building process, but utilising more complex conceptions and gestalts; often resulting in discontinuous shifts in understanding, in the manner described by Kuhn (1970).

Although Hargreaves and Hammersley express different emphases, many of Hargreaves pragmatic views resonate with respect to my research aims, as does Hammersley’s emphasis on thoughtful consideration. I left the field of the classroom and teaching ten years ago, but still consider myself a practitioner in the field of education. As such, I consider my research to be similar to the action research carried out by a school or classroom based practitioner. The scale of my research is larger and by its nature has to be generalisable in its outcomes in terms of large scale assessments; however the purpose is to address a given educational problem: how to evidence the comparison of reliability and validity between old and new forms of assessment.

Blake (1997) suggested that the ‘relevance gap’ between educational research and practice could be in part addressed by closer collaborative relationships between researchers and practitioners. He was ostensibly referring to universities and schools, and used the phrase ‘local universities’ to describe this partnership. Although the scale of collaboration would be different, I would advocate that the same symbiotic relationship should be fostered and developed between universities and exam boards. This does happen to a small extent, particularly with regard to the curriculum and pedagogic developments. However, given the significance and impact assessment has in education and on society in general, it would be mutually beneficial for these two institutions to be better informed about each others work, utilise their complementary experiences, expertise and skills and develop robust research and analytical approaches to address assessment related issues.

Apart from discussing the purpose of my research question and approaches, I should make it clear that my research needs to be teleological. My research outcomes have to be fit for purpose in terms of providing appropriate evidence to influence the policy and practice of assessment and therefore make a significant contribution to the practice of education, as described in the aims and requirements of this doctoral programme.

The assessments developed and used in my research will most likely find a place in high stakes assessments. ‘High stakes’ in this context refers to assessments that are used for selection or entry requirements for students (eg. GCSE and GCE), or those used for accountability purposes for schools (eg. GCE, GCSE and National Curriculum tests) in England. As such, these assessments are governmentally regulated in terms of their development and outcomes, and are dominated by statistical modelling techniques. Any potential changes to the status quo (for example moves to replace the mode and nature of assessments from paper to computer) will need to satisfy regulatory requirements in terms of establishing comparability and equivalence and ‘maintaining standards’.

My teleological stance might seem to indicate that the research outcomes only need to be empirically based, and there is no doubt, as I have described, that statistical evidence is key to influencing policy decisions by regulators and exam boards. However, there are research and pragmatic reasons why I am also interested in incorporating interpretive approaches and analysis to my research, in particular to address my research questions focusing on validity rather than reliability aspects.

Hodkinson (2004) advocated the importance of the quality of the interpretation, rather than relying on the objective purity of data, and although analysis based on quantitative data seems to provide more clear-cut solutions than qualitative analysis, and is therefore often held in higher esteem in policy decisions, few people would want to make significant decisions solely on statistics. Even David Blunkett (2000) conceded that qualitative methods may be a useful adjunct to quantitative methods. The key principle from my point of view is to ensure that the qualitative approaches and analysis are as robust and rigorous as the quantitative approaches. This view is supported by Hammersley (2005) and Nash (2005) who argue for more high quality, systematic approaches to qualitative research.

Notwithstanding the quality of the associated strands of quantitative and qualitative approaches to my research, there are pragmatic reasons why a qualitative strand to my research is appropriate and necessary. Not only are qualitative approaches more appropriate to address my research questions concerning assessment validity, but also there is far more emphasis in gathering public support and confidence in assessment related initiatives than there used to be, (although there are exceptions, which I will discuss later). Public consultations on initiatives from governmental agencies or exam boards have some currency in policy decisions. Unfortunately, these consultations usually have poor response rates, however there is at least some acknowledgement that opinions can matter to informing policy decisions.

In the case of exam boards, any significant changes to assessment systems in the first instance have a very simple barometer of success; schools and teachers will either opt in, if they feel that either logistically or for assessment related reasons, they have something to gain from a change or else they will vote with their feet and go elsewhere for assessments that suit their requirements. As such, significant financial and systematic investment by government or exam boards must be matched by the potential interest, and take up of the end users. Even if I could prove that my research assessments had far greater validity and reliability in on-screen rather than paper-based form, unless schools actively support them, change will be slow at best. Therefore interpretive evidence from teachers and students may be more significant than might be assumed.

1.2.3 Research inferences

My third issue concerns the choice of my research questions and approaches and ideas about the different language used and the inferences drawn from their outcomes. Positivist and interpretive research approaches are often presented using different forms of language to describe and account for their processes; positivists using scientific terms such as robustness and validity, and interpretivists preferring terms such truthfulness and values. Although there is a difference in the language and tone used in these approaches, when used appropriately, they all contribute to the critical analysis of evidence.

My research has a strong emphasis in empiricism, allied to a complementary qualitative strand, and therefore adopts a mixed methodological approach.

Sometimes the choice of methodology is used as a means to a pre-determined end, be it political or social. Foucault (1977-78) exposed this using the term ‘political arithmetic’ to describe the way governments can use empiricism for their own devices. He raised questions about the role of researchers and research approaches and gave numerous examples of how extreme views of different research approaches can be used for particular purposes.

The following quotations from Foucault give contrasting examples of these extreme viewpoints.

> *I think that the modern age of the history of truth began at the moment when empirical knowledge itself, and on its own, allowed access to the truth. That is, from the moment when without asking anything of the subject, without the being of the subject having to undergo any modification or alteration whatsoever, the philosopher (or scientist or anyone looking for the truth) was capable of recognising in him or herself the truth and had access to the truth by the mere act of empirical knowledge (Foucault, 1981:19).*
>
> *It is hard to see what kind of objectivity is achieved by the statistical analysis of a questionnaire examining the lies of school age children and their playmates. At the end of the day, the results are re-assuring, we learn that children lie mostly to avoid punishment, but also to boast of their exploits etc. We can be sure by virtue of these findings, that the method was quite objective. So what? These are those obsessive peeping toms who, in order to look through a plate glass door, peer through the keyhole (Foucault, 1957:58).*

In Chapter 2, I describe how assessment empiricism does not always bear too close a scrutiny in terms of its own supposed reliability and validity (eg. Newton, 2005; Wiliam,1995). Lather (2004) exposed similar issues associated with testing and accountability programmes in the US that have sprung up in the light of the No Child Left Behind Act in 2001. 'Empiricism' and 'positivism' are terms that are used in a variety of ways and not necessarily applied in common forms (Hammersley, 1995). Furthermore, a belief in the importance of objectivity as a guiding principle in research approaches does not necessarily guarantee the validity of the outcomes (Hammersley, 2004). 'Objectivity' should not be confused with a stereotyped notion of empiricism with no story to tell. It should just describe a set of defendable research principles and approaches, where research approaches are selected to be valid for their intended purposes

In contrast to the language of empiricism, interpretive enquiry prefers to use truthfulness as a functional equivalent to the positivist term of validity (Reason & Rowan, 1981).

Bridges (1999: 597) argues that educational research has to be 'concerned in some sense with the truth in relation to the matter which is the focus of its enquiry'. He also suggests that where truth is not a goal, or criterion of enquiry, educational research 'probably collapses into incoherence'. However, a focus on truth is easier said than done. Walker (1985) suggests that interpretive enquiry tells *a* truth, but not necessarily *the* truth. This will be the case with my mixed research approaches. They will be my interpretations of interpretive and empirical evidence.

There is probably a greater opportunity and need to adopt reflexive approaches when utilising interpretivist research approaches. This requires the honesty to acknowledge one's own experiences and values and how they might influence either the research approaches or the interpretation of the acquired evidence. Likewise, the experiences and values of the participants also need interpretation, and are much harder to take into account. Part of the skill of interpretive enquiry is of course, the way the researcher-participant relationship is managed. At the same time, although empirically based, the analysis of some of quantitative data will also be influenced by my previous professional experiences, and will also therefore have some reflexive input.

In combination, my research questions can be best addressed using different research approaches, which although using different forms of evidence, language and tone, add to the overall validity of my research.

1.2.4 Changing assessment principles and practices

The fourth and final theme that I want to explore here are ideas about changing assessment principles and practices. This is clearly at the heart of my research interest. I return here to the quote given at the start of this chapter:

Every time a substantial change is made to an examination program…a thorough study of the impact of the change upon the students, sub groups of students and others who may be affected by the results of the testing program must be made prior to the implementation of the change. (Messick, 1989: 42).

These sentiments underpin my interest and research in comparing computer and paper-based assessments. There are a number of issues that are worth exploring a little at this stage, about the nature of change in relation to the field of education, and my role as an agent of change.

When Ken Boston (2005) Chief Executive of the Qualifications and Curriculum Authority (QCA) announced that he expected on-screen assessments for all new qualifications by 2009, and that e-assessment should be a routine provision in this country by that time, he and QCA had a number of motivations. It was clear then, and has continued to be the case ever since, that there is the need to incorporate 21st century technology into education generally and assessment specifically. At one end of the spectrum is the need to modernise the logistics of examination management in this country (Ripley, 2004). The number of externally marked tests and exams have increased exponentially over the last five years, and the 'cottage industry' of scripts being sent through the post to the doorsteps of examiners is at the least insecure and at most, costly, and time and people intensive.

This motivation should not be seen as the sole driving force for assessment changes, however one should never underestimate the quality, robustness and efficiency of assessment operational systems in the perceived success or failure of national assessments. This can be exemplified by the numerous 'debacles' of key stage test delivery failures and the continued reporting of missing or stolen exam scripts. Boston (2007) extended the argument towards the movement towards on-screen systems, offering more assessment related reasons like personalised learning, assessment on demand, rapid feedback of results and the need to explore issues of assessment reliability and validity related to on-screen assessments.

My interest has far more to do with the assessment related arguments than those of efficiency, although it might be argued that the latter reason initialises the movement towards computer-based assessment.

The aspirations and expectations expressed in 2005 were ambitious in the extreme, and unsurprisingly, as we have now reached 2010, have not been met. This is in part due to the lack of IT infrastructure in schools, but also largely due to comparability and equivalence issues. These terms are described and discussed in more detail in Chapter 3. There has been renewed governmental interest in computer based testing entering into 2010, with Kathleen Tattersall, (2009) chairwoman of Ofqual announcing that on-screen tests would be a fairer way to test students who have grown up in a computer and digital age, and they should be universally available in all high stakes assessments in England over the next ten years.

However, as much as the government, Ofqual and QCDA want efficiency and assessment related enhancements, movements to computer based testing would initially not involve specification and curricular change, and therefore for accessibility reasons, for a number of years, schools and students would have a choice concerning the mode of assessment, ie. paper or computer-based. This would limit any significant changes to the assessments.

As the only established equivalence between paper and computer based tests are found in basic multiple choice formats, and most, if not all large scale, high stakes assessments in England are **not** predominantly in this form, the lack of research in comparability and equivalence of assessments in different modes of styles used in England has resulted in assessment stasis, where perceived problems of maintaining 'standards' have presented a significant hurdle in terms of incorporating new assessment technologies (Wheadon and Adams, 2007).

It is interesting how this governmental caution and sensitivity is selective. Major curriculum reforms have regularly been implemented in England with no research, piloting or trialling stages, whereby entire cohorts have effectively been used as research guinea pigs. Oates (2007) gives examples of such developments, including the implementation of the National Curriculum, Key Skills, GNVQ's and Curriculum 2000, raising significant ethical issues that have largely been ignored by policy makers. The Diploma could now be added to this list.

DES officials throughout the 1990's repeatedly stated that to test out a full curriculum offer with a selected group of pupils would constitute tampering with their futures. '…you can't experiment with things which, if things don't go as you plan, might compromise their future lives…' The result was to experiment with every pupil.

Those who question the ethics of educational trialling and piloting in non-live assessment situations should consider the large scale harm caused as a consequence.

Computer-based assessment has not been allowed to go down this route as assessment methodology at this time is considered to be reflection of the curriculum, and not a major educational change in its own right which would allow the 'standards' clock to be reset and restarted from scratch.

In comparison with whole-scale untried and tested curricular reform, computer based assessment caution and inertia might therefore not necessarily be a bad thing, until it is more thoroughly researched and established.

There are consequences to this inertia. As I have mentioned earlier, one of the intended aims of assessment is to measure how effectively the curriculum is delivered. However, there is an increasing mismatch between the ways computers are utilised in teaching and learning opportunities for teachers and students, and the restricted ways in which the curriculum is assessed on paper (Heppell et al, 2004; Tattersall, 2009). This raises assessment validity issues, and therefore whilst the first steps to computer-based testing might be to improve the efficiency of assessment systems and to establish comparability and equivalence of similar styles of assessment, the longer term aim must be to utilise the potential and power of computers to assess a broader range of skills, which reflect essential construct skills, but not assessable through the medium of paper-based exams. This shift would require including the resetting of the 'standards' clock.

This staged progression was outlined by Bennett (1998), and is discussed in more detail in Chapter 3; whereby initially the same construct areas can be assessed on computer as on paper, but moving to a position where computer-based assessments can assess constructs not possible on paper. This will clearly take time, considering the factors I have briefly outlined here. While this may be frustrating for some, it may be that a more measured approach to assessment change is appropriate. My role in this research and assessment movement is a responsibility not to be taken lightly, and there is clearly a reflexive input between my current and previous professional roles, and their associated interests and values. This entails the development of innovative, engaging and fit for purpose assessments, while at the same time working within the confines of appropriate validity and reliability measures. As such, I am just as interested in highlighting the pitfalls and problems of new assessment technologies, as I am in establishing equivalence and proof of concept evidence.

1.3 Summary

This chapter has set out my research aims and questions. The location of my interest has been discussed through four interlinking themes which have offered a reflexive position from which my research can now proceed. I will now move on to explore how the terms reliability and validity are applied to high stakes paper-based assessments in England.

Chapter 2

Reliability and Validity Explored

2.1 Introduction

School students in England are subjected to more mandatory assessments than any other country (Wiliam, 2001b). Most of these consist of externally set, externally marked paper-based tests and examinations. Confidence in the reliability and validity of these assessments therefore underpin public confidence in a testing and examination culture (Newton 2004). However, there are significant differences in what the terms reliability and validity mean to different stakeholders, and how they are applied in high stakes educational assessments. This chapter sets out to explore the meanings of these two terms and then discusses the issues surrounding their application. The term high stakes is often used to describe assessments that are used to categorise and select students, and sometimes grade or rank schools. They can therefore be said to carry with them substantial consequences for stakeholders (Stobart, 2008).

This chapter focuses on high stakes, school paper-based assessments in England and how their reliability and validity are measured or interpreted. This provides a baseline for the current status quo. My thesis however, is concerned with the changing assessment practices in relation to movements towards computer based assessments that will impact on students over the next ten years and therefore how interpretations of reliability and validity may also change.

The first person to be credited with establishing a mental measurement scale was Galton (1884). He hypothesised, using crude physical measurements and anecdotal evidence that there was a normally distributed attribute of 'intelligence' across a population. From that point developed the discipline of 'psychometrics', which attempt to apply statistical principles and models to the measurement of particular mental dimensions or traits.

Gardner (1992) describes how Binet's early attempts at developing intelligence tests for predictive purposes in Paris in the early twentieth century led on to the developed of IQ (intelligence quotient) and SATs (Scholastic Aptitude Tests) in the United States, which have proliferated into the national measurement technique preferred for college and university entry. Rather than focus on these supposed raw 'content independent' mental ability measurements, this chapter focuses on the validity and reliability of paper tests and examinations set across subject areas in England. The first step therefore is to define what these terms actually mean.

2.2 Validity

The most common and simple definition is probably one of the first developed; the extent to which an assessment 'measures what it purports to measure' (Garrett, 1937, p324 cited in Wiliam, 1994). Another straightforward definition is from Anastasi (1990), that validity is concerned with 'what the test measures and how well it does so'.

However, educational texts abound with the breakdown of validity into component parts, rationalising that a test may contain more than one form of validity. There are a large number of available validity types, with Brown (1980) identifying at least thirty, but the ones I will briefly describe are the key areas of Content, Construct, Predictive and Concurrent validity.

2.2.1 Content Validity

This a measure of how an assessment matches the content and learning aims of a particular syllabus or specification. The assessment should be inclusive of key relevant subject matter, sampling these areas fairly. Both the content covered and the cognitive or skill level of the test should be considered. Overall, the concern is to consider what the test appears to assess, and whether it actually does.

Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn (Messick, 1989, p16).

A test or examination is the result of selecting possible questions from a pool of available items and also has to be a sample of the specification. Therefore:

Content validity is based on professional judgements about the relevance of the test content of a particular behavioural domain of interest and about the representativeness with which item or task content covers that domain (Messick, 1989, p17).

2.2.2 Construct Validity

This is also a measure of what a test measures, but instead of looking through the lens of subject content, it is concerned with a more broad view of a latent trait or domain. This could include verbal reasoning, numeracy, or spacial awareness. It might also include part or all of a subject ability such as English, maths or science, particularly in terms of a specified school curriculum.

Construct validity is evaluated by investigating what qualities a test measures, that is, by determining the degree to which explanatory concepts or constructs account for performance on the test (Messick, 1989, p16).

Therefore construct validity may incorporate other forms of validity:

There is often no sharp distinction between test content and test construct...content-related inferences and construct-related inferences are inseparable (Messick, 1989, p36).

2.2.3 Predictive Validity

Usually this is described as a forward inference correlation between a result of a test (eg for selection) and future performance. Therefore the generalisation of a test result acts as an indicator to another outcome criteria. It could be argued that IQ tests are practical applications of predictive validity (James, 1998). It could also be suggested that many high stakes assessments in England act no more than predictive indicators of future performance.

2.2.4 Concurrent Validity

This is the validation of one test alongside another. In a non statistical method, if a test (eg. a national curriculum test) has a high correlation to performance measured by a teacher over a period of time, it could be argued that the national curriculum test has concurrent validity. It concurs with other measures in the same subject or construct domain. Statistically, concurrent validity is often used in the construction of new tests. Their performance is compared against the old test to gain a statistical correlation. This only works on the supposition that the old test had its own concurrent validity. Predictive and concurrent validity are often referred to as criterion related validity, the ability to predict performance on a particular criterion.

Having described the component parts of validity, which serves the function of considering the construction and uses of assessments, it is now necessary to view them in a more unitary manner.

As far back as 1955, Cronbach considered a holistic view of validity:

One does not validate a test, but only a principle for making inferences (Cronbach & Meehl, 1955 p297).

Over the last twenty years researchers such as Messick, 1989; Cronbach, 1988; Gipps, 1994 and Wiliam, 1993, have encouraged educators to consider the outcomes of assessments more than their constituent parts:

Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment (Messick, 1989 p 13).

For a fully unified view of validity, it must also be recognised that the appropriateness, meaningfulness and usefulness of score-based inferences depend as well on the social consequences of the testing. Therefore, social values cannot be ignored in considerations of validity (Messick, 1989, p19).

Therefore:

Validity is not a property of tests, nor even of test outcome, but a property of the inferences made on the basis of these outcomes (Wiliam, 2000b, p2).

The issue of validity being viewed in the context of the consequences of a test lies at the heart of the assessment dilemma in England. I will return to this theme later.

2.3 Reliability

A unified approach to validity includes reliability, as confidence in inferences made from assessments will include administration, marking and grading procedures (Stobart, 1999). I will however, similarly to the section on validity, outline the component parts of reliability measures in order to illustrate pertinent issues.

One of the underpinning notions of reliability is that the results provide a particular yet consistent rank ordering of ability of a certain domain or trait. An ideal measurement of reliability would be that if a group of students took the same test twice, their result and the rank order would remain the same (Nuttall and Willmott, 1972).This is clearly an absurd notion in the context of assessment.

In theory, in order to measure reliability we would need to brain-wipe a set of candidates and make them do the test again, with no memories of questions or answers from their previous attempt, or tiredness or change of mood. Impossible, of course (Schagen, 1999, p 28-29).

As discussed in the previous section, the primary purpose of most assessments lies not in the actual test items themselves, but in their generalisation to some wider domain (Nuttall 1987). Therefore there is an implied public trust in assessment reliability, and its accuracy (Newton, 2004). Any academic or professional discussion on the form and nature of reliability in assessments make clear the view that there is no such thing as a completely accurate and reliable test. Wiliam (2001a) sets out the three major sources of unreliability: factors in the test itself, factors in the candidates taking the test and scoring factors (particularly who marks tests, and how well they do it).

Classical Test Theory (CTT) provides the simplest and most practical way with dealing with reliability issues within tests and is the most commonly applied statistical tool applied (Bartram, 1990), and dates back to the work of Charles Spearman in the early 20th century. It is usually represented by the following formula:

$$X = T + E$$

Where:

X is the observed score (the actual measurement obtained)

T is the true score (what the measurement would be if there were no error)

E is the error score (the influence of error on the measurement, also known as measurement error)

Wiliam (1993) considers classical test theory as an attempt to capture the idea of ‘signal-to-noise ratio’ for assessments. This is based on the assumption that an individual’s score contains error (noise) which can be decreased but never totally eliminated. It also assumes that the error is random and normally distributed. Another point to bear in mind that the true score does not mean a true measurement of ability, it is just an supposed average score that an individual would achieve over repeated taking of the same or very similar test.

Using this simple equation it is clear that when errors are small in comparison with the actual scores, a relatively high reliability is achieved, and when the errors are large in comparison with the actual scores, there is low reliability (Wiliam, 2001a).

The key formula to apply is:

Standard error of measurement (SEM) = $s\sqrt{1-r}$

Where:

s= the standard deviation for the test

r = the internal reliability coefficient for the test.

A reliability coefficient of 1 means that the standard deviation of the errors is zero and there is no error, so the test is perfectively reliable. A reliability coefficient of 0 means that the standard deviation of the errors is the same as that of the observed scores- the scores obtained by the individuals are all error, so there is no information about the individuals at all. If a test had a reliability of zero the result of the test would be completely random.

Using the standard deviation and the internal reliability coefficient of a test, the standard error of measurement (SEM) can be calculated. This is an estimate of the error when using and interpreting an individual test score. The larger the SEM, the less reliable a test or test score will be.

In a high stakes, public examination, William (2001c) argues that the internal reliability needs to be over 0.9 for the test to be considered reliable, however even high reliability measures can result in significant misclassification of students. At a high reliability co-efficient of 0.9 (for example Cronbach's alpha- which is explained on the next page), William (1995a) calculated that 30% of KS3 pupils could be classified at an incorrect national curriculum level. This error could be further compounded by marking reliability error, which is discussed later.

The main sources of error that threaten reliability are that:

- Students may perform better or worse depending on the particular questions chosen on a test
- Students perform better or worse on different days

- Different markers may give different marks for the same piece of work

(Black & Wiliam, 2002)

There are other variables that may contribute to the error component, some of which will also be discussed later.

The first source of error, that there will be a variance of performance depending on the questions chosen in a test, raises a number of issues. The perceived problem is that a test can only sample a small number of items and therefore gives an incomplete picture of a student's knowledge and understanding (Bartholomew, 2000). How often have we heard a student or indeed ourselves say that if only a certain topic had come up in an exam, they would have done better. One solution of course would be increase the length of the test, in order to increase the overall spread of scores, and thus reduce the mean error mark (Wiliam, 2000a). This notion is unviable as there is continued governmental pressure to decrease the length of summative assessments, and also the fact that in order to increase reliability coefficients in this way would necessitate doubling or even tripling the duration of tests. The counter argument to this is that if the test is correctly constructed it will sufficiently sample the correct constructs to make the result generalisable (Goldstein, 1994).

Classical test theory deals with this issue by calculating the internal reliability co-efficient of a test. The KR20 (for multiple choice tests) or Cronbach's alpha (for other forms of test) are the measures usually employed. These reliability co-efficients work on the principle that if a student scores highly on a particular item, this performance should be consistent with that student's performance across the whole test. Once all the scores from every component item by every pupil is entered, Cronbach's alpha is able to explore performance by splitting the test scores into any given way so that consistency of marks can be calculated. The maximum reliability coefficient is 1, with high stakes tests achieving usually between 0.85-0.95. This is a statistical indicator, estimating the probability that a given mark might be in error by given amounts (Black & Wiliam, 2006).

This all sounds impressive, but there are a number of provisos attached to internal reliability measures:

- The coefficients works on the assumption that the same construct is being tested throughout. This may be desirable or not depending on the subject area being tested and the nature of the assessment

- The coefficients are distributed evenly across the mark range achieved by the cohort. This is highly unlikely, as depending on the length of the test, the number of questions answered correctly will be significantly different

- The coefficients take no account of marking or grading errors and any other variables which may have affected a student's performance in a test

The only way there is an attempt to reduce the variable of students having good and bad days is to spread assessments or tests across a number of days. As described earlier, there is a limited opportunity to do this in a 'one-shot' testing culture.

The variable of inter and intra-marker reliability is the one given the most time and attention by researchers, national testing agencies, exam boards and the public at large (Wolf & Silver, 1993). Inter-marker reliability refers to how far different examiners mark work in the same way. Intra-marker reliability refers to how far an examiner marks equivalent work in the same way.

It is interesting, however inevitable, that marking reliability seems to have the highest profile when issues of reliability are considered. They receive the most coverage in the press (Newton, 2004), and are perhaps an easy target when simple addition errors or non-adherence to a mark scheme results in a particular grade or a pass not being achieved.

Within the marking procedures of exam boards, training and standardisation marking meetings and exercises are designed to ensure examiners work within a mark tolerance when applying a mark scheme, and if not, appropriate weightings to marks are applied. The costs of increasing reliability by applying double marking procedures are deemed prohibitive in terms of cost and time, and are therefore not used in large scale, high stakes national assessments.

As for the other areas of reliability measures, there are a range of statistical measures that can be applied to marking accuracy. It is unsurprising that the lowest figures of reliability are found in essay and extended writing questions that are open to subjective judgement. It is also unsurprising, that apart from research studies set up to study intra and inter-marking reliability, high stakes and public examinations cannot put figures to these errors. The only data they have is the number of appeals, remarks and adjusted marks in an exam series. This may be a small fraction of the actual errors. Newton (2003) has pointed out that even if marker reliability is very high (0.98), there may be up to 15% misclassification of grades. As high as this seems, Black & Wiliam (2006), consider this error small in comparison with the other reliability factors discussed.

2.4 The Relationship between Validity and Reliability

It is clear that the concepts of validity and reliability are not independent of each other. Reliability is the property of the assessment procedures themselves, whereas validity is a property of the information they produce. While both components are important features to consider, there is an inevitable tension between them. One view is that an assessment without high reliability cannot have high validity. If there is uncertainty about the accuracy of the assessment, then the extent to which it measures what is intended to measure must be uncertain.

Even those investigators who regard reliability as a pale shadow of the more vital matter of validity cannot avoid considering the reliability of their measures. No validity coefficient and no factor analysis can be interpreted without some appropriate estimate of the magnitude of the error of measurement (Cronbach,1951 p179).

According to classical test theory, the maximum validity for a test is the square root of the reliability (Magnusson, 1967). The problem with this argument is that the desire to increase reliability generally means the production of very restricted forms of assessments, response types and marking mechanisms. Many writers on this subject take the view that validity is the more important element, as there is no point measuring something reliably unless it is clear what is being measured. However, validity would appear to be more problematic because it is harder to measure (Crooks et al, 1996) and even philosophical in nature (Clausan-May, 2001). It is also commonly viewed that if validity is increased by extending assessment types, particularly in relation to higher order thinking skills, reliability is likely to fall, however this may not be a bad thing as the assessment type may be fit for the intended purpose.

As Sadler (1989) comments:

Attention to the validity of judgements about individual pieces of works should take precedence over attention to reliability of grading in any context where the emphasis is on diagnosis and improvement. Reliability will follow as a corollary (p122).

2.5 Dependability

The relationship between validity and reliability means that although each can be described separately, they are only manifested in a combined manner. This has led to a more unitary approach (Stobart, 1999) of the concept of dependability. This is expressed as:

Reliability + Validity = Dependability (Wiliam, 1993; James, 1998)

The relationship between validity, dependability and reliability can therefore be described thus:

- Validity is the extent to which inferences within and outside the domain of assessment are warranted
- Dependability is the extent to which inferences within the domain of assessment are warranted
- Reliability is the extent to which inferences about the parts of the domain actually assessed are warranted

(Wiliam, 1993)

Dependability is the intersection of validity and reliability (Gipps, 1994).

Using dependability acknowledges that the selection of an assessment methodology purely on the basis of gaining the highest reliability measures or one that would appear to have the greatest validity might result in an assessment that is not fit for purpose. Consideration of dependability suggests an essential trade off between reliability and validity to best effect for the assessment of a particular subject.

The problem with a unified approach to validity (dependability) is that there is no way to calculate a value, and the concepts involved are complex. It is argued that this results in the continued neglect of validity when assessments are monitored (Stobart, 1999) and that validation will only flourish if approaches are developed which help to organise our thinking about important validation questions and to identify issues which need particularly close scrutiny (Shepard, 1993).

Exam boards have been lucky not to have been engaged in a validity argument. Unlike reliability, validity does not lend itself to sensational reporting. Nevertheless, the extent of the boards neglect of validity is plain to see once attention is focused…the boards know so little about what they are assessing (Wood, 1991, p 151).

2.6 Unified approaches to Validity

Using a broader view of validity, one that now subsumes validity and reliability, and also one that acknowledges that validity is as much about the uses, inferences and consequences of an assessment, this chapter now sets out to describe two contrasting models to explore a unified approach to validity. The first model was developed by Messick (1980) and uses a theoretical framework to demonstrate the relationship between the basis and function of assessments. The second was developed by Crooks et al (1996) and proposes a functional framework to evaluate the threats to the validity of assessments. Rather than repeat similar points, I will give an expanded discussion on the latter model, particularly as it provides a more straightforward translation into current education assessment theory and practice.

2.6.1 Messick's framework for validity

Messick (1980) constructed the following simple matrix framework model shown in Figure 2 below:

Figure 2: Messicks theoretical framework matrix

		Function	
		Result Use	Result Interpretation
	Evidential basis	Construct validity **1**	Construct validity and relevance/utility **2**
Basis	Consequential basis	Value implications **3**	Social consequences **4**

Figure 1 shows the upper row indicating the technical conceptions of validity, while the lower row shows the consequences of the inferences of the outcomes of assessment.

If we explore each box within this matrix, we can see why great care needs to be taken both in the construction of high stakes assessments, but also in the inferences that are drawn from them.

Box 1

The construct validity of an assessment (particularly if it is high stakes) underpins the basis of the assessment itself and also the functions or implications made from the results. If the evidential basis of the test is taken from limited areas of the course of study, it is said not to adequately represent the intended domain, and therefore the threat to validity is the ‘construct underrepresentation’ and the limited confidence that can be generalised for the results. This issue often centres around the use or not of authentic assessment tasks, the argument being that many high stakes assessments are so narrowly constructed that they lack authenticity (Wiggins, 1993).

Box 2

This box explores the issue of how the basis of the assessment is related to usefulness of the result. If the construct validity is high, then the result will be a good indicator of future performance in the domain, but if there is construct underrepresentation, or lack of authenticity, then the utility of the result lacks validity. It would have limited uses to predict future performance.

Box 3

This box links the consequential basis of the assessment with its interpretation. This explores the issue that the nature of an assessment and the implications made from the results place an implied value on the assessment. If constructs are underrepresented or there is a lack of authenticity, it implies that certain areas of study or process skills are not valued. This threatens validity.

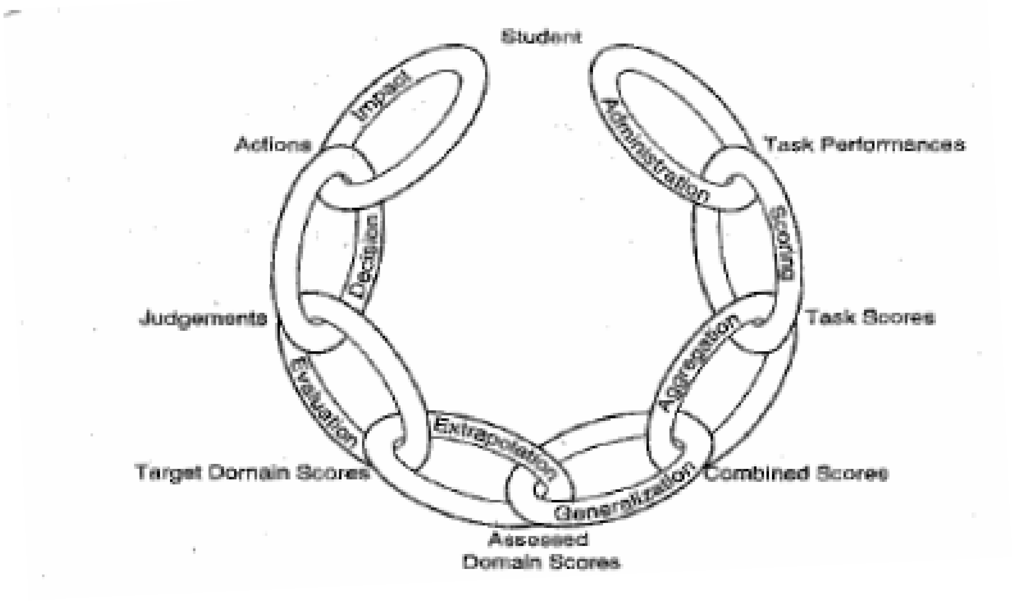
Box 4

This box then takes the premise of the lack of value of certain areas of a domain, or associated process skills to its natural conclusion. The consequences of distorted values of constructs actually assessed, rather than those intended to feature as a part of a course of study are that they often not taught at all, or lack emphasis because they will not feature as part of the assessment. This may not be an intention of the assessment, but it ends up as a consequence of the assessment, and therefore validity is compromised.

2.6.2 Crooks et al framework for validity

The second model was developed by Crooks et al (1996) and is shown in Figure 2 below. This model built on the work of Kane (1992) and Shepard (1993), who both discussed the idea that by working through the combined effects of the inferences and assumptions of an assessment, a view of validity could be established. Crooks et al moved this discussion on to create an eight stage linked model shown below in Figure 3. This model is not appropriate for all forms of assessment, however it is suitable in the large scale, high stakes assessment environment that this research study is based.

Figure 3: Crooks et al model of educational assessment for use in the validation and planning of assessments.



In Figure 3, assessment is depicted as divided into eight conceptually distinct stages, with validation then based on careful scrutiny of each of these stages. The eight stages are depicted as links in a chain, the strength of which is determined by the weakest link. An interesting dimension of this model is that it can be used (though adapted) to look at the validity of formative or summative assessment. It can also be looked at from steps 1 – 8 in terms of validating an existing model of assessment, but equally it can be considered in reverse order, looking at the intended outcomes of an assessment first and when working through the steps that underpin these. For the purposes of this paper I will work through the links from steps 1-8

A brief description of the eight links of the chain follows, together with some discussion and supporting exemplification.

2.6.2.1 Administration (link 1)

The first link explores the circumstances under which students actually take assessments. Four threats are identified:

Low motivation; inferences from performance only works if students have engaged and applied themselves to the tasks in hand. If students think they are going to fail anyway, or are disengaged with what they perceive to be meaningless activities, the validity and the inferences drawn for an assessment would be invalid

Assessment anxiety; this is the antithesis of low motivation. High anxiety can paralyse performance, and unfortunately can be exacerbated in high stakes assessments. There are some students however, who actually perform better than normal in these stressful one-shot testing conditions.

Inappropriate assessment conditions; This aspect involves the importance of correct procedures being applied, instructions read, necessary equipment and space available and correct allocation of time provided.

Task or response not communicated; this aspect covers a range of possible interferences. These include instructions and rubrics in assessments being unclear, ambiguous language in an assessment, and lack of accessibility to a certain domain or construct through particular disabilities, unrelated to the construct being assessed.

Stobart (1999) defends the position of national curriculum tests with regard to this link by suggesting that as they are high profile and high stakes assessments, and therefore there is no lack of motivation by schools or students to do well. Clesham (2004) describes the stages involved in the development of national curriculum tests, which are quite unlike any other examinations in this country as they are pre-tested and also reviewed by expert, teacher, EAL, SEN and accessibility groups before they are used. This gives some assurances that there will be no surprises, unforeseen or unconsidered issues arising from student performance. GCSE and GCE examinations do not however incorporate these elements in their development and therefore this does pose threats to validity.

Stobart (2000, 2005) examines the problems of presenting fair assessments in a multicultural society and suggests it is naive to assume that the content or assessment methods do not create bias. Stobart and Gipps (1998) give the following advice:

We need to encourage clearer articulation of the test/examination developers' constructs on which the assessment is based, so that the construct validity may be examined by test takers and users. Test developers need to give a justification for inclusion of context and types of response mode in relation to the evidence we have about how this interacts with group differences and curriculum experience (p48).

There is plenty of evidence, (see for example, Gipps and Murphy, 1994) which indicate differences between the performance of boys and girls on open ended and closed questions and differences in coursework forms of assessment. The lack of oral elements in assessment is often cited as one that discriminates against ethnic groups who value verbal forms of communication (Rudduck, 1999).

Wiggins (1993) gives a comprehensive airing of what is needed to produce 'authentic' meaningful assessment tasks, rather than judging performance in simplified and de-contextualised ways, which contribute little to identifying and measuring constructs. Even more worrying is the evidence from Pollitt et al (2007) which suggests that as a general rule, nervous, anxious, borderline (NAB) students operate at two years less than their chronological age in high stakes test conditions. If this is the case, validity is compromised at the very start of this eight stage model.

2.6.2.2 Scoring (link 2)

This link explores the errors that may be implicit within an assessment's scoring mechanisms. Five threats are identified:

Scoring fails to capture important qualities of task performance; this includes mark schemes and marker training not recognising creditworthy responses that may be perfectly valid, or more commonly scoring being based on a narrow set of skills (to increase marker reliability) and therefore missing important features of either a domain or process skill (eg in oral reading, evidence could be gained by a large number of factors that each contribute to the end result).

Undue emphasis on some criteria, forms or styles of response; Care needs to be taken that assessments do not place undue emphasis on perhaps the use of standard English, or punctuation, spelling or grammar, if the domain being assessed is of a different nature.

Lack of inter or intra-rater consistency; it is important that markers apply a mark scheme consistently across all the work they mark, and that all markers do the same, using the same standards. Ensuring this consistency increases reliability, but can also narrow the focus of assessments as complex skills often require expert or professional judgement.

Scoring too analytic; this threat deals with an assessment taking a task apart and marking on a micro level, rather than assessing the effectiveness of the outcome, even if it has taken a very different route to get there.

Scoring too holistic; this is the opposite argument, and suggests that if an overall grade is given, particularly for a substantial piece of work, little formative use can be applied in highlighting strengths and weaknesses of performance.

In high stakes, one-off summative assessments, the reliability of mark schemes and markers are critical elements. Stobart (1999) and Clesham (2004) describe the procedures used in national curriculum testing that attempt to ensure high accuracy in scoring processes. Pre-testing of items allow student responses to appear in mark schemes and extensive marker training exercises provide quality assurance measures on marker reliability to apply both the letter and spirit of the mark scheme to apply.

A major issue to discuss here is the acknowledgement of marking unreliability. There is no lack of research and evidence to explore the factors associated with marker reliability. Willmot & Nuttall, (1975); Murphy, (1978, 1982); and Baird et al, (2002, 2003), are just a few of literally hundreds of studies that investigate how different styles of questions and mark schemes are applied by examiners. As mentioned previously, there are obvious trends in reliability in terms of more closed questions having higher reliability measures, however Newton (2003) points out that even with very high values of marker reliability, the proportion of candidates likely to be incorrectly graded is still likely to be large.

Baird and Mac (1999) calculated that a near perfect reliability measure of 0.98 can result in 15% of candidates getting an incorrect grade, while a reliability of 0.90 (still a high figure) can result in 40-50 % misclassifications. This is due to the fact that grade boundaries lie on a particular mark point, so one mark either way can be significant.

Newton, (2004) argues that there should be better public understanding of unreliability and therefore a more considered view on the importance and limitations of reliability in high stakes assessments. Interestingly, as much as Wiliam (2001a) describes the consequences of marker unreliability, he considers this source of error is actually not as large as those associated with the test construction itself.

This link therefore, exposes the tension between reliability and validity. Where holistic marking operates, with the intention of improving greater validity, marker unreliability is at its highest, and there are the highest number of marking reviews. This is particularly evident in essay based subjects, such as English and History.

2.6.2.3 Aggregation (link 3)

This link explores the issues of scores on individual tasks that are then aggregated to produce subscale or total scores. The threats are:

Aggregated tasks too diverse; If an assessment has many diverse components to it, which may add to its construct validity, the aggregation of the component scores or grades into one summative grade or level may render them meaningless. An example that is often cited to exemplify this is that national curriculum science reports one single summative level of attainment, where it subsumes the performance of the three separate sciences (biology, chemistry and physics). Similar arguments can be applied to English and Mathematics, that each consist of component aspects of the subject.

Inappropriate weights given to different aspects of performance; if aggregated scores are to be used, it is important that the weights of component parts are weighted to reflect their relative importance. This might be to do with weightings associated with different areas of a domain, or it may be to do with the question type e.g. different weightings given to multiple choice and essay type questions.

Unlike high stakes examinations in the USA, there are no high stakes assessments in England that rely solely on a narrow type of response like multiple choice. However, it is still evident that summative exams have to take place in limited time scales and rely on written performance of a sampled selection of a syllabus or specification. Wiliam (2000b) argues that the narrowing of the assessment of a construct into formulaic summative tests can invalidate the assessment and the inferences one can draw from it. On the other hand, using statistical measures, it can be seen that internal reliability increases very slowly with time duration. Wiliam (2001a) calculated that KS2 test would require more than 30 hours of testing (instead of the current time of two hours) for each subject to decrease the possible misclassification of pupils by 20%. This is an unworkable solution for 11 year olds (and probably any other age group). Newton (2003) argues that as learning outcomes in schools are defined by programmes of study or specifications, they effectively act as the construct. Therefore as long as the specification is adequately sampled, then construct validity is maintained. Cresswell (1996) describes how the aggregation and weighting of distinct facets of a domain through a sampled assessment can actually enhance the assessment's validity.

2.6.2.4 Generalisation (link 4)

This link essentially deals with reliability, in particular whether a student's mark on a particular assessment provides a dependable measure of attainment in the target domain. Threats are:

Conditions of assessment too variable; student performance may be dependent on a large number of variables, the time allowed, when in the day they take place, the format of the assessments and the levels of control applied by teachers. Failure to control these factors may invalidate generalization.

Inconsistency in scoring criteria for different tasks; there needs to be a level of correlation between the scores of different tasks within an assessment. If the scoring criteria are similar, the reliability improves.

Too few tasks; reliability is decreased if the assessment only samples small portions of a subject domain.

This link exposes yet another paradox. The very things that aid generalisation, hinder it. As Stobart (1999) explains, national tests and exams set down strict constraints in terms of when, where and how they are conducted. This assures standardisation and an amount of equality across the country, but clearly it sometimes discriminates against students who are having a bad day or any particular difficult circumstances (Wiliam, 2001c).

The other important element here is the role of psychometric application and measurement. General principles of the use of classical test theory have been described in the first part of this paper. Quinlan and Scharaschkin (1999) describe the statistical measures used in national curriculum testing in order to achieve measures of reliability, however at the same time these measures equally expose possible inaccuracy and error. Psychometric statistics do not attempt to hide these issues. Psychometric texts (Aiken & Groth-Marnet, 2003; Lewin, 1997; Bartram, 1990; Schagan, 1994) describe levels of confidence that can be applied to assessments, with the assumption that perfect reliability is not possible. Researchers have done much work to interrogate, discuss and publish sources of error and suggest solutions in terms of alternative models and systems of assessment (Wiliam, 1995b; Bartholomew, 2000; Brooks and Tough, 2006). Harlen (2004) highlights this issue as a key aspect of the need to promote educational literacy to politicians, schools, teachers and parents. Black (2003) comments that:

There are no serious attempts to research the effects of public examinations, let alone publish the results of such research. If this were to be done, it seems likely that the revelations of the chances of error would cause public concern (p75).

Wiliam (2000a) argues that exam boards and government agencies do not do enough in terms of publishing reliability measures for particular high stakes assessments. While this is still true of exam boards, QCA do now publish Cronbach's alpha for national curriculum tests.

Although the limitations of psychometric measurement have been mentioned earlier, Goldstein and Heath (2000) suggest that psychometrics has provided the only sustained attempt to provide formal frameworks for addressing key reliability issues.

Broadfoot (1996) however, has been a consistent critic of tests and psychometrics in the setting and maintenance of standards. It is '*the necessity to make inferences from the small particular to the larger whole that lies at the heart of the myth of measurement'* (p 207). Her view is that there is little scientific justification in the way psychometrics are applied, they are not applied objectively, however, their real danger is that people believe in them (Broadfoot and Black, 2004), and even worse that they '*have inhibited the positive and creative use of assessment to promote, rather than to measure learning'* (p219).

2.6.2.5 Extrapolation (link 5)

This link deals with how effectively an assessment samples the target domain. The threats are:

Conditions of assessment too constrained; this may be evidenced by a biased sample of questions that do not adequately sample the domain or possibly by selecting a question type, (eg multiple choice) that also do not allow essential elements of a domain to be assessed.

Parts of the target domain not assessed or given little weight; this re-emphasises the point that if a construct is underrepresented (Messick, 1989), validity is compromised.

Most of these issues have been discussed in other areas of this chapter. Newton, (2003); Stobart, (1999); and Black, (1998), have argued that a specified syllabus or programme of study effectively act as a construct in school based curricula as they specify content and skills expected are expressed in grade criteria or level descriptors. As mentioned before, Goldstein (1994) believes that as long as the sampling of the domain is weighted correctly in terms of relative importance, validity is not threatened. Barthlomew, (2000); Broadfoot, (2006); and Wiliam, (2000a), are among many commentators who dispute that sampled assessments can ever claim to be fair and representative measures of attainment. They argue that apart from the thorny issue of what is a construct, sets of questions will always favour some students and not others, types of question perform differently, and even question order can affect performance. Pollitt et al (1985, 1999, 2007) and Wolf (1991) amongst others, have carried out empirical studies to demonstrate how question wording, style and structure can affect the way students respond to them. It is questionable, therefore, how these can therefore have a generalised dependability attached to them.

2.6.2.6 Evaluation (link 6)

This link is concerned with whether those interpreting assessment information really understand it and are aware of its limitations. Threats identified here are:

Poor grasp of assessment information and its limitations; this is probably at greatest risk when teachers or schools interpret information from assessments they have not written, i.e. standardised or end of unit/module tests. No test would claim to deal with absolute values, but schools often report them using degrees of accuracy that do not actually exist, as discussed throughout this paper. Even if a teacher has written a test, care needs to be taken that a result is not claimed to assess a subject domain if it only actually assesses small knowledge based portions

Inadequately supported construct interpretation; this essentially makes the same point. Care needs to be taken that large inferential leaps are not taken from performance to imply construct validity. If, for example an assessment of physical education was taken only using written evidence from students, the construct would clearly be inadequately supported.

Biased interpretation or explanation; this can work in a number of ways. Sometimes, a teacher may rightly moderate a judgement of a performance using their professional judgement of a pupil, but this can lead to a student not getting the acknowledgement or credit for high performance or conversely a teacher applying 'the halo effect' where a low performance is discounted because the teacher believes a student to be better than their assessment shows.

Wiliam (2000a) uses a powerful apocryphal tale to demonstrate how poor application and understanding of information can be dangerous bedfellows. Objective measurement is clearly appropriate in many areas, but it often dominates agendas and belies the fact that many important and pertinent factors cannot be measured in this way. This point is well illustrated by what is referred to as 'Macnamara's Fallacy'. This is named after a US Secretary of State during the Vietnam War. He argued that the ratio of Viet Cong/North Vietnamese Army losses to the US/Army of the Republic was an important measure of military effectiveness. '*Things you can count, you ought to count. Loss of life is one*'. Charles Handy (1994) cited in (Wiliam, 2000a), described this strategy thus:

The MacNamara Fallacy: The first step is to measure whatever can easily be measured. This is OK as far as it goes. The second step is to disregard that which can't easily be measured or to give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily really isn't important. This is blindness. The fourth step is to say that can't be easily measured really doesn't exist. This is suicide (p111).

Wiliam uses this theme throughout his critiques on national testing systems:

*We start out with the aim of making the important measurable, and end up making only the measurable important (*Wiliam , 2001b, p58*).*

This serves as a useful tale to analogise what can happen when data is sought out, without really engaging with what the data is telling us and what it actually means.

2.6.2.7 Decision (link 7)

This step involves deciding what actions are taken using the results of an assessment. There is clearly a close link between this link and impact. A decision could be selection for a particular course and therefore should be consistent with the information from the assessment. Again, two threats to validity are identified:

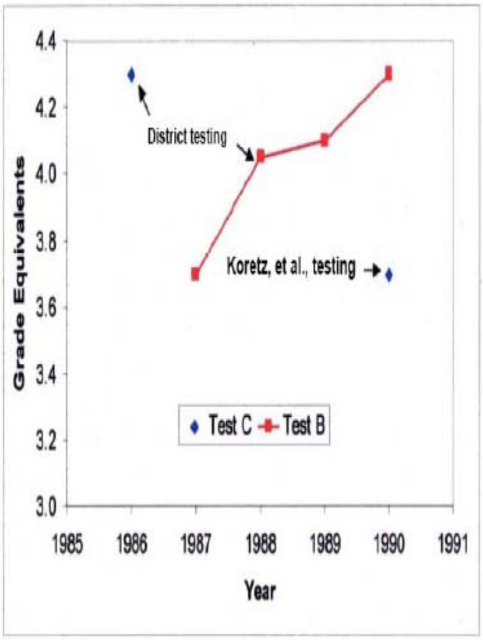
Inappropriate standards; if a decision is taken on the basis of a particular grade being achieved, there must be confidence that cut scores for grades have a sound foundation. This can be dependent on the outcomes of the assessment. If the assessment is simply a filtering device (eg a selective school test), all that is required is a test which fairly rank orders students on a given construct. If the grade is designed to indicate a particular place on a progressional scale, care needs to be taken that where these boundaries are set are consistent with understood identifiers of performance at particular levels.

Poor pedagogical decisions; this is concerned with actions taken as a result of an assessment. How teachers interpret and feedback information to students and parents is a crucial element of assessment. If information is 'cherrypicked', then feedback may be misleading and unnecessarily positive or negative. Another problem can be that the results of assessments simply do not seem to offer any progressional information and so have little or indeed negative effects on pedagogy.

The 'standards' debate looms large over any discussion concerning the processes and outcomes of high stakes assessments. Massey, (1995); Stobart,(1999); Shorrocks-Taylor, (1999); and Newton, (2003), all describe the reasons and significance of the shift in the theory and practice of the application of standards in national curriculum assessment in the mid 1990's. Rather than using tightly constructed performance criteria, they shifted to more loosely defined constructs based on a best fit model. Therefore standards moved from a 'specific competence' model to one of 'general competence'. This model also describes how grade boundaries in GCSE and GCE are established. This shift in thinking and approach already exemplifies how definitions of standards change, and therefore the maintenance of standards over time is a fairly meaningless proposition (Brooks and Tough, 2006).

Black (1998) illustrates how the educational measurement of standards can distort educational intent. In 1987, all fifty states in the USA achieved above average results in national standardised tests. Unlike the UK, states in America can select tests from a range available. This was coined the 'Lake Wobegon Effect' after Garrison Keillor's mythical town where all the women are strong, all the men are good looking, and all the children are above average. Many states and schools had a good track record and did not need to change teaching and learning strategies to achieve good results. It was the strategies in poorly performing schools and states that illustrate the point. They had pressure to improve their results, which they duly did by selecting a test, and only teaching material on that test. The students clearly got better on the performance in that test, as the national data showed, but nothing else, as was proved when they took another test. Wiliam (2007) describes a specific example of this in an American state system.

Figure 4: Test performance over time in an American District



SOURCE: Adapted from Koretz, Linn, Dunbar, and Shepard (1991)

Figure 4 shown above shows that in 1986, this district was administering Test C, with a high degree of pupil performance (grade average 4.3). In 1987 a new test was administered (Test B). performance fell sharply (to 3.6), but then over a number of years rose to 4.3 once more. At this point Test C was re-introduced with predictable consequences (it fell to 3.6 grade average). Rises in achievement seem to have more to do with familiarity with a form of test, or teaching to the test than rising standards.

Tymms (2004) demonstrated the same apparent effect in performance in national curriculum performance over time. He used evidence from comparabilty studies commisioned from QCA and comparabilty of the changes to the assessment models used over time to conclude that apparent rises in standards over time were illusionary, and had more to do with differences in marking tolerances and standards and the fact that test coverage had changed, assessing different areas of subjects (eg mental mathematics).

Wiliam (2001a) concludes therefore that the notion of apparent rises in standards in high stakes assessments is fallacous as they do not actually serve as proxys for wider achievement and potential, but rather, effective teaching to a narrow range of skills.The phenomena of 'teaching to the test' is a good example of Goodharts law. Charles Goodhart was a chief economist at the Bank of England. The example he used was the relationship between inflation and money supply. Economists had noted that increases in the rate of inflation seemed to coincide with increases in money supply, although neither seemed to have any relationship with economic growth. This led to the simple assumption that controlling money supply would also control inflation. Unfortunately the effect of this policy was a huge slump in the economy.

The very act of making money supply the main policy target changed the relationship between money supply and the rest of the economy (Kellnor, 1997) cited in (Wiliam, 2001b, p60*).*

Goodhart's law exposes the consequences of selecting particular performance indicators to act as a proxy for overall improvement. Manipulability of performance indicators destroys the relationship between the indicator and the indicated. In other words, the clearer you are about what you want, the more likely you are to get it, but the less likely it is to mean anything (Wiliam 2001b).

2.6.2.8 Impact (link 8)

This step is distinct in that it deals with the consequences of an assessment rather than the assessment itself. This supports Messick (1989), who calls this the consequential basis of validity.Two threats to validity are identified in this stage:

Positive consequences not achieved; this includes using information gained from assessments to aid progression, provide useful feedback to teachers and students on teaching and learning and to act as a motivating factor

Serious negative impact occurs; this is the flip side of the coin. Potential negative consequences would be that students might lose motivation, be excluded from further learning opportunities and self efficacy is reduced. These consequences would be bad enough if the assessment was valid, but if it isn't the continued use of a particular assessment cannot be justified.

All the previous seven stages lead to the impact of an assessment, but it has been argued that this stage actually is the main driver in high stakes assessments and the whole eight stage model could effectively be applied in reverse as the meanings and consequences of tests and examinations dominate the application of a validity model.

The first section of this chapter set out an emergent idea that validity is as much about the inferences and uses that are put to the outcomes of an assessment rather than the content and circumstances of the assessment itself.

The notion of multiple purposes of assessment is of course nothing new. Black (1998) describes how payment by results was a feature of 19th and early 20th century education in England, and so clearly, from an early stage of national educational provision, the outcomes of pupil performance had more riding on it than purely educational achievement and progression.

Newton (2007) identified three main purposes of an assessment system:

- To generate a particular kind of result, eg. to rank students in terms of their end-of-course level of attainment. This purpose is about the ways assessment results are represented;
- To enable particular kind of decision, eg. to decide whether students have learned enough of the basic material to allow them to enrol on a higher-level course. This purpose is about the uses of assessment results;
- To bring about a particular kind of impact, eg. to require teachers to align their teaching with a national curriculum. This purpose is about the consequences of assessing.

These purposes may appear manageable to consider, however if we consider just one of these purposes, eg. the uses of assessment results, Newton (2007) identified 22 different uses, listed below. The highlighted decisions show the 14 multiple uses of national curriculum tests alone:

1. **Student monitoring**
2. **Formative**
3. **Social evaluation**
4. Diagnostic
5. Provision eligibility
6. **Screening**
7. **Segregation**
8. Guidance
9. **Transfer**
10. **Placement**
11. Qualification
12. Selection
13. Licensing
14. Certification
15. **School choice**
16. **Institution monitoring**
17. **Resource allocation**
18. **Organizational intervention**
19. **Programme evaluation**
20. **System monitoring**
21. **Comparability**
22. National accounting

Ken Boston, (2007) chief executive of QCA, called this the 'swiss army knife model, one knife, but multi-purposed'. This then is the problem; if we are to consider the validity and fitness for purpose of an assessment, the question is, validity for which use? The consequence of this question is that an assessment may be perfectly valid for one purpose and totally invalid for another. '*A test is never valid, only valid for a particular purpose' (Wood, 1991).*

Teachers, generally, are quite complimentary about the structure and content of national tests (eg.QCA, 2002, 2003, 2004). However they dislike national testing because the accountability purpose dominates their school agenda.

If you have a system in which you take those tests, put them into league tables and send Ofsted inspectors in to hold people accountable, schools will test a lot more (Tymms, 2007).

This returns to the point that Wiliam (2000a) makes, that instead of an assessment being a sample of a construct, it becomes the construct, and teaching to a test for accountability purposes results in a loss of meaning and any real educational validity.

2.7 Summary

All the articles and extracts discussed in this chapter demonstrate that reliability, validity, dependability and fairness are elusive concepts in current high stakes, paper-based assessments.

By now it should be clear that there is no such thing as a fair test, nor could there be: the situation is too complex and the notion simplistic (Gipps and Murphy, 1994, p273).

Newton (2007) gives some clarity to the dilemma of high stakes assessments by suggesting that stakeholders need to ask themselves whether the positive impacts outweigh the negative, and for whom? The key message being that assessment is all about approximation, indeterminacy, trade-off and compromise- as long as the outcomes serve the greatest good.

Therefore in conclusion, it seems that there is not, nor ever has been a golden age of reliability and validity in terms of high stakes assessments in England. Perhaps a change of assessment practices in terms of examinations and tests moving to computer-based forms can result in assessment materials and outcomes no worse than the current policies, procedures and practices produce and could potentially open up possibilities of a system more fit for purpose.

The next chapter will discuss how computer-based assessments have developed over time and how their potential is tempered by accessibility, comparability and equivalence issues.

Chapter 3

Computer-Based Assessments

3.1 Introduction

This chapter reviews the literature regarding computer-based assessments, starting from early forms of multiple choice tests, (MCQ) tests, particularly in the context of comparative studies where paper and computer-based forms of the same assessments were available. Equivalence issues and the potential movement towards innovative and simulatory computer assessments will then be discussed.

3.2 Computer-based assessments pre -1995

Computer based testing and assessment systems seem such a 21st century concept and technology, however these forms of assessment date as far back as the early 1970's where computerized military and psychological testing systems were first used in the USA. The drive then was partly the same as now in terms of efficiency; the need to speed up the marking and feedback of results and the reduction of marker error by the use of automated scoring. From those early beginnings, it has remained a focus of research attention to question whether there is equivalence between computer and pen and paper modes of assessment, particularly in basic testing formats, where there are no assumptions of differing constructs being assessed.

Much of the early literature on equivalence discussed in this section concentrated on accessibility, familiarity and computer efficacy issues. As computers were a relatively novel resource, it was surmised that there could be differences in performance between computer and paper-based assessment on the basis of test mode.

However, early comparability studies indicated that there were not simple relationships between performances in these two different assessment modes. Mazzeo and Harvey (1988) reviewed 38 studies covering 44 tests. Eleven of these tests performed better on computer, eighteen indicated no difference, and fifteen showed better performance on paper. They looked at patterns underlying differences and found significantly more omissions on computer tests and difficulty reading figures off graphs. Reading long passages on screen also seemed to disadvantage students, and there appeared to be a tendency for random errors on the computer, perhaps caused by pressing incorrect keys.

Bunderson et al (1989) analysed 23 comparability studies. In these, three studies showed higher achievement on computer tests, eleven showed no significant difference in performance although paper did perform very slightly higher, and nine showed higher performances on paper based tests.

Bergstrom (1992) compared and synthesised 20 studies from 8 research reports comparing performance on computer and on-screen tests. While the tests were generally comparable, the means scores on paper were consistently higher than computer-based tests.

Mead and Drascow (1993) carried out a meta-analysis of 28 studies incorporating 159 tests on a variety of timed power and speeded tests for young adults and adults, largely taken from military aptitude tests. Power tests are characterised by being content knowledge led, whereas in speeded tests, time taken to work through questions is measured. They found no significant differences in the power tests, but students performed less well in timed (speeded) tests on computer.

It is probably not surprising that there was so much variety across these four large synthesised studies. The range of subjects, skills, interface design and question types assessed and the types of student, ages and computer experience varied greatly, however there were a few overarching generalised findings. One simple outcome was that any form of equivalence cannot be assumed, and needs to be verified and evidenced for any particular assessment; another that there are different forms of equivalence, which will be discussed later.

3.3 Computer-based assessments post -1995

After these initial studies and reviews up to the mid 1990’s, there were two significant changes in the drivers for computer based assessment. One was the technological hardware advancement apparent in schools. This was evident by large increases in the numbers of computers in schools and allied to this, increased internet access and coverage, latterly enhanced by broadband. The other significant driver lay in the increased profile of local and national assessments, both in formative and summative forms.

Since the mid 1990's the number of computer based testing and assessment programmes increased significantly in the USA educational system. They have subsequently become mainstream in many diagnostic and learning programmes, but more significantly, they have become mainstream in many areas of high stakes school, college and University assessments. Computer-based assessments, using predominantly objective questioning techniques have proved to be fast and efficient in terms of administration, marking and providing results and performance feedback to centres and individuals and for selection procedures. One of the consequences however, in such a highly litigious society, has been the need for a vast increase in comparability and equivalence studies to defend grading and selection outcomes.

Even though the early work carried out by Mazzeo & Harvey, Bunderson et al, and Mead & Drascow were based in largely different contexts and covered a range of different forms of psychological and aptitude tests, they were useful in identifying performance differences in certain areas. The number and range of results and evidence gathered in early studies also made it clear that there were not always straightforward relationships when considering comparability, and that it was necessary to consider each assessment type in its own right, and also consideration of the interface design, question and response types in detail when establishing equivalence.

In the USA, high stakes assessments such as the SAT's (Standards Assessment Tests), and GMAT's (Graduate and Managerial Admission Test) have computer as well as paper-based versions. There are many K12 (18 year old) testing programmes that are used for grading and selection procedures. As computer access and familiarity has increased, there has been more consistent, comparable and equivalent scores achieved in either medium.(eg, Russell & Plati, 2001; Poplum et al, 2002; Choi & Tinker, 2002; Pommerich, 2004; Wang, 2004). Another suggested reason for increased comparability is the improved navigational tools available to the computer users, which simulate paper test taking strategies. These include facilities such as skipping items, going backwards and forwards in a test and the ability to change or amend answers.

The strategies mentioned above are largely associated with answering multiple choice questions, and while there have been more comparability studies and research carried out with these question types, there has also been a number of studies carried out using other forms of objective questioning, short and extended answers formats.

Alternative and innovative objective questions, either through the stimuli given or the type of response required seem to result in inconsistent patterns of performance. Pommerich (2004; 2007) has suggested that the more complicated it is to present information on-screen or respond to a question, the greater the possibility of test mode effects, particularly for lower attaining students. This may be an efficacy issue (eg. Horkay et al, 2006), or it may be that there is more cognitive workload presented to students through certain on-screen question types that are not active factors for able students, but disadvantage the less able (Noyes et al, 2004).

Russell (1999) investigated open responses in science, language arts and maths and found significant differences in science, favouring performance on computer, a little difference in maths, favouring performance on paper, and no differences in performance in language and arts.

Nichols, (1996); Russell & Haney, (1997); and Russell, (1999) all showed that students tended to write more on computer open-ended questions than on paper, although the writing was not better quality or more creditworthy. However, Russell and Hanley (1997; 2000) also concluded that students who were confident and familiar with writing on computer scored significantly higher on computer in maths, science and language arts. One area that has remained persistent in disadvantaging pupils taking computer tests involves the reading and comprehension of long reading passages. Even with enhanced page layout and navigation tools, the working strategies of highlighting sections or phrases, locating information and general navigation throughout the text seems to make the two modes of delivery not equivalent (Murphy et al, 2000; O'Malley et al, 2005).

Given the increased accessibility and familiarity of computers and computer based assessments within education, it might be supposed that the mode of delivery at this point in time would no longer be an issue, however the only area that this can be generally accepted as true is for very basic multiple choice tests where the questions and response types are presented in identical formats in differing modes.

3.4 Computer-based testing in England

The studies and research discussed so far have been based in countries other than in England. This has been partially due to differing views of assessment methodologies. The educational system in England has continually resisted large-scale use of multiple choice testing for high stakes assessments, arguing that the content and construct validity and general fitness for purpose are compromised. As most research and studies pre-1995 were carried out in multiple choice comparability, it is understandable that the level of national interest in England was limited. Attitudes to computer-based high stakes assessments have developed since 2000 from a position of academic interest, however non-engagement in the desire to change existing tried and trusted paper-based custom and practice, towards a high level of interest in solving assessment validity and logistical problems that have resulted from large scale governmental testing programmes.

Ken Boston's aspirations and expectations for e-assessment (2005; 2007), were ambitious in the extreme, and unsurprisingly, were not met. This is in part due to developmental lead in time, but also largely due to equivalence issues. As most large scale, high stakes assessments are not predominantly in a basic multiple choice format, the lack of equivalence and maintenance of established 'standards' have presented a significant hurdle in terms of incorporating new assessment technologies (Wheadon and Adams 2007).

There have been a small number of small scale research studies in England investigating on-screen question types. Johnson & Green (2004) conducted maths tests to primary students in paper-based and computer forms. Even though there were no significant differences in overall performance between the tests, there were differences between individual items, particularly those that involved different working strategies between paper and computer. Where working out was required, computer performance fell below that of paper.

Thelfall et al (2007) also looked at maths assessments, and converted selected KS2 and KS3 question types into an on-screen format. Similarly to the Johnson and Green study, while the overall performance between modes of tests were comparable, the differences in certain question types were marked, and indicated that equivalence in terms of scoring or validity could not be established. They use the term 'affordance' to describe the effect of the interaction between the student and computer interface on their response to a task. Affordance is therefore a key issue in comparability and equivalence.

Small scale use of computer multiple choice testing alongside a large paper-based cohort within an exam board has been reported by Wheadon and Adams (2007) and displayed differing performance between the two modes, with students performing better on the computer test. However, as the numbers were so low taking the computer versions, no generalised conclusions about equivalence could be claimed.

3.5 Accessibility, comparability and equivalence issues

Apart from empirical national statistics indicating that there are still sections of society and schools that lack easy access to computers (e.g U.S. Department of Commerce, 2002; Becta 2006), there are clearly qualitative reasons for the lack of comparability between particular computer and paper-based tests which make generalised findings difficult to establish. These include age and background of student, the subject being tested, the presentation of the questions on-screen and the response type, computer experience, familiarity, anxiety, efficacy and attitudes (eg. Luecht et al, 1998; Taylor et al, 1999; Levine et al, 1998; Chua et al, 1999; Brosman, 1998; Al-Gahtani and King 1999; Singleton et al, 1999). Each one of this non exhaustive list can lead to some difference in performance, and therefore the elimination of all these variables to establish proven consistent and reliable equivalence is unlikely in the near future.

Bennett (1998) discussed the probable development and potential of computer assessed environments in education. He described a progression through three generations of testing systems:

1st Generation

This generation was categorized as the use of traditional skills and test formats and designs that mimic paper-based versions, utilising limited technology.

2nd Generation

The second generation will use new item formats including multimedia and constructed responses. These might assess and measure new constructs.

3rd Generation

The third generation will integrate instructional and assessment electronic tools that can sample performance repeatedly over time. This generation will use complex simulations including virtual reality that models real environments and allow more natural interaction with computers. These will assess and measure new skills and constructs.

There are therefore two conflicting drivers at work. The first is the desire by many to promote and utilise the full power and potential that computer assessment can increasingly offer, creating assessments that support learning and instruction that paper cannot (Bennett, 2002). Equally there is the desire, through bitter experience, not to simply invent new technologies that recycle current ineffective practices in assessment, particularly the measurement of narrow and constricted skill sets (Ripley, 2004).

Most high stakes assessments are dominated by the use of psychometric measurement (Goldstein, 1994). A simple description of psychometrics is that of a discipline that attempts to apply statistical principles and models to the measurement of particular mental dimensions or traits. The problem lies in the fact that most of these mental traits were developed using behavioural psychology of the early 20th century (Shepard, 2000). Although the sophistication of psychometric measurement and analysis has advanced enormously over the last twenty years, Bennett (2001) argues that we are basically measuring the same things over and over again, and ignoring the measurement of other cognitive constructs that have increasingly been acknowledged over the same twenty years. These include knowledge organisation, problem representation, mental models and automaticity (Glaser, 1991).

The conflicting driver is the need to establish equivalence with current assessments and the equally important issues of accessibility, inclusion and lack of disadvantage (Ripley 2004). As this literature review has indicated, other than in basic multiple choice formats, there are equivalence issues between computer and paper-based tests. This lack of equivalence may be the result of modal differences and lack of familiarity, experience and confidence within them, or they may be the result of different constructs being assessed, and therefore direct comparability and equivalence is simply not an appropriate methodology.

Although it would solve many issues to take this stance and effectively draw a line in the sand of current high stakes assessment measurement and 'standards' equivalence, it is most unlikely that this will actually happen. As McDonald (2002) points out, the inherent conservatism of many educational systems, and England is certainly one of them, has resulted in slow uptake of new assessment technologies, and change is at best gradual. The assessment and accessibility consequences are that any assessments provided in a computer format also have to be made available in paper form, and will probably co-exist as chosen options for some time. This situation ensures that equivalence cannot be ignored and has to be a significant feature of exam and test development and awarding methodologies.

Mead and Drascow (1993) indicated that there were two forms of equivalence, and it is essential to differentiate their forms, causes and possible solutions. The first relates to scoring equivalence. If apparently identical (apart from the medium) tests showed consistent score differences, the equivalence could be established through linear or equi-percentile equating. So, it might be that 70 marks on paper equates to 65 marks on computer. The mark difference may not be consistently 5 marks difference across the mark range, but equating graphs show what the difference is at each point. This equivalence can be confirmed if the overall rank order of candidates does not change.

The second form of equivalence relates to the construct validity of a test or assessment. This is a measure of what the test is designed to measure:

Construct validity is evaluated by investigating what qualities a test measures, that is, by determining the degree to which explanatory concepts or constructs account for performance on the test (Messick, 1989, p16).

Are computer and paper based assessments assessing and measuring the same things? Differences in performance between two tests and differing item statistics may indicate that something other than what was intended, is being assessed. This can be indicated by looking at discrimination values and internal reliability measures of items and tests respectively, and is confirmed if the rank order of candidates is affected between the two tests (McDonald, 2001).

The problem with the consideration of construct validity is that constructs are often difficult to identify, and performance differences can be attributed to 'construct irrelevance' issues: that is, factors that seem to be affecting performance (eg efficacy, attitude, anxiety) that are not active assessment constructs.

These two measures of equivalence lie at the heart of the computer assessment dilemma. The only assessments that consistently show equivalence between modes are basic multiple choice formats, and therefore they are the only ones that are found in high stakes assessments. In terms of the three generations of testing systems (Bennett, 1998), this has resulted in stagnation in the first generation, where there is limited use of technology and tests simply mimic paper test versions (Wheadon and Adams, 2007).

It would appear that in order for high stakes assessments to move through to the second and third generations of large scale, high stakes educational assessments, there will need to be two significant movements. One is for construct irrelevant factors to be eliminated. Sutton (1997) is one of many observers who suggest that computer accessibility issues, alongside any anxiety, efficacy, familiarity or attitudinal factors are short term obstacles, and this literature review has indicated that differences in performance have reduced in the past ten years. However, the continued inconsistency of research data would also suggest that there are significant construct irrelevant factors in operation. Negroponte (1995) suggests that the answer to these issues is in the development of such good interface design that effectively makes them 'go away'.

Once the construct irrelevant factors can be identified and eliminated, the next step will then be to identify and accept that different constructs are evident when comparing assessments in different modes, or to actively design on-screen assessments that are assessing different constructs to paper based versions, with paper-based assessments being taken by exception rather than by choice. While this notion may seem fanciful, there are for example, areas of high stakes assessment in England where it is readily acknowledged that paper- based assessments cannot assess essential process skills, such as scientific inquiry (Roberts & Gott, 2006).

3.6 Summary

This chapter has discussed how computer-based assessments have developed over the last 40 years and how there are still comparability and equivalence issues when they are considered alongside equivalent paper versions. The potential for education and assessment through the use of computers has been described, together with the regulatory issues that inhibit this movement.

The research study reported in this dissertation sets out to explore different modal science assessment types set in the context of the national curriculum. The next chapter describes how science education and its assessment have developed over time and how computer-based modes may enhance the construct validity of science assessment.

Chapter 4

Science Assessment

4.1 Introduction

The nature of science education and its assessment has been a contentious issue in England over the past 150 years, and many of the issues that were contentious then, continue to be so now. This chapter sets out to discuss some of the central issues surrounding science education and assessment, in particular focussing on the place of investigative science. The role of computer based simulations to teach and assess scientific enquiry skills will be explored from theoretical and practical perspectives.

4.2 A little bit of history about science education in England

Many science education writers and commentators (eg Layton, 1973; Jenkins, 2007) have drawn parallels between the social, political and economic pressures that initiated science into educational curricula in the 19th century, and similar pressures still operating today in terms of:

- What factors influence the science curriculum?
- Who designs it?
- Who is it for?
- How should it be assessed?

The 19th century is a good place to start to look at the development of science education in England, the first half of the century demonstrating a status quo in terms of education generally, and then the second half leading the way to the revolution of science education and from then on the continual conflicts of its place and purposes.

Any available education up to the mid 19th century was the preserve of the elite; dominated by the education of noblemen and gentlemen. Latin and Greek formed the sum of the curriculum, a platoic education based on the implicit assumption that education was linked to the symbolic control of society, and not linked in any way to production (Ross, 1999). The classics were so vocationally useless that they were a badge of honour or a symbol of a gentleman's education, as they by definition did not have to work for a living (Lawson and Silver, 1973).

It was not that science was not of interest to anyone. There was lively interest, debate and research going on in universities, but it had little or no impact on the schools that fed it. This was due to the fact that entry requirements for university education were entirely classically based and therefore the inertia of universities was a contributory factor to the lack of any science education development.

However, the 19th century was a period of radical change in society. Private and non-conformist academies emerged which taught practical experimental science, and these institutions gave alternative routes to university and to the professions. They encouraged a vocational climate in which the Industrial Revolution was fostered and they trained some of the workers who pioneered it.

Key drivers that sparked interest in science education included:

- The publication of Darwins 'Origin of the species' in 1859;
- The emergence of industries which required some technical knowledge and understanding;
- The public views on scientific education by eminent scientists of the day,

These scientists included Lyon Playfair, T.H. Huxley, Richard Dawes and Herbert Spencer. Playfair was a leading chemist of the day, who later became Secretary and Inspector of Science and Art. After a visit to The Great Exhibition in Paris in 1851, he expressed a view on the lack of science education in England:

... *we English are weak. Philosophy we have in abundance. Manual skills we possess abundantly. But we have failed to bridge the interval between the two. On the contrary, there is a dead wall separating our men of theory from our men of practice (*Playfair, a leading 19th scientist, quoted in Green, 1999, p56).

Later, in 1867, after he had visited the Industrial Exhibition in Paris in 1867, he again expressed concern that English manufacturing superiority was in decline compared to other nations, attributing it to the lack of available technical and scientific education.

This body of scientists became known as the 'scientific movement' (Evans, 1985). They advocated the science of common things, a plea for realism and a utilitarian view of science education within society. They publically challenged the general neglect in science education in England and endorsed the view that a sound foundation in science was essential in Elementary schools in order for science to flourish. The introduction of science in the curriculum offered in public and grammar schools can be partially credited to this group. On the other hand, there were other scientific factions that lobbied for science to be taught without regard to its applications. Robert Hunt, from the influential Governmental School of Mines, argued that the study of science for its own sake was an exercise which tended 'to the refinement and elevation of every human feeling', and that the emphasis on science as useful knowledge would prove harmful to the progress of science (Layton, 1973, p.136).

As the debate concerning science education progressed through the late 19th century, the Devonshire Commission of 1872 laid out some guiding principles:

*The true teaching of science consists, not merely in imparting the facts of science, but in habituating the pupil to observe…to reason… and to check… by further observation and experiment. It may be doubted whether any other educational study offers the same advantages for developing and training the mental faculties (*Devonshire Commission (sixth report, 1875) quoted in McClure, 1986, p 108).

Therefore as the 19th century drew to a close there was recognition not only of the intellectual and rigour that science demanded, but also its usefulness; in particular its capacity to develop the power of observation (Evans, 1985). However, even though the initiation of science education had much to thank the scientific movement, the assessment regime largely undermined it.

In order to assess and 'measure' attainment, and ensure standards were achieved and maintained, a Payment by Results system operated by which schools funding and teachers payment were directly linked to pupil rote learning, memory and recall to external inspectors who visited schools. The seeds of the philosophical conflicts between the theory and practice of science education and assessment had been firmly sown.

This science curriculum and assessment history could be followed through various stages of the 20th Century, with similar issues emerging; concern about the nature of science education offered, differences in provision across the population and the effect assessment had on the curriculum. I will therefore move forward to the late 20th Century to discuss how these issues manifest themselves now and how on-screen assessments may contribute to more valid forms of science education.

4.3 Back to the Future: Science Education in the National Curriculum

If we now fast forward to the issues surrounding the late 20th and early 21st century science education and assessment, the concerns are redolent of the 19th century issues concerning the place and purpose of science education. The position of science within the curriculum is secure and no longer peripheral; and is indeed core in terms of educational provision and governmental policy, however the key questions surrounding who influences and designs science education, who is it for and how it should be assessed remain central to academic, educational and public debate.

In order to address issues of patchy provision and quality of science education throughout England (APU, 1984), the National Curriculum, initiated in 1989, had the intention to provide all 5-16 year olds with a broad and balanced entitlement science curriculum, including scientific enquiry and content elements. Post-14 courses, delivered mainly through GCSE, already had assessment regimes to certify attainment. This was then allied by national key stage tests in science at key stages 2 and 3 (11 and 14 year olds respectively) and through teacher assessment at KS1 (7 year olds).

This initiative was a major governmental initiative, and was widely supported by the science community as a boost to the status of science in schools and as a potential remedy to the falling numbers of students studying sciences post 16.

The 'importance of science' statement from the National Curriculum documentation was laudable and set out expectations about the nature and purpose of science education:

> *Science stimulates and excites pupil's curiosity about phenomena and events in the world around them. It also satisfies this curiosity with knowledge. Because science links direct practical experience with ideas, it can engage learners at many levels. Scientific method is about developing and evaluating explanations through experimental evidence and modelling. This is a spur to critical and creative thought. Through science, pupils understand how major scientific ideas contribute to technological change- impacting on industry, business and medicine and improving quality of life. Pupils recognise the cultural significance of science and trace its worldwide development. They learn to question and discuss science-related issues that may affect their lives, the direction of society and the future of the world (*The National Curriculum for Science, p15)

However, there have been unfortunate consequences to the 'Science for All' agenda.

- Even though the three separate sciences were subsumed under the mantle of ‘science’, there were considerable ‘chunks’ of biology, chemistry and physics to be covered, often at a reasonably high level of demand. This has resulted in the active disengagement in their attitude and interest towards science of many pupils (Leach et al, 2001)
- The inclusion of considerable content was underpinned by the notion that it would be a good thing to have a comprehensive foundation of science knowledge and understanding, but many educators have subsequently asked of this curriculum coverage ‘what is this coverage for and who is it for?’. Osborne et al (2000) described most pupils as *consumers* and *users* of science, rather than *producers* of scientific knowledge. They were not going to go to university to study science, or go into scientific careers, yet their curriculum coverage was seemingly written for that intention. Indeed, throughout the 20 years of national curriculum science, progression onto science A level courses did not rise, nor onto physical science degrees.
- As with the analogy to 19th century science, the inhibitor to the nature and purpose of science education as laid out in the statement above, lay in the assessment methods employed, particularly with regard to scientific enquiry.

The form and style of assessment with regard to scientific enquiry differs between national key stage tests (11 and 14 year olds) and KS4 qualifications such as GCSE.

In the KS2 and KS3 science national curriculum tests, scientific enquiry has had inconsistent attention. The early versions of key stage tests did not include assessment of scientific enquiry skills at all. In addition to the national tests themselves, there was also a ‘Teacher Assessment’ component to the national curriculum which was designed to take account of practical elements of science. However, as performance in the tests and national and local league tables became dominant measures of accountability, only those elements of science assessed through the summative tests became embedded in the custom and practice of the teaching of science (eg. Green and Nickson, 1997; Black and Wiliam, 1998).

The concern is that the emphasis on the use of test results for accountability purposes may diminish the role of Teacher Assessment to a point at which the full programmes of study are not being adequately assessed (Stobart, 1999, p1).

This effectively came to pass, not just in the assessment of science, but in the curriculum offered to pupils. The assessment system effectively undermined investigational approaches to science because the teacher assessed Sc1 levels were seen as having a lower status and importance than the content tested in the standard tests (Watson, 1999).

This issue has been explored in great detail by commentators (see for example Wiliam, 2000a, 2001a, 2001b; Gipps,1994) who have argued that summative assessments, instead of being a sample of a construct, actually becomes the construct, and therefore teaching to a test for accountability purposes results in the loss of meaning and any real educational validity.

Ofsted reports have continually commented on the lack of breadth in the teaching of science, and in particular, the lack of attention paid to the development of transferable enquiry and process skills (Ofsted, 1999-2006).

Even when aspects of scientific enquiry were included in national curriculum tests,(from 1996) the emphasis was perceived to be on piecemeal approaches to process skills as well as the content, almost 'pub quiz' science (Sturman, 2003). Allied to this, Osborne et al(2000) suggested that when Blooms Taxomony (Bloom,1956) is applied, most test items were based on low level cognitive skills, usually involving recall and occasional application, but rarely delving into comprehension, explanation, evaluation or application to novel contexts.

The situation at KS4 differed in as much as the national curriculum, as transmitted through GCSE qualifications included a proportion of the assessment given to the measurement of students 'doing' science. However, this masked the reality of scientific enquiry assessment, which was dominated by written Investigation reports by students (House of Commons, 2002; Gott and Duggan, 2002) and the routinized approaches to the development and assessment of process and enquiry skills (see for example Bryce & Roberstson, 1988; Millar & Driver, 1988).

The place and justification of investigation and enquiry in science education had been discussed and established going back to the foundations laid down by Dewey (1910), and had carried on through the century, for example Bruner (1960), Kuhn (1993), and right up to Newton et al (1999). However somehow, the development of science education and an entitlement for all became distorted by the nationalisation of the curriculum and assessment regimes, and the resulting narrow interpretation of the subject by many schools.

Leading scientists of the day had significant influence regarding the nature of science education in the mid 19th century. By the end of the 20th century, this influence had become state controlled with the consequences outlined in Chapter 2.

4.4 Science for the 21st Century

The growing debate over the purpose, nature and assessment of science culminated at the start of the 21st century with concerted desire for change, supported by government and influential bodies, for example QCA and Ofsted, and the re-emergence of input from academia and scientists. Reviews such as Millar and Osborne (1999) and Osborne et al (2000) emphasised that the science curriculum and assessment did not serve the needs of many students, who will be potential *users* and *consumers* of science, rather than potential *producers* of scientific knowledge. They recommended that problem-solving, scientific literacy and the ability to critique information and ideas should be a substantial feature of KS4 science courses. Millar (1996) coined the notion that the guiding principle for science education should be '*do less, but do it better'.* Duggan et al (1994) suggested that in order for pupils to develop procedural knowledge, concepts of the nature of evidence and an understanding of the nature and purpose of scientific investigations should be firmly established as early as possible.

It is evident from the drivers for curriculum and assessment change that scientific investigational work is only one strand of scientific enquiry (Osborne et al, 2000) and even when investigations are used, they should emphasise the inherent uncertainty of science, rather than promote the notion that science, and particularly the scientific method is only about proof and confirmation (eg. Solomon, 1999; Ravetz, 1997;). Jenkins (2007) suggests that adherence to 'scientific method' actually misrepresents science education, as it bears little relation to the diverse nature of modern scientific disciplines (eg. molecular biology, astrophysics, bioinformatics), and the ways in which scientists actually work; as Einstein phrased it, often *loose opportunism* (quoted in McNally, 1999, p10).

This research study is focussed in part on the use of differing stimuli, response types and interactive simulations to support the teaching and assessment of science. The roots of the approach of using integrated, holistic realistic scenarios lie in a constructivist view of learning, and its supporting assessment. Many areas of science content and enquiry require students to use models to develop meaning and conceptual understanding and much scientific theory is dominated by the use of abstract theory to represent and explain natural phenomena (eg magnetic fields, electron orbitals). The problem is that much of this modelling is often counter-intuitive and even unnatural in its nature (Wolpert, 1992), and therefore consequently challenging.

On the other hand, students do not go into science classrooms without their own ideas about how the world around them operates; they do not passively learn and record information (Osborne & Wittrock, 1983). They have their own personal experiences, ideas and constructs of how the natural and physical world works and unless they are allowed to express these ideas and experience phenomena which challenge their inherent beliefs, deep learning and understanding will not be achieved (eg, Driver, 1983; Driver and Bell, 1986).

Cognitive psychology as applied to assessment and testing interprets this view in a complementary fashion. If a particular skill or construct is being assessed, the context used needs to be as relevant and meaningful as possible, otherwise there is a danger that unintended and ideosyncratic interpretations by pupils will result in construct invalidity (Pollitt et al, 1985; 1999).

4.5 The use of Computer Simulations

Interactive computer simulations in teaching, learning and assessment have been used over a long period of time in professional settings, particularly associated with technical or problem solving expertise (Akpan, 2001). There is evidence of a number of educational and logistical advantages. These include experiencing learning goals beyond traditional instruction methods (Thomas and Hooper, 1991), facilitating conceptual development in ways not possible by other means (Andre and Haselhuhn, 1995) and identifying relationships between components in a system and controlling the system (Gagne et al, 1981). These process skills can also be largely taught and assessed through active participation, however, simulations have a number of distinct advantages; they always work, unlike many classroom experiments, they can be used to experiment and investigate situations that are too expensive, dangerous or logistically difficult to set up when required. One of the major advantages is that they take far less time to set up and run.

A meta-study and analysis of 30 military training programmes that used simulations by Oslansky and String(1979), showed that the students achieved equal or better attainment using simulations rather than hands on methods, and the courses took 30% less time. This factor alone is significant. One of the criticisms of school investigative work is the issue of available curriculum time. Complete practical investigations can take four hours to set up and run in schools (Roberts and Gott, 2006). This therefore limits the number and complexity of investigations that can be accommodated into available curriculum or assessment time. Simulations therefore can overcome these time obstacles, and yet still enable students to actively engage with investigative processes, analysis and evaluation.

Kubicek (2005) argues that the use of computer based enquiry can actively enliven science, and if used creatively, can relieve the stagnation that school investigative work often leads to. Interactive programmes demand the active participation of the student as investigator (Tapscott, 1996). They also allow many different types of enquiry, from the active manipulation of variables within a system to the use of articles and data to demonstrate how science is evaluated and communicated. Development of conceptual understanding through modelling is a significant feature of computer simulators, allowing students to create multiple variable environments, test, run and discuss them. These approaches can therefore support a constructivist view of learning as described by Driver et al (1996).

The last issue to address is one of construct validity. Is using an interactive simulator a proxy for practical science, either in teaching, learning or assessment?

While most of the literature concerning the educational use and value of computer science simulators is positive, there are alternative views about their construct validity. Schrok (1984) called simulators *counterfeit science*, in that they isolate students from real-world experience. Bross (1986) expressed the view that simulations are not scientific because they imitate nature with programming and graphics and not from natural laws. Therefore students become computer literate, but science illiterate. Bross's view that simulations are no substitute for the real thing and carry no weight compared with hands-on laboratory demonstrations that are live, captivating and authentic, encapsulate the fear that simulators might displace essential experiences and development of science process skills.

In the main, however, the prevalent view seems to be one of addition and supplement to practical scientific enquiry, rather than its replacement (see for example Murphy, 1996; Kubicek, 2005; Akpan, 2001).

Simulators can allow experimentation to remain authentic while eliminating the tedium and errors made in gathering results (MacKenkie, 1988), and at the same time they can engage and motivate students in ways that traditional methods often fail (Hogarth et al, 2005).

4.6 Summary

This chapter has discussed how the place and position of science education and its assessment has been problematic over the last 150 years. Although science is a core entitlement in schools in England, the content of the science curriculum and how it should be assessed continues to be a challenge and contested by key stakeholders. The potential to address and improve the construct validity of science assessment through innovative computer-based assessments has been described and discussed.

This research study sets out to explore how paper and onscreen science tests and investigations compare and contrast with each other in the valid and reliable assessment of high stakes assessments in science. The next chapter describes the methodology underpinning this research.

Chapter 5

Methodology

5.1 The nature of my research question

My research is a study of the appropriateness of on-screen science assessment materials compared to paper-based versions, and how any potential change in assessment might affect assessment practices. 'Appropriateness' in this context is conceptualised in terms of comparing the reliability and validity of the form and performance of science assessments presented in on-screen and paper-based modes. As I will describe, some of this conceptualisation will take the form of empirical research; other parts through naturalistic enquiry.

It is therefore the pursuit of evidence concerning reliability and validity that inform my methodological decisions. Critically analysed, the empirical and naturalistic strands will provide complementary and triangulated evidence to address my research question.

As I have described in Chapter 2, educational measurement reliability is most easily described as consistency; that if a group of students took the same test twice, their scores and rank orders would remain unchanged (Nuttall, 1972). There are a number of threats to assessment reliability, some random and others systematic. My methodology has the primary concern of measuring the internal consistency reliability of equivalent paper-based and on-screen science tests using classical test theory, the calculation of Cronbach's alpha in each case and in addition, the use of Rasch latent trait modelling. These measures are statistically estimated, and therefore this aspect of reliability will be attained through a quantitative methodological approach.

There are other aspects of reliability that are not statistically measurable, and are often subsumed under the general banner of validity:

- attitudes to different assessments by students
- the content, form and outcomes of differing assessments

For these aspects, the most appropriate research approach is qualitative, using questionnaires and interviews with both students and teachers.

One of my research aims is to find out what students and teachers think of on-screen assessments in terms of style, content and appropriateness. This centres on the perceptions, understanding and experiences of students and teachers. The approach taken to address this question lies within an interpretative framework, one that favours interview, observation and questionnaire as research procedures.

Interviews and questionnaires are flexible methods that can be used by researchers whose philosophies are embedded in any of the research paradigms. For example, a highly structured survey questionnaire or interview can result in quantitative outcomes that can be tested statistically for significance. At the other end of the continuum, the use of open-ended interviews and questionnaires can impose little or no structure at all. My questionnaires and interviews are semi-structured in this sense.

Consideration of the face, construct and concurrent validity of the two equivalent tests in different modes are component elements of the comparative nature of much of my methodological approaches, and will provide the basis for a critical evaluation on the possible dependability of using on-screen tests in high stakes assessments; dependability being the intersection of validity and reliability (Gipps, 1994).

‘Positivism is dead. By now it has gone off and is beginning to smell’ (Byrne, 1998 cited in Robson, 2002, p26). This would appear to be a premature obituary with regard to my area of educational research. My research is not exclusively positivist, however positivism will provide a significant element of my thesis. As I will describe, my methodology comes under a mixed methodology banner, and I have avoided being dogmatic about methodological paradigms, philosophies or approaches.

5.2 Triangulation

Cresswell (1994) uses pragmatic reasons to suggest the use of a single paradigm in research studies, including the extensive time and expertise required to operate combined paradigm approaches and the potential scope and size of such a study. However, if a mixed methodology of both quantitative and qualitative approaches is appropriate and feasible for a particular study, Cresswell (1994) also acknowledges that it should be followed, and indeed suggests that mixed methodologies can be highly compatible and complementary.

Denzin (1978) used the term *triangulation* to argue for the use of mixed methodologies. This is based on the assumption that bias from one data source or method can be neutralised when used in conjunction with other data sources and methods (Jick, 1979).

Greene et al (1989) used triangulation as one of five reasons to favour the use of mixed methodologies:

- triangulation in the classic sense of seeking convergence of results
- complementary, in that overlapping and different facets of a phenomenon may emerge
- developmentally, wherein the first method is used sequentially to help inform the second method
- initiation, where contradictions and fresh perspectives emerge
- expansion, wherein the mixed methods add scope and breadth to a study.

In my research study, all of the above reasons are pertinent. There may be concensus between the empirical data and the qualitative evidence, in which case convergence is established. However, there may be differences and conflicts between the performance data of assessments in differing modes and the preferences expressed by students and teachers. Any emerging contradictions need to be explored, and thus different strands of evidence provide analytical, discursive and developmental opportunities.

In summary, my research can therefore be characterised as a mixed methodological approach where the outcomes of classical test theory and statistical modelling will empirically compare the whole test and item level performance of assessments taken in different modes. At the same time, this evidence will be compared to both student and teacher views on the appropriateness of on-screen science assessments compared with paper-based versions. The next sections of this chapter describe and justify the methodological approaches I have considered and utilised.

5.3 Quantitative Methodology

This section discusses why and how empirical methodological approaches have been incorporated into my research in order to address my research question.

Evaluation of how ‘good’ or ‘bad’ either items or tests are is a challenging task, and cannot be made solely on intuition, guessing or custom (Sax, 1989). Therefore psychometric tools have been developed to set down common parameters of comparison and indicators of the effectiveness and quality of assessments, and also to regulate and standardise them (Kehoe & Jerard, 1995).

Item and test analyses are methods of evaluating the quality of tests or examinations by looking at their constituent parts (items) and their performance as a whole (Thompson & Levitov, 1985). My study is focused on the comparative analysis of paper-based and on-screen science assessments, and therefore statistical measures need to be applied in the tests in both modes and then compared.

My quantitative methodology uses three forms of item and test analyses; classical test theory, Cronbach’s alpha co-efficient and Rasch modelling. The selection and justification of these statistical measures will be described and discussed in this section.

5.3.1 Classical Test Theory

This method provides the simplest and most practical way of dealing with reliability issues within tests and is the most commonly applied statistical tool applied to summative assessments (Bartram, 1990). Classical test theory dates back to the work of Charles Spearman in the early 20th Century, and is usually represented by the following formula:

$$X = T + E$$

Where:

X is the observed score (the actual measurement obtained)

T is the true score (what the measurement would be if there were no error)

E is the error score (the influence of error on the measurement, also known as measurement error)

Wiliam (1993) considers classical test theory as an attempt to capture the idea of 'signal-to-noise ratio' for assessments. This is based on the assumption that an individual's score contains error (noise) which can be decreased but never totally eliminated. It also assumes that the error is random and normally distributed. This issue is discussed in more detail in Chapter 2. Another point to bear in mind is that the true score does not mean a true measurement of ability, it is just a supposed average score that an individual would achieve over repeated taking of the same or very similar test.

Using this simple equation it is clear that when errors are small in comparison with the actual scores, a relatively high reliability is achieved, and when the errors are large in comparison with the actual scores, there is low reliability (Wiliam, 2001a).

Classical test theory concentrates on two key statistics within assessments; item facility and item discrimination.

5.3.2 Item Facility

This is essentially a measure of difficulty of an item; a high facility indicating an easy item and a low facility indicating a difficult item. This is given by the formula:

$$\text{Fac}(X) = \frac{X}{X\text{max}}$$

Where Fac(X) = the facility value of question X

X = the mean mark obtained by all candidates attempting question X

Xmax = the maximum mark available on the question

Another simple method to work out facility values of single mark questions is to divide the number of students answering an item correctly by the total number of students answering the question (Crocker & Algina, 1986). Thus an item answered correctly by 85% of examinees would have a facility value of 0.85.

It is important to note that facility value is a behavioural measure, and not an absolute measure of difficulty (Thorndike et al, 1991). This means that a facility measure in a test simply measures the comparative performance of items by a particular group of students. Facility values are important however, to design tests that differentiate across the ability range of the test. A test developer would normally want a test to contain questions that has a range of facilities, to ensure accessibility at one end, but also differentiation of outcome for students of different abilities. In general, in National Curriculum testing, a facility score of approximately 0.6 at the target level is considered to be desirable. This can only be achieved however, by pre-testing items and tests.

5.3.3 Item Discrimination

If a test as a whole, and the items within that test are measuring the same construct, then it would be expected that students who do well on individual items would do well on the test as a whole and vice versa. This correlation between performance on an individual item and performance on the test as a whole is called item discrimination (Aiken and Groth-Marnet, 2006). The item discrimination index or discrimination coefficients can be used to measure this.

The Item discrimination index (D) is used to compute a very simple measure of the discriminating power of a test item. The top and bottom 27% of scores are collected. D is the number of students in the top 27% who answered the item correctly minus the number of students in the bottom 27% who answered the item correctly, divided by the number of students in the larger of the two groups (Wood, 1960). The two ends of ability are used to maximise the differences over the normal distribution, while providing sufficient numbers for reliable analysis (Wiersma & Jurs, 1990).

Whilst the Item level index is an effective measure of discrimination, it is restrictive in terms of the sample of students used. I want to use a method that includes the data from all students, and also one that replicates the methodology used in high stakes assessments in the UK.

Discrimination coefficients differ from the item discrimination index in that they calculate discrimination values for all students taking a test. Most summative assessments taken in the UK use the coefficients for this reason.

Biserial correlation coefficients are usually used when there is a simple dichotomy of answer; right or wrong (Ebel & Frisbie, 1986). The items in my science tests were designed to capture and analyse responses in more than a binary fashion, and so this coefficient was not used.

The Point biserial correlation is used to look for correlation between an item facility and a total test score for a whole cohort taking a test. Henrysson (1971) suggests that this index is more informative about the predictive validity of the total test score than other measures, as it is a combined measure of item-criterion relationship and of difficulty level.

Point biserial correlations are also the most commonly used indexes in UK summative assessments, and so this confirmed my decision to use this measure of discrimination.

Discrimination correlation values for items can range from +1, where there is a perfect relationship between students scoring high marks on an item and their overall test score, to -1, where there is a perfect *inverse* relationship between students' scores on an item to their overall test score.

Discriminations should always be positive as this indicates that an item is measuring the same construct as the test (as it should). Negative discriminations indicate that, for example, a student scoring highly on an item scored very low on the test as a whole. This essentially means that the item was assessing something different to the rest of the test. There is very occasionally a case to be made for such an item, if there is a new aspect of a subject included within a curriculum or test specification (Lord and Novick, 1968). However, too many of these items in a test will interfere with classical test analyses.

The Measurement and Evaluation Center at the University of Texas (DIIA, 2003) offers guidelines, shown below in Table 1 for interpreting item discrimination values:

Table 1: Interpretation of Item Discrimination Value Guidelines

Discrimination	**Description**
0.40 or higher	Very good items
0.30 to 0.39	Good items
0.20 to 0.29	Fairly good items
0.19 or less	Poor items

Source: The Measurement and Evaluation Center, University of Texas

Massey (1995) carried out a large scale analysis of test and examination data in England, and concluded that items with discrimination values below 0.2 were weak and generally should be excluded from tests, whereas values above 0.4 were very good items. This is, of course, easier said than done, as exam items are not pre-tested and performance is only recorded retrospectively. He also pointed out the effect of extreme facility values on item discriminations. Easy questions at the start of a test for example usually have a very high facility scores. This will result in very low discrimination values, but be perfectly justifiable. Likewise, very difficult questions, which will be accessible to a small number of students, will result in very high discrimination values, but should not take up a large part of a test which needs to be accessible to a wide range of abilities.

In my research study I will not be able to pre-test items in order to construct tests with pre-established facility or discrimination values. However, my background and experience in the construction of science tests and examinations is helpful in writing and selecting items that will be accessible, yet cover a range of facilities and discriminations values.

5.3.4 Internal Reliability Measures

Reliability is the extent to which the measurements obtained in a test are consistent. As described previously, classical test theory suggests that test measurement is made up of a true score and an error component (Wiliam, 2001a). The reliability is therefore the amount of variation in the test scores; the higher the reliability, the lower the amount of error variance in the test. Therefore, the higher the reliability, the better the items and the test as a whole perform.

Various methods of estimating reliability have been used in UK assessments. In the early years of National Curriculum tests, test-retest (where the reliability is the correlation between a student's first and second score) and the use of parallel tests (where the reliability is the correlation between the scores on both tests) were used (see Schagan, 1993; Schagan & Hutchinson, 1994). These reliability measures were expensive and time-consuming to conduct, and were eventually replaced. I have therefore opted to use the measure which became the established method of estimating internal reliability in National Curriculum tests (Newton, 2007). This reliability measure was developed by Lee Cronbach (1951), and is commonly referred to as Cronbach's coefficient alpha or Cronbach's alpha. This is a measure of the amount of measurement error associated with a test score; the correlation between the test and all possible tests measuring the same construct (Massey, 1995).

Cronbach's alpha is scored from 0 to 1, where the higher the value, the more reliable the test is considered to be. The measure indicates how well items within a test are related to each other, and are therefore measuring the same construct.

The Measurement and Evaluation Center at the University of Texas (DIIA, 2003) offers the following guidelines shown below in Table 2 for the interpretation of Cronbach's alpha values:

Table 2: Interpretation of Cronbach's Alpha Value Guidelines

Reliability Measures, using Cronbach's Alpha	Interpretation
0.9 and above	Excellent reliability; at the level of the best standardised tests
0.8 to 0.9	Very good for a classroom test
0.7 to 0.8	Good for a classroom test; in the range of most. There are probably a few items which could be improved.
0.6 to 0.7	Somewhat low. This test needs to be supplemented by other measures (e.g. more tests) to determine grades. There are probably some items which could be improved
0.5 to 0.6	Suggests need for revision of test, unless it is quite short (ten or fewer items).
0.5 or below	The test definitely needs to be supplemented by other measures for grading.

Source: The Measurement and Evaluation Center, University of Texas

National Curriculum tests are one of the only high stakes assessments to have published Cronbach alpha values over the years in the UK. They generally generate values between 0.85 and 0.95 (Black & Wiliam, 2006), which provide fairly positive evidence of reliability. I will calculate Cronbach's alpha for my science assessments, and consider what evidence they provide with respect to internal reliability.

Interpretations of Cronbach's alpha are not always straightforward, as different styles of assessments are not necessarily suited to this methodology, and produce significantly different differing results (Nunnally & Bernstein, 1993). Extended writing or essays will always produce lower Cronbach's alpha measures than multiple choice questions. They may however, be preferred as a more construct valid form of assessment.

Cronbach's alpha measures are often combined with the standard deviation (SD) of marks from a test to calculate the standard error of measurement (SEM). This measure gives confidence intervals that marks attained on a test are normally distributed, showing how much variance of a true score there will be across all the observed scores. I will calculate the SEM for the computer and paper-based and tests and investigations, and compare this variance with any other variances found between assessments in different modes.

It is important to note that the use of Cronbach's alpha and the Standard Error of Measurement (SEM) measures the reliability of constructs within a test and the behaviour patterns of students. It does not account for any marking unreliability. I will deal with this aspect of my methodology later.

It might appear that the use of classical test theory and the calculation of test reliability would be sufficient for the quantitative element of my methodology. These measures apply more statistical measures than most public examinations in the UK. However, there are a few problems associated with the use of classical test theory. These include:

- The perceived ability of students is determined by the difficulty of a test. This means that if a test is difficult, facility scores will be low, and students will appear to be of low ability. Different tests may therefore be incomparable.
- The reliability of test scores does not remain constant across the ability range. This means that standard error may be different for different abilities across the normal distribution curve. This means it is difficult to measure the relative abilities of students (Lord, 1984).

5.3.5 Latent Trait Models

Latent trait modelling attempts to overcome these problems. It is based on the assumption that there is a relationship between the observable test performance of students and an underlying trait or ability (Hambleton and Cook, 1977). Item characteristic curves that latent trait models produce are therefore designed to be independent of the ability of a particular group, and this then means that measurements of students can be equated across test forms that are not parallel. This is called *invariance* of item and ability parameters.

There are two types of latent trait methodologies; Item Response theory and Rasch modelling. Either would produce item characteristic curves that I could use for my study. However, Item Response Theory (IRT) is often used in strictly hierarchical levels of difficulty of items, as found in a subject like mathematics or in tests where there is choice of items allowed by students. Science has more of a variance between perceived item difficulties in questions, and the tests developed contained no choice of items for students, and so I opted for the use of Rasch modelling.

Rasch, like IRT, seeks to demonstrate the ability of candidates in terms of their performance ability on the construct of an assessment. However, Rasch is a simpler and more manageable form of latent trait modelling as it uses only one parameter- difficulty of item, compared against one person parameter-ability (MacCann & Stanley, 2006). Rasch is useful in comparing non-equivalent groups or assessments in order to make useful equating comparisons or standard setting decisions year on year. Rasch is a model in the sense that it represents the structure which data should exhibit in order to obtain useful measurements; it therefore provides a criterion for successful measurement. A distinction between Rasch and other statistical models is that usually, statistical models are used to describe sets of data. In contrast, where the Rasch model is used, the objective is to obtain data that fits the model (Andrich, 2004).

For the purposes of my research, classical test theory, reliability measures and Rasch modelling are not being used to award levels or grades or maintain any given standard. They are being used simply as statistical tools to make useful comparisons on the performance of tests taken in the two different modes; paper-based and on-screen.

5.3.6 The control of other threats to reliability

As described earlier in this chapter, reliability error within assessments can be categorized as systematic or random (Brennan, 2001). Systematic causes of error concern the nature of the assessment, the constructs tested or not, and the assessment process itself. It might be argued that any given assessment does not test the breadth of the curriculum or all available constructs. These issues are not the concern of reliability measurement. Random causes of error are concerned with variance within and between assessments, and it is in this area that reliability measurement is focused. Random causes include:

- The luck of the draw in terms of the questions asked within a test
- Students may perform better or worse on different days or in different circumstances

- Different markers may give different marks for the same piece of work

(see Wiliam, 2001a; Black & Wiliam, 2002).

Some of these issues are the inevitable consequences of a 'One-Shot' testing culture (Wolf and Silver, 1993); particularly any potential difference in student performance on different days. My research model does not attempt to accommodate this variable. Likewise, my research is not primarily focused on the measurement of marker reliability; this would be a research study in its own right. However, I will briefly describe what is meant by marker reliability and explain how I minimised this variable in my research.

Marker reliability can be categorized as inter or intra reliability. Inter being how reliably different markers apply a mark-scheme, intra being how reliable and consistent each marker is in applying a mark scheme (Wolf and Silver, 1993).

In order for consideration of internal reliability measures to contribute towards a robust discussion of the appropriateness of on-screen assessments compared to paper-based versions, it will be useful to minimise the effects of marker error as much as possible. This will be achieved by:

- The use of automated marking technologies. Over 80% of the computerised tests are automatically marked by the computer. As correct and accepted answers to questions are pre-determined and programmed, there can be no inter and intra marker error, either in the marking itself or the transcription of marks

- All the paper-based tests and the open ended sections in the computer tests will be double marked. There are only two people involved in marking, and therefore after standardisation procedures (training of mark scheme and trial marking exercises), marking reliability should be fairly high. However, if scripts are marked independently by the two people, and any differences discussed and marking decisions agreed, marker reliability error is significantly reduced. The costs of double marking are deemed to be prohibitive in high stakes testing in the UK (Wiliam, 2000a). However, for the purposes of my research, the elimination of marker unreliability was useful in focusing the research outcomes to discussion of items within the tests.

Therefore, discounting the measurement of the second and third given causes of random error, it will be the first cause, the selection of questions within a test that my statistical analyses will focus on.

This section has discussed how and why decisions on appropriate empirical methodologies have been considered and implemented.

The next section will consider of appropriate naturalistic approaches to address my research question.

5.4 Qualitative methodology

As I have discussed earlier, my methodology contains quantitative and qualitative strands. The quantitative elements will arrive at comparative statistical measures of internal reliability. The qualitative strand will canvas views and interpretations of paper-based and on-screen assessments and use an interpretive paradigm.

The qualitative part of my research is exploratory and descriptive, as Robson (2002) elaborates:

Exploratory

- to find out what is happening
- to seek new insights
- to ask questions
- to assess phenomena in a new light
- usually, but not necessarily, qualitative

Descriptive

- to portray an accurate profile of persons, events or situations
- requires extensive previous knowledge of the situation etc. to be researched or described, so that you know appropriate aspects on which to gather information.
- may be qualitative and/or quantitative

These categories are not necessarily tightly bound and my enquiry contains many aspects of both. Given the large sample numbers in the quantitative element of my methodology (n=1000), I wanted to utilise the sample to the maximum in order to canvas views. Case studies would not have provided the representative range of opinion I required, whilst an effective survey strategy would. At the same time I wanted a degree of hybridisation (Robson, 2002) in that I was looking for insight and detailed feedback from a few participants (exploratory) to go alongside larger amounts of general factual and impressionistic information from a much larger sample (descriptive).

This model was operationalised in the following ways:

- all students who took the paper-based and on-screen assessments, completed questionnaires immediately afterwards, and a comparative (paper/on-screen) questionnaire following their last assessment
- A selection of students were interviewed in groups
- A selection of individual teacher were interviewed

Each of these three approaches will now be discussed.

5.4.1 Questionnaires

The purpose of the questionnaires was to collect factual and impressionistic information on the same scale as the assessments taken by students. In some respects, this was to address validity issues alongside those of measurement reliability; were the assessments considered appropriate and fit for purpose? The use of such large scale questionnaires would also elicit general outcomes to be explored further in the group and individual interviews.

Several defining features of the questionnaires required for my inquiry required consideration; their administration, structure and question types.

In the context of my enquiry, there were a number of administrative options available, including postal, self administered or group questionnaires (Cohen et al. 2007; Robson, 2002). Each of these has associated advantages and disadvantages.

Postal questionnaires can be more economical in terms of researcher time and effort, and although response rates are always a concern, they can be as least as good as other survey instruments if conducted with care and good planning (Hoinville and Jowell, 1979). Self-administered questionnaires can be taken with or without the researcher being present (Cohen et al, 2007). However, the presence of the researcher (particularly if a group takes a questionnaire together) can be helpful as data can be collected from many respondents simultaneously, and any queries or ambiguities can be clarified at the point of completion. Self-administered questionnaires without the researcher present can also be advantageous as they remove particular time pressures or any effect the presence of the researcher may have on the minds of the respondents.

As the questionnaires in my enquiry were completed by students after completing assessments, group self-administered questionnaires were used most of the time. On a few occasions, when assessments were taken remotely by schools without a researcher present, the class teacher acted in the position of the researcher, administered the questionnaires and sent them back by post.

The structure and question types in the questionnaires required careful consideration. A general rubric when considering these issues is the sample size (Oppenhiem, 1992). For reasons of time and effective coding procedures, the larger the sample size, the more structured, closed or numeric the questions should be. The sample size in my enquiry was 1000, and students were to be asked to answer a number of questionnaires, which necessitated a highly structured approach. At the same time, one of the purposes of the questionnaires was to elicit in-depth, rich responses in an unknown, exploratory area (Bailey, 1994). With this in mind, open ended text boxes were inserted alongside closed responses to allow students to make comments if they wished.

As the respondents were 14-15 year old students across a wide range of ability, and there was limited time for the questionnaires to be completed, accessibility and ease of completion were paramount in their design.

Short, closed questions are recommended and were used with these restrictions in mind (see Wilson and McLean, 1994; Edwards & Talbot, 1999; Blaxter et al, 1995). The language used was reading level assessed, which considers not only vocabulary, but also sentence structures.

As one purpose of the questionnaires was to provide comparisons between paper-based and on-screen tests, Likert scales were used on some questions. Likert scales, developed by Rensis Likert in the 1930s, provide a range of responses to a given question or statement, usually strongly agreeing or disagreeing (Oppenheim, 1992). These are useful as they can generate numeric comparisons, but also differentiated responses (Cohen et al, 2007). Likert scales often use a 5 or 7 point scale. There is however, a tendency for many respondents to opt for the mid-points on these scales. With this in mind, I opted for a 4 point scale, to prevent respondents sitting on the fence.

Even though the plan was for questionnaires to be administered immediately after assessments were taken, asking students (and even adults) about the details of individual questions afterwards often results in patchy recall in areas of interest (Ericsson & Simon, 1984).

Calderhead (1981) and Lyle (2003) discuss the use of stimulated recall in order to overcome this problem. This strategy involves playing back video or audio sequences of an activity to participants to reflect on their practice. This allows the participant to relive their experiences and verbalise their thoughts. My use of this approach was to include screen shots of particular questions of interest in the questionnaires to aid recall. This increased the length of some of the questionnaires beyond what Edwards & Talbot (1994) and Blaxter et al (1995) recommend, but only in the provision of visual stimuli.

Chapter 6 will describe in more detail the structure of each of the five questionnaires designed for this research study. In summary, they were designed to be quick and easy for students to complete, using combinations of closed Likert scales, space for open responses and screen shots of particular question types to aid memory and recall.

5.4.2 Designing Interviews

As described throughout this section, there are a number of components to my methodology, including quantitative and qualitative methods. The quantitative statistical modelling would be complemented by analysis of questionnaires taken by the same students. All this data and information would then be further expanded by interviews with a sample of students and teachers.

Kerlinger (1973) wrote that interviews allows the gathering of information on peoples' knowledge, feelings and attitudes, beliefs and expectations, intentions and actions and reasons and explanations.

My interest in students and teachers' understandings, knowledge, beliefs, explanations, actions and expectations of the validity and comparison of assessment in different modes concur with Kerlinger’s ideas and therefore this form of evidence will be a valuable asset for my research. The interviews were conducted immediately after the tests were taken and before the students or teachers received any results from the trials. Therefore their views were concerned with the comparative face validity of the assessments; were the science assessments accessible, fair and engaging?

5.4.2.1 Structure of Interviews

Interview structure and the nature of the questions asked are closely related. Hook (1981) described interviews as varying from rigidly structured to very unstructured. A structured interview consists of a schedule of questions to which short answers would be required; an unstructured interview has more flexibility with questions emerging from the answers and comments of the respondent.

Closed, highly specific questions have the disadvantage that they invite superficiality of response, because participants are not permitted to expand upon their answers. However, commonalities between interviews do allow responses to be more easily compared statistically. Potter and Whetherall (1986) argue that highly structured questions require participants to reduce their views to gut-responses that they are not permitted or required to expand on or justify. They argue that a highly structured 'Yes or No' question essentially asks 'I know that there is a lot of discussion to be had, but, off the top of your head what do you think, Yes or No?' Their research also showed that when interviewees are asked to elaborate upon such initial responses they are likely to contradict them. They argue that structured questions do not collect the depth of data that can be gained from questions that allow participants to elaborate. This view concurred with my own as I wanted to add understanding and depth to initial views on questionnaires for students and gain fresh insight from teachers on their views on the validity and comparisons between paper-based and on-screen assessments, therefore my intention was to adopt an unstructured interview approach.

Lomax and McLeman (1984) state that unstructured interviews, circumscribed only by a general topic to be discussed, enable respondents to impose their interpretations on that topic while allowing the interviewer to ensure that all dimensions of the issue are considered. One type of unstructured interview is the focused interview (Merton and Kendall, 1946) in which the topics of interest have already been decided upon. I will expand on the essential features of this approach and why it best suited my research intentions.

5.4.2.2. The focused Interview

Merton and Kendall (1946) describe four features of the focused interview:

1. the researcher has provisionally analysed significant elements of the issue, e.g. through initial surveys as information collecting exercises. In my case this was positivist data;

2. a sample is selected who have been involved in a particular situation;

3. an interview guide is developed addressing the major areas of inquiry and hypotheses suggesting features of important data to collect;

4. interviews are carried out, with a value on subjectivity, to test the hypotheses and be open to fresh hypotheses.

These four features were useful guidelines for me in the planning phase for the interviews. I am familiar with issues of reliability, validity and comparability between paper-based and on-screen assessments through literature reviews and clearly these would be the dominant issues discussed in the interviews. My experiences as a teacher also impacted on my areas of interest, including the students' ease of engagement with an unfamiliar mode of assessment (on-screen) and the teachers' views on assessment in general. Early response data from questionnaires and a piloting exercise in advance of the main research informed the general direction that the focussed interviews would take.

The choice of participants was determined by students' and teachers' willingness to be interviewed and BERA ethical guidelines (2004) were adhered to in all phases of the research. Legal requirements for interviewing school children vary between Local Authorities, however best practice recommended by governmental agencies advise that children are interviewed in small groups rather than individually. Apart from the necessity to protect the researcher and children from any issues of impropriety, group interviews with schoolchildren have distinct advantages in terms of reducing stress, anxiety and promoting confidence and active group discussion.

Thirdly, Merton and Kendall suggest that an 'interview guide' is developed. An interview guide was defined by Hook (1981) as 'a loose collection of topics and possible questions'. An interview guide was developed in the light of prior knowledge and understanding of pertinent issues, information gathered from the pilot study, and emerging data from questionnaires.

The interview guides for students and teachers focused on differing aspects of my research questions. Student interviews focused on the comparative face and construct validity and the 'usability' of the assessments, whilst the teacher interviews focused on the authenticity, appropriateness and fitness for purpose of assessments in differing modes. Student and teacher interview guides included introductory and ethical considerations, followed by areas of interview focus. Chapter 6 will describe in more detail the structure of the interviews.

Although the two interview guides had different foci, they had the same intention of being flexible, subjective yet probing. Kerlinger (1973) and Cannell & Kahn (1967) were helpful in the consideration of questions. They suggest the following criteria:

- Questions are related to the research problem and objectives
- The type and language of the questions is right and appropriate
- The questions are clear and unambiguous
- The questions are not leading
- The questions demand knowledge and information that is within the respondents frame of reference
- There is one idea per question
- Questions occur in a logical sequence
- There are open and closed questions
- The researcher is sensitive to personal or delicate material that the respondent may resist
- The researcher considers the pressure to give a socially acceptable answer
- The researcher pilots the interview

As I have described, these considerations were used in the planning and implementation of the interview phase of my methodology.

5.5 Summary

In summary, in this chapter I have explored research approaches, both conceptually and in the context of my research questions. Empirical and naturalistic approaches have been considered, discussed and selected in order to gather the different forms of data and evidence that are necessary for my research study.

In the next chapter, I will describe in more detail how this methodology was operationalised as a method.

Chapter 6

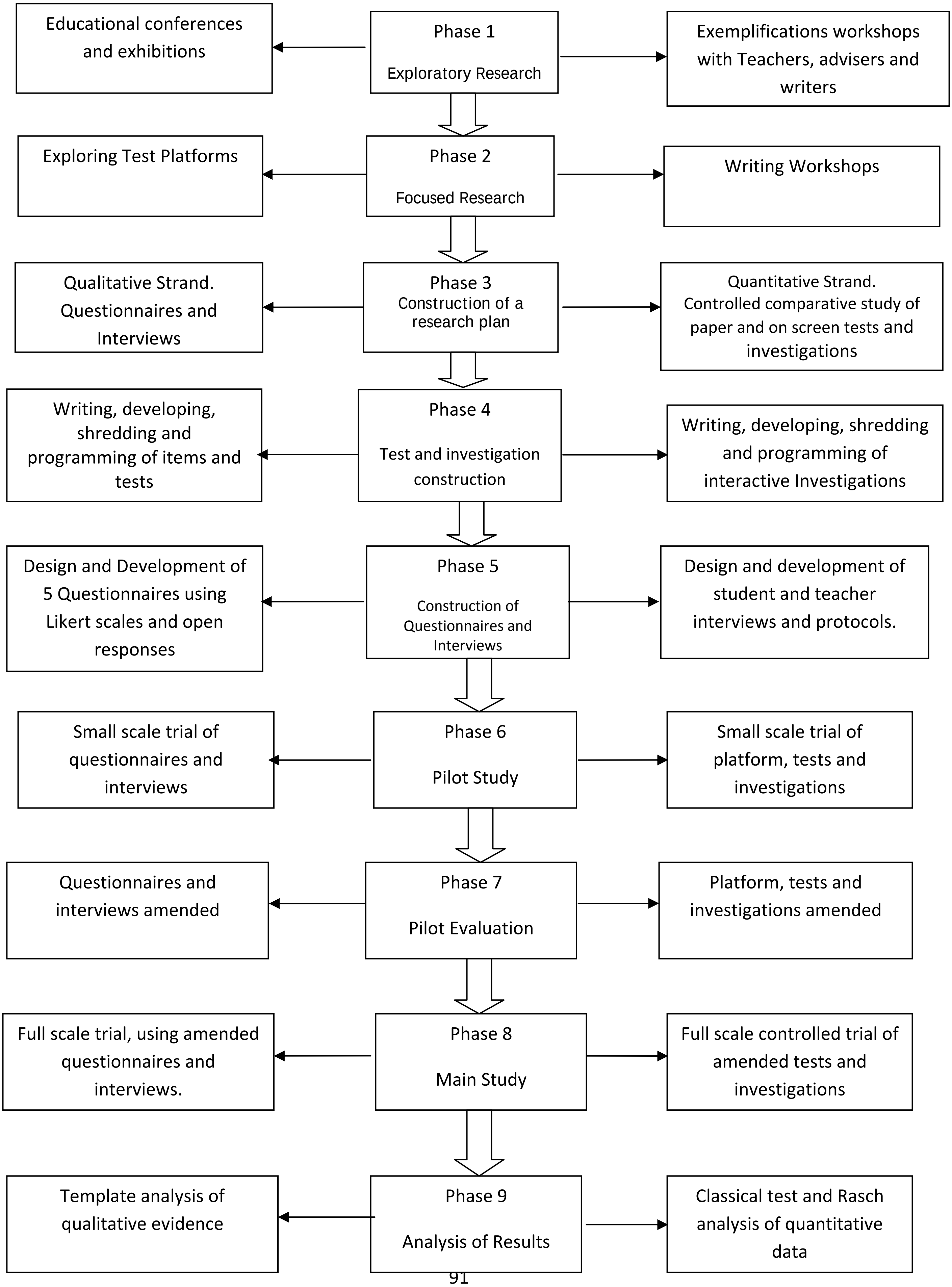
Method

6.1 Introduction

This chapter describes how the methodology was operationalised as a method, starting from early research into the design of on-screen assessments, moving onto the construction of on-screen and paper-based tests and investigations, setting up a controlled research plan, piloting the quantitative and qualitative research strands, and finally the collection of the live research evidence.

In the previous chapter, the rationale was explained for the collection of both quantitative and qualitative evidence. Rather than describe how these elements developed as separate strands of research, the nine stages of my method will be discussed, integrating quantitative and qualitative strands where appropriate. These nine stages are shown below in Figure 5 as a flow chart, and then described in more detail.

Figure 5: Research Plan Flow Chart


Educational conferences and exhibitions
Phase 1 Exploratory Research
Exemplifications workshops with Teachers, advisers and writers
Exploring Test Platforms
Phase 2 Focused Research
Writing Workshops
Qualitative Strand. Questionnaires and Interviews
Phase 3 Construction of a research plan
Quantitative Strand. Controlled comparative study of paper and on screen tests and investigations
Writing, developing, shredding and programming of items and tests
Phase 4 Test and investigation construction
Writing, developing, shredding and programming of interactive Investigations
Design and Development of 5 Questionnaires using Likert scales and open responses
Phase 5 Construction of Questionnaires and Interviews
Design and development of student and teacher interviews and protocols.
Small scale trial of questionnaires and interviews
Phase 6 Pilot Study
Small scale trial of platform, tests and investigations
Questionnaires and interviews amended
Phase 7 Pilot Evaluation
Platform, tests and investigations amended
Full scale trial, using amended questionnaires and interviews.
Phase 8 Main Study
Full scale controlled trial of amended tests and investigations
Template analysis of qualitative evidence
Phase 9 Analysis of Results
Classical test and Rasch analysis of quantitative data

6.2 Phase1. Early Research

This phase had two objectives. The first was to find out what on-screen assessment packages were currently available; the second to consider what authentic on-screen science assessments might look like.

The first objective was realised by going to various educational conferences and exhibitions, particularly those showcasing new and emerging teaching, learning and assessment packages and programmes. They provided an excellent opportunity to observe, try out and discuss products with educational assessment providers, exam boards and governmental agencies.

Educational e-Assessment packages fall into two main categories. One consists of teaching and learning activities which do not attempt to empirically measure performance. Others use simple summative multiple choice questions to assess knowledge, and attribute levels or grades of attainment on the basis of a narrow range of marks and skills.

A significant part of my work at QCA was the development of both the KS3/KS4 science curricula and supporting assessments. The science national curriculum, the KS2/KS3 science national tests and GCSE's and GCE examinations had been through major changes in terms of re-appraising the role of science education in society. From 2003, scientific enquiry skills had re-emerged as essential features of the taught curricula and in the instruments of assessment. However, I saw no evidence of meaningful or authentic assessment of scientific enquiry in the e-Assessment research I had undertaken, and therefore highlighted this as an area I could focus on in terms of developing on-screen science assessments. Chapter 3 emphasised the lack of these forms of assessments.

The second objective at this initial research phase was to meet with a range of teachers, advisors and writers to share ideas about the possibilities that could be explored through the development of on-screen science assessments. Before I organised these meetings, I developed a draft model of how an aspect of scientific enquiry could be assessed in an on-screen environment. I wanted to address a number of issues at this early stage and used the draft model for exemplification purpose and to encourage dialogue. Contributors could either draw on the exemplification example and enhance and improve it, or they could reject it, using reasoned argument as to why a different approach would be better.

For exemplification purposes, I decided to look at the area of investigative science, and consider how a model of an interactive, authentic and contextual investigation could be approached. My aim was to put science into a context that KS3/KS4 students would find stimulating, interesting and relevant, engaging students into taking on the role of the investigator, not merely being the recipient of secondary data.

I used the context of a ‘stomp rocket’ for my exemplification piece. Stomp rockets are plastic rockets that can be fired considerable distances by stamping on an air-filled bag attached to them. The distance the rocket is fired depends on two variables, the force of the stamp and the angle of launch of the rocket. This investigation provided an excellent opportunity for authentic scientific experimentation, and also, stomp rockets are inexpensive summer or beach toys unconnected with a traditional view of school science. I videoed my daughter using a stomp rocket in a local park and then developed a storyboard of how this context could then be used for experimentation. The construction of a storyboard would be essential to translate my ideas to a programmer who would construct an on-screen working item. The storyboard template would also be the method of getting teachers and writers to develop their ideas for on-screen assessments.

My exemplification piece was also a vehicle for exploring what the current technology within an exam board could do. I wanted to explore the potential of incorporating the following aspects into on-screen assessment items:

- Videos
- Simulator models, so that pupils could choose variables, carry out their own investigations and collect, record, present, analyse and evaluate their own data
- Different forms of closed response and marking mechanisms
- Open response opportunities for pupils to communicate their scientific ideas and understanding

The ‘Stomp rocket’ assessment item consisted of a video of a student playing with a stomp rocket in a park to put the question into a context, and then moved on to a screen consisting of a stomp rocket simulator. Pupils could experiment with the simulator before making any decisions about what they were going to investigate. Once familiar with the simulator, they were given the option of which variable they were going to explore. They then had to systematically collect data, firing rockets using either different forces at a given angle, or a given force over a range of angles. Finally, they had to collate their collected data, construct a graph of their data, and analyse and evaluate their evidence. This storyboard can be found in Appendix A.

A programmer subsequently set about turning my storyboard into a working on-screen assessment item. With programming work initiated, and initial thoughts documented, I set out to meet the groups of interested stakeholders mentioned earlier.

I met with groups in London, Kent, Bedford, Cambridge and Doncaster. Participants at these meetings included people who had been involved with national curriculum science assessment, and others who had been vocal in their criticism of such science assessments. I also included people who had been recommended to me by local authorities as innovative and creative teachers who had expressed interest in getting involved with this research. I was also keen to include people who had no background in writing or developing national science assessments in order to gain new perspectives on assessment approaches.

There is never an easy substitute for the experience of item writing and development. In exam circles it is often referred to as a dark art, and there is no doubt that the writing of focussed assessment items is a lot harder than it looks. I was eager to discuss with these groups their views on how assessment items could be improved and the opportunities that on-screen assessment could take advantage of in terms of forms of stimuli and interactivity. The exemplification item I had written provided useful material as a stimulus for discussion, and these meetings confirmed to me where there might be assessment opportunities to explore in an on screen environment.

At this stage, my research was exploratory in nature. I was interested in moving into the next phase, where ideas about the potential of on-screen assessments in terms of their structure and construct coverage could be developed into a range of item types.

6.3 Phase 2: Writing Workshops and Testing Platforms

This phase also consisted of two elements. The first was the opportunity for any of the people I had met with to write ideas or develop on-screen items for possible inclusion in the research study. I provided item writing guidance and templates which were designed to provide a framework if required, however the templates were not compulsory if writers found that they inhibited creative ideas or the writing process. Simpler word templates were also made available. The on-screen writing template and initial writing guidance are found in Appendices B and C respectively. The number of people who actually wrote and developed ideas compared to the number of people involved in workshops and discussion was small. This came as no surprise. The art of focussed and concise item writing is challenging using paper-based systems. In many ways, on-screen writing is even harder as the size of the screen limits the use of words, instructions and information to a much greater extent than on paper. When students use paper-based examinations they often use information that is located somewhere on the page or the facing page. This is not possible in on-screen assessments and therefore the writing and development of items has to be very carefully designed. Apart from there being less available space on- screen compared to paper pages, there is evidence that students find scrolling between pages difficult and it can therefore inhibit student performance.

In my own test development experience, paper-based writing workshops usually result in significant writer drop-out, and so it proved with on-screen writing. Some of these participants still had a role to play later on in trialling phases. Those people who did write items usually developed an initial idea, sent it to me for feedback and then developed the idea into a working item using either the on-screen or word templates.

The second element of this phase was to revisit the programming of the stomp rocket item in order to evaluate the available programming skills and platform technologies available within my exam board. It became evident that my assessment ideas and items could not be provided or supported by the existing and available programming or operating platform. Existing on-screen tests were only available using simple multiple choice formats and did not use any of the aspects I listed earlier as essential elements in the creation of authentic, interesting and stimulating items. Progress in programming the stomp rocket item was slow, and it became clear that the current programming skills could not develop complex item types and the operating platform could not support interactive items, had limited methods of presenting and marking closed response questions and could not allow the collection of open responses to questions.

These problems would have to be solved for the research to progress. I was fortunate to meet a programmer who offered skills that could address the problems I had identified. The outcomes were two-fold. I had found a route where innovative items could be programmed and a customised platform built to allow students to take web-based on-line assessments, incorporating automatic and open response marking mechanisms.

6.4 Phase 3: Construction of a Research Plan

This phase was characterised by the creation of a focussed research plan. The initial phases 1 and 2 were largely interpretative and exploratory research. This type of approach could have continued in this way and my research route would have been more open-ended as a result. However, I was concerned that my research into on-screen assessments had tangible and practical outcomes in terms of contributing empirical evidence to towards a gradual national shift to providing assessments in on-screen modes rather than current paper-based versions.

Whilst QCA had expressed ambition to modernise the testing and examination systems in England and make use of 21st century technology, it was becoming clear that moving high stakes assessments into on-screen environments carries with it significant issues and risks. Notwithstanding technological reliability issues, assessment comparability issues were emerging. There is little or no evidence in this country to address questions such as:

1. If the same closed questions (largely multiple choice) are taken on-screen and on paper, would performance be the same?

2. If paper-based questions are adapted for on-screen in terms of altering stimulus or answering mechanisms, would performance differ and could it be assumed that they are assessing the same constructs?

3. If stimuli or the presentation of questions show significant differences between on-screen and paper-based versions (eg through the use of videos or pop-ups) would performance differ, and could the assessment itself be considered better in terms of fitness for purpose?

4. If interactive questions were to be offered on-screen, how would they compare with experimental, exploratory questions presented on paper; are they assessing the same constructs and could they be considered better in terms of fitness for purpose?

Therefore, with these four questions in mind, a controlled research plan was constructed, in order to gather quantitative and qualitative evidence to contribute towards an understanding of the reliability, validity and comparability issues between assessments in paper and on-screen modes.

The research plan consisted of a comparability study of the performance of on-screen assessments compared to paper-based assessments. There were two strands to this research; one strand comparing standard test questions in paper and on-screen modes; the second strand comparing investigational assessments in paper and on-screen modes. These strands would run concurrently; the aims of the research for both strands were to consider:

- whether the mode of science questions results in any difference in performance by students or measures of reliability?
- whether the on-screen science assessments are considered more authentic and fit for purpose than equivalent paper based versions?
- whether the modes of assessment appear to assess the same constructs?

6.4.1 Test Design

Two parallel equivalent versions of on-screen science tests were written, developed and programmed. The research rationale for two versions is explained on page 98. There were three forms of questions to be incorporated into the on-screen tests:

- Questions that were comparable in form and response to paper-based questions (apart from the response mode). These were automatically marked;
- Questions that differed from paper based questions in terms of type of stimuli provided (eg, video, models, pop-up information boxes) or in the form of response required by pupils (eg manipulation of on-screen icons or tools). These were also automatically marked;
- Questions that require open responses, these ranged from single words, short phrases to longer responses. These were not automatically marked; they were captured and marked by a science specialist.

It was not my intention to take a paper test and translate it to an on-screen format. I wanted to start from an on-screen perspective where possible, and then translate it to a paper format. Therefore, once programmed, paper-based versions of the on-screen assessments were then produced. These were as close a comparative match to the on-screen versions as possible, although the mode of interaction and response type was different.

6.4.2 Investigation Designs

Two parallel equivalent versions of on-screen science investigations were also written, developed and programmed. There were three forms of items incorporated into the investigations:

- Experimentation, using an interactive simulator, to include trialling and data collection;

- Questions that differed from paper based questions in terms of the form of response required by pupils (eg manipulation of on-screen tools).These were automatically marked;

- Questions that required open responses, ranging from decisions on the range of data collected, to explanatory text. These were not automatically marked. They were captured and marked by a science specialist.

Equivalent paper based assessment for each of the Investigations would also be produced. These were designed to be as close a comparative match to the on-screen versions as possible, although the mode of interaction and response types were different.

6.4.3 Research Model Rationale

Two equivalent versions of tests and investigations were necessary to allow a four parameter model to operate (as shown in Table 3). This enabled each group to take an equivalent test and investigation in each mode, which could then be compared, which is a minimum requirement in this type of research study

A representative sample (in terms of ability) of 1000 students was split into two cohorts, with each cohort taking a pair of equivalent tests and investigations. The four parameter model could also allow for students taking the tests in different orders:

Table 3: Four Parameter Research Model

Computer Based Test and Investigation (CBT and CBI)	**Paper-Based Test and Investigation(PBT and PBI)**
Test and Inv 1	Test and Inv 2
Test and Inv 2	Test and Inv 1

- 250 students will take CBT and CBI 1 first, and then PBT and PBI 2
- 250 students will take PBT and PBI 2 first, and then CBT and CBI 1
- 250 students will take CBT and CBI 2 first, and then PBT and PBI 1
- 250 students will take PBT and PBI 1 first, and then CBT and CBI 2

All students would complete a questionnaire on the on-screen and paper-based tests that they have taken and a comparative questionnaire. A sample of students would also be interviewed. Student questionnaires and interviews collected evidence on their attitudes to the different modes of assessment and their working practices in such assessment types. A sample of teachers was also interviewed using a similar set of questions in order to get their thoughts and views on the validity of the assessment items.

The time allowed for a test and investigation would be one hour. The high student numbers would allow statistical packages to apply classical test analysis with high degrees of reliability with analyses including facility, discrimination, SEM, reliability measures and differential item analysis. In addition, Rasch Modelling would also be carried out.

Test and investigation type items and questions would be developed using the research carried out during phases 1 and 2 of my research, and the developmental platform designed during those phases would allow pupils to interact and navigate their way through the assessments, record all their responses and mark the closed question formats. Open response questions would be marked remotely by an expert marker, and the marks recombined to give an overall score. Paper-based versions would be taken by students in a traditional manner, and sent back to me for manual marking.

Before the main study, a pilot study would be carried out to ensure the assessments themselves were working correctly and suitable for students, that the platform and data collection and marking systems would work, and that the questionnaires and interviews included all aspects of required information

6.5 Phase 4: Test and investigation Construction

After the research plan had been established, phase 4 consisted of the writing, development and programming of two equivalent on-screen tests and investigations.

From the ideas initiated from the first two phases of research, over a number of months, items were written and developed. Some of these items were fairly straight-forward multiple choice questions in order to compare performance of like questions on-paper and on-screen. Some questions differed in the use of particular forms of on-screen closed question answering mechanisms such as the use of drop down lists and dragging and dropping. Some on-screen questions used video and colour photography to enhance the forms of stimulus presented, and the advanced interactive questions allowed engagement and interaction on-screen that was not possible in paper-based versions.

Programming could not take place until items were at an advanced stage of development due to the complexity, time and costs involved with programming items. Writing and development therefore had to be as pro-active as possible in working to on-screen design principles. As much of this work was testing new ground, it was difficult to know with certainty what effect certain decisions would have, however the principles of sound test and exam development were applied wherever possible in terms of layout and language used.

Exam and test development involve a process called shredding. Once an item is written and developed, it is presented to experts for review. This can be a protracted process as every problem or possible flaw in a question or its associated mark scheme is exposed for discussion and criticism. My group of shredders were limited in this project in terms of on-screen experience, however the on-screen assessments were thoroughly shredded and the purpose and working processes of every marking point was established, and the layout and language used was addressed in the context of every individual item. Once questions were shredded, they were amended and then reconsidered until there were no outstanding issues. At this point they were passed to a specialist programmer.

Even though the items had been thoroughly reviewed before this stage, they still had to go through a series of iterations to produce fully worked up working items. As expected, the investigations required most attention, as they did not really operate in the way of a normal test item, and contained a number of sophisticated mechanisms such as the production of anomalous data and the drawing mechanism for lines and curves of best fit. The simulator aspects were quite complicated, as were decisions about what and how to mark the work produced by students. The finalised versions of the on-screen tests and investigations are found in Appendix D, both in screenshots and on CD.

Once the on-screen versions had been produced, the paper-based versions could then be constructed. These would be designed to be as close a match to the on-screen versions as possible in terms of the focus and intent of questions. Answering mechanisms were mirrored in terms of being matched to the on-screen versions, although closed responses have a different format on paper than they do on-screen. The finalised paper-versions are found in Appendix E.

6.6 Phase 5: Construction of Questionnaires and Interviews

Once the form and style of assessments had been established, draft questionnaires and interview questions could be drawn up. There were various attitudinal aspects that I wanted to explore. To gain a comprehensive insight into comparisons between tests and investigations presented in different forms, feedback would be required from pupils and teachers. For the purposes of the pilot, five questionnaires were developed for students, and interview questions for students and teachers. The five questionnaires were focussed on different aspects, but their design was dictated by a few key factors; there would be limited time for their completion, and they would have to be very easy and straightforward for 14 and 15 year old students of differing abilities to access, read and respond to. For these reasons Likert scales were used throughout the questionnaires as they are easy to use as long as the questions themselves are well constructed, unambiguous and clearly written. This form of data would also allow large scale quantitative analysis to be applied. The questionnaires also gave space for pupils to elaborate their views and opinions in open response boxes. This qualitative evidence would be analysed through collation of evidence, coding and template analysis into major themes. A brief description of each of the questionnaires follows:

6.6.1 Paper-based Test Questionnaires

These were designed to gather information about student attitudes to the apparent ease, enjoyment, likes and dislikes about being tested in science in a traditional paper-based method. Students were also invited to comment on whether they thought paper-based tests were a fair way of assessing their scientific ability. These measures would be of interest in their own right, however they would be of even greater value when compared to similar questions based on equivalent on-screen tests. The questions on the questionnaire would be a mixture of Likert scaling and open-response comment. These questionnaires would be taken straight after students took a paper-based test, and would be expected to take 5 minutes to complete. An identical Test paper A and Test paper B questionnaire were constructed. The paper-based test questionnaire can be found in Appendix F.

6.6.2 Paper-based Investigation Questionnaires

These questionnaires were designed to gather attitudinal information about answering an extended paper-based investigation. They were set up in a similar fashion to the paper-based test questionnaire in that they asked students about the apparent ease, enjoyment, likes and dislikes of answering a question of this type. As this was an investigation, I also wanted to ask about their familiarity with the context of the investigation, and the ease of drawing graphs and lines of best fit on paper. They were again asked about whether they thought a paper-based question was a fair way of assessing their investigational skills and they were finally asked to give a preference between the paper-based test and investigation. The test and investigation would be run concurrently so this measure would give a snapshot viewpoint on their preference. The questions were a mixture of Likert scaling and open-response comment, and would be expected to take 5 minutes to complete. Customised versions for Investigations A and B were constructed. The paper-based investigation questionnaires can be found in Appendix G.

6.6.3 On-screen Test Questionnaires

These questionnaires were designed to gather attitudinal information about the ease, enjoyment, likes, dislikes and the fairness of the on-screen assessment method in the same way as the paper-based questionnaires. The similarity of the initial questions in the questionnaires would ease any comparative measures applied to their analysis. The on-screen questionnaires then went into detail about the answering mechanisms used in questions and the stimulus types used in the questions. Even though these questionnaires would be taken straight after the on-screen test was taken, screen shots of relevant questions were inserted into the questionnaire to avoid any ambiguity of understanding of a question and to jog memories if required. The questions were a mixture of Likert scaling and open-response comment, and would be expected to take approximately 5-10 minutes to complete. Bespoke versions of onscreen Test A and Test B had to be constructed due to the insertion of relevant screenshots. The on-screen test questionnaires can be found in Appendix H.

6.6.4 On-Screen Investigation Questionnaires

In the same way as the other questionnaires, these questionnaires started off with questions on student perceptions on the ease, enjoyment, likes and dislikes of the assessment presented in this way. The questionnaire then goes into detail about the context of the investigation, the stimuli presented on-screen and the skills required to carry out certain tasks. In the same way as the on-screen test questionnaire, relevant screen shots were included to ensure that pupils understood what they were being asked about, and to remind them of relevant features and tools within the investigations. As the on-screen test and investigation were to be taken concurrently, the students were also asked about whether they thought completing an investigation in this way was a good test of their process skills and as the on-screen test and investigation were taken concurrently, they were asked to give their preference between the on-screen test and investigation. The questions were a mixture of Likert and open response comment, and would be expected to take approximately 5-10 minutes to complete. Bespoke versions of the on-screen investigation 1 and 2 were constructed to include particular screen shots from each version. The on-screen investigation questionnaires can be found in Appendix I.

6.6.5 Final Comparative Questionnaire

This questionnaire was designed to give students the opportunity to compare the tests and investigations presented in paper-based and on-screen versions. All pupils would take this questionnaire after the last assessment they completed, whether it was paper-based or on-screen. As well as comparative measures of ease, preference, fairness and enjoyment, this questionnaire also asked about relative anxiety levels prior to pupils taking the different forms of assessment. The questionnaire was a mixture of box ticking preference and open response and would be expected to take approximately 5 minutes to complete. This questionnaire can be found in Appendix J.

The information gathered through the questionnaires would be considerable, and allow both quantitative and qualitative analysis. To supplement the questionnaire information, semi-structured interview questions were also constructed for pupils and teachers. The purpose of the pupil interviews was to allow them to expand on any views expressed in the questionnaires and to give them more of a voice to give their opinion on various aspects of the assessments. Oral interviews would allow feedback of a different nature to the questionnaires, allow more freedom of expression for some pupils and allow a group dynamic to operate.

6.6.6 Student Interviews

In advance of the visit to the school, the class teacher was asked to select a small group of students who would like to take part in a short interview with myself or another colleague if I was unable to attend. An interview protocol was produced to ensure that interviews would be carried out with consistency and clarity. A portion of the student interview protocol is found in Appendix K.

6.6.7 Teacher Interviews

An interview protocol for teacher interviews was also drawn up. The teachers had not completed any questionnaires, however they had been present during the paper-based and on-screen assessments. Wherever possible, teachers were given prior access to the assessments before students took them in order to familiarise themselves with the assessments, be prepared for any questions from students during the trials and also in preparation for an interview. The intended focus for interviewing teachers would need to be different to the input from students. I was interested in their perceptions of the purposes and uses of on-screen assessments, both in a test and investigation form. Their views on authenticity of the science and fitness for purpose of the assessments were high on my agenda. The basic teacher interview protocol is found in Appendix L, using a semi-structured approach.

6.7 Phase 6: The Pilot Phase

At this stage, the questions, platform, questionnaires and interviews had all been prepared and were now ready for the pilot phase.

Schools were chosen using a number of criteria. Some of them were specialist science schools, some were GCSE examination centres, others had teachers who had contributed to this project through writing or attending meetings. Approximately 300 letters were sent out at this stage inviting schools to participate in the comparability trial. A copy of the initial invitation letter can be found in Appendix M. The intention was to start to engage with schools to take part in the main study, but I also wanted to get a few schools interested in being our pilot study, in order to check that all our preparation and materials were robust enough for the main study. Three schools who had volunteered to take part were selected for this pilot stage.

Once the schools had expressed interest in taking part in the pilot phase, dates were agreed when the paper-based and on-screen tests and investigations would take place. As promised to the schools, the systems involved were fairly straightforward. The schools selected classes to participate and sent in the class lists so that individual passwords could be assigned to each student. Teachers were also asked to provide us with either a national curriculum test result in science or a teacher assessed level for each student. This data helped in the comparability data analysis carried out later. Where possible, the schools were visited on the day of the on-screen tests and investigations. This enabled the schools to have hands on help if there were any problems with accessing the tests from the web and allowed for the interviews. All the questionnaires and paper-based tests were sent to the schools in advance, and teachers were allowed access to the on-screen tests and investigations to familiarise themselves with the questions and style of delivery. All students were issued with a guide sheet with their individual passwords attached and instructions on how to access the test website and the test and investigation versions they were taking. This guide sheet can be found in Appendix N. All the paper-based tests, taken before or after the on-screen tests were sent back to us for marking.

On the day of a visit, the teacher responsible for each class was asked to sign a form allowing the quantitative and qualitative data to be used for research purposes. This form can be found in Appendix O.

6.8 Phase 7: Pilot Evaluation

This phase was concerned with evaluating the pilot study in order to amend or re-consider any of the constituent elements involved, in order to ensure the smooth running and operation of the main study. These elements included the items in the tests themselves and their surrounding administration. I will go through and address each of the emergent issues in turn.

6.8.1 The Tests and Investigations

Feedback concerning the test questions and the investigations was positive. The tests appeared to be written and presented in an unambiguous manner and through observing students taking the tests and talking to students and teachers, it was pleasing that all the time and effort taken in question writing and construction was well spent. The tests used a variety of answering mechanisms, and it was evident through observation that the students required little or no help in using different answering techniques, either on the paper-based or on-screen versions. The investigations were more complex in their construction, and some students needed to be shown how to manipulate the on-screen simulator in order to obtain different readings. This was not an unexpected finding and in many ways I might have expected more problems. The widespread use of gaming technology has given students an intuitive feel for using keyboards and mice to carry out instructions. Evidence for these views emerged as different strands of data were combined.

The main issue emerging from the review of the tests and investigations was that the tests in particular were too long. Secondary school timetables are usually arranged in 1 hour sessions and it became clear that this was not enough time for students to comfortably complete a test and an investigation. It would be an important element of this comparability study to look in detail at overall test and item level performance, and the results could be contaminated if students simply ran out of time or had to rush through questions. Therefore the number of questions in the test sections would need to be reduced for the main study.

No data analysis was carried out from the pilot study as the sample numbers were not sufficient to reliably apply statistical programmes.

6.8.2 The Platform

This was the aspect of the whole study that carried the most risk. A decision had been taken not to use existing test platforms and data capturing software as they would not have been able to support or mark the types of questions I wanted to investigate. The creation of a bespoke research platform that could cope with innovative items and response types was an exciting prospect, however untried. This pilot phase was the first opportunity to find out if it could cope with multiple users at the same time using web access, and whether all the necessary data would be successfully captured.

The results from the pilot phase were successful. Students had no problems logging in and gaining access to the tests and investigations and most were able to navigate their way through and submit their assessments to the server. Once back from the pilot centres, the data allowed automatic and open response marking, which could then be collated and analysed through SPSS.

Two problems were identified during the pilot phase. The first was perhaps inevitable. For a few students, while working through the on-screen assessments, for no apparent reason, the test or investigation crashed. As the data was not captured until an entire test or investigation was submitted, this resulted in the loss of all the work of those students. They could log in again, however they then needed to go through the entire test or investigation again. Depending on the time available or the mood of the student, this was not always possible. The platform was subsequently altered, so that once an item was completed, the data was captured and submitted immediately, and not only once the entire test or investigation was completed and submitted.

The second problem was more significant. In one centre, which was set up with a wireless network, students took the on-screen tests and investigations on laptops. This would appear to be an ideal scenario for school based assessments. The on-screen assessments however exposed a major potential flaw in the system. If, at any point during an assessment window, wireless signals dropped below a certain threshold, which is not unusual with this technology, the connection to the server was broken, and all the data lost. This happened in one particular centre. There was not enough time to restart the whole assessment, and therefore the whole comparability exercise was compromised as the paper based tests were not useable for comparability purposes without the on-screen element. The lesson learnt here was harsh, but informative. For the main study, schools would need to use LAN based connections. The amendment to the platform in terms of capturing data throughout the assessments would however help to alleviate data loss problems.

6.8.3 The Questionnaires and Interviews

Students seemed to have few problems answering these styles of questions. Data from the questionnaires were checked through, captured and discussed, however, no formal data analysis was applied at this stage as the sample numbers were not large enough. The purpose was to ensure that the questionnaires seemed to cover all aspects of interest and that they were understandable and manageable to complete.

The test and investigation paper-based and on-screen questionnaires did not appear to have omitted anything significant and the use of screenshots in the on-screen versions were commented on by students as being helpful in remembering certain types of questions.

The main issue emerging from the questionnaires concerned manageability. The amount and type of information to be collected was comprehensive, however the window of time available to go into schools, for students to take the tests and investigations and to complete questionnaires and interviews was problematic. The priority for every school was to complete the tests and investigations as this data was essential for comparability data analysis. The strategy for questionnaires and interviews would need to change. If the time provided by schools in the main research study was not tightly restricted, the complete range of questionnaires could still be used. However, if time was an issue, schools were advised to concentrate on the comparative questionnaire. The other questionnaires and interviews were carried out on a sampled basis, and therefore it was not necessary to apply them to every student and teacher to gain representative views.

The interviews carried out were successful. No one had any objections to be recorded and so valuable views and opinions were electronically captured. Interviewing students in small groups allowed fluid discussions and interactivity of ideas that were not evident from the individual questionnaires. Even on a sampled basis, this evidence will complement the quantitative data analysis applied to test and investigation performance.

6.8.4 School Issues

The significant school issue was the difficulty in getting them to participate in the comparative study. Out of an initial 300 letters to schools, only 5 schools were willing to participate. There were a number of possible reasons for this including the need to give up 2 hours of curriculum time, restricted access to computer rooms or suites, uncertainty about any benefits to the department or schools and any perceived administration or arrangement difficulties for the teachers. The research project had been carefully planned to be minimally disruptive to schools and administratively light on participating teachers. The method was accordingly amended in order to enable schools to be more flexible in how passwords were allocated and when the tests and investigations were taken. Each science department was also paid the equivalent of a half day teacher supply rate in order to acknowledge their time and co-operation in taking part in this study.

6.9 Phase 8: The Main Research Study

Once amendments were made on the basis of the outcomes of the pilot study, the content of the tests and investigations were finalised and prepared for roll-out. Similarly, the questionnaires and interview schedules were adjusted and finalised.

The final sample of 14 schools was representatively selected out of approximately 20 who had expressed interest, and arrangements were made for them to carry out the assessments in the presence of a researcher where possible or remotely if required. In either case, the tests and investigations were completed by students, together with questionnaires. As described previously, student and teacher interviews were carried out on a sampled basis. Schools were asked to take their assessments in a pre-determined order, to account for any suggestion that test or investigation order would influence performance. This was possible for some schools, however others could not always adhere to the given order due to computer-room access restrictions. This resulted in a non exact, (however statistically adequate) distribution of students taking tests in the order set out in the four parameter research model.

As schools took the tests and investigations, their on-screen responses could be viewed immediately from the server, which gave re-assurance that the tests, investigations and platform were working correctly and also gave some initial insight into students responses.

Paper-based tests and investigations and all questionnaires were either collected, if the schools had been visited, or posted back to us if not.

The marking method was the same as outlined in Phase 3. Once all the tests and investigations were automatically or human marked, the data was collated, and then analysed. The questionnaire and interview data and evidence were also collated and then analysed.

6.10 Phase 9: Analytical framework

Chapter 5 described the rationale for the collection of qualitative and quantitative evidence. Once both of these strands of data and evidence were collected and collated, they were then analysed.

6.10.1 Quantitative Analysis

The quantitative performance data from the tests and investigations were put into spreadsheets, and all the data was cleaned in order to ensure that analysis would only be carried out on students submitting both paper and on-screen assessments and that all the usable data was free from any omissions or errors.

This cleaned data was then analysed using two separate approaches. The data was imported into SPSS, which calculated all the classical test analyses traits, eg. mean marks, standard deviations, facility and discrimination values, Cronbach's alpha and DIFs. The SEM was also calculated using these figures. The data was also analysed through Rasch, which demonstrated how groups and individuals performed in whole tests and investigations and also demonstrated performance in each item. The quantitative evidence is shown in Chapter 7.

6.10.2 Qualitative Analysis

All the Likert scaled data from the questionnaires were entered into spreadsheets, and then analysed from the spreadsheets. The open-ended responses from the questionnaires were also entered into spreadsheets using codes to categorise responses into key categories. Once the open responses were entered in this manner, it was easier to determine how prevalent particular views were in relation to others. This form of template analysis (King, 1998) helped to identify the main emergent themes from the questionnaires, which were then considered alongside the Likert data.

Evidence from the interviews was similarly analysed. All interviews were transcribed, and then comments were categorised though coding. The evidence from the interviews was then also considered alongside the questionnaire evidence.

6.11 Summary

This chapter has discussed how my methodology was operationalised as a method. The nine phases have been described from early research stages to the analysis of the data gathered in the main study.

The next two chapters show the quantitative and qualitative outcomes of the main research study in Chapters 7 and 8 respectively. Finally, the quantitative and qualitative evidence will be considered together, and analysed in terms of supporting or contradictory outcomes. This analysis can be found in Chapter 9.

Chapter 7

Quantitative Results

7.1 Introduction

As described in Chapter 6, once all the performance data was collected from the testing platform, it was all transferred into excel spreadsheets. Once in this format, the data could then be cleaned and then statistically analysed.

Cleaning refers to a process whereby any data sets that are incomplete or non-matched can be removed from the sample. Incomplete data consisted of student data sets where, for some reason, part of the data had not been transferred correctly or is missing. Sometimes this data could be retrieved from the testing platform, however if the data could not be recovered, all the test data from a particular student or group of students had to be removed from the data set. There were few instances of this issue in this study. The second requirement for cleaning data came where students took a test and investigation in one mode, but then were absent for the test in the alternative mode. When this happened, a matched pair of modal data was not available for analysis, and therefore all the data from those students were removed from the sample. Overall, cleaning reduced the sample size from 1313 to 989.

This chapter will show the outputs of classical test and Rasch analysis on the performance of the two groups of pupils (Group 1 and 2) on two equivalent tests and investigations (A and B) in paper and on-screen modes.

A total number of 1313 pupils took part in this study. However, as described, I was only able to use students who had completed both paper-based and on-screen tests and investigations and also only fully populated data sets. Of the original sample of 1313 students, there were 989 fully populated matched pairs. Therefore classical test analysis using SPSS and Rasch analysis using RUMM 20/20 was carried out on the matched pair sample size of 989.

7.2 SPSS and Rasch

In Chapter 5 I described the central features of classical test and Rasch analyses and justified their role as complementary components of my quantitative approaches.

Once all the data sets had been cleaned, they were imported into SPSS and Rumm 20/20 for classical and Rasch analysis respectively. SPSS was used to produce a range of test and item analyses. Mean test scores and standard deviations were calculated as were the internal reliability co-efficient (Cronbach's alpha) and the Standard Error of Measurement (SEM) for each test and investigation in paper and on-screen mode. Within each test, facilities and discriminations were calculated for all items and any differential item functioning (DIF) determined for mode and gender. A DIF indicated where there was a significant difference between any given test or item performance by either mode or gender. Although significance is usually determined by any figure above a significance of $p <0.05$, I used the figure $p <0.005$ in my research to focus on high levels of significant differences.

The cleaned data was also imported in Rumm 20/20. This analysis enabled me to explore how the tests and items performed in a one parameter latent trait model. The single parameter in Rasch being the test or item difficulty, in relation to the ability profile of the student cohort. In particular, this analysis enabled me to confirm any test or item DIF analysis, and in addition, explore any differences in performance across the ability profile of the students in a clear visual manner.

Before the SPSS and Rasch analyses are shown, the student sample is described.

7.3 The student sample

Table 4 below shows the distribution of students taking each set of tests

Table 4: Total student numbers who completed a matched computer and a paper test

Student Group	Mode and Test version taken	Student Numbers
Group 1	Paper Test A; Computer Test B	479
Group 2	Computer Test A; Paper Test B	510
Groups 1 and 2		989

Table 5 below shows that the teacher assessment profiles of Group 1 students alongside those of Group 2. There were more students in Group 1 attributed national curriculum levels 6 and 7 than in Group 2, although the number of students teacher assessed at levels 5 and above were nearly identical for both groups (388 and 383 respectively).

Mann -Whitney and Kolmogorov-Smirnov tests were carried out to establish whether there was any difference in the distribution of teacher assessment levels between the two group samples. A significance of $p=0.000$ in both of these tests revealed that the student sample of Group 1 was significantly higher in ability than Group 2 based on their teacher assessment national curriculum levels.

Table 5: Teacher Assessment levels of students in Groups 1 and 2

Teacher Assessed (TA) Levels			Group1 students	Group 2 students	Totals
Level	2	Number Count	5	7	12
		% of Group	1.0%	1.4%	1.2%
	3	Number Count	16	25	41
		% of Group	3.4%	4.9%	4.2%
	4	Number Count	68	93	161
		% of Group	14.3%	18.3%	16.3%
	5	Number Count	141	195	336
		% of Group	29.6%	38.4%	34.1%
	6	Number Count	184	136	320
		% of Group	38.6%	26.8%	32.5%
	7	Number Count	63	52	115
		% of Group	13.2%	10.2%	11.7%
Total		Number Count	477	508	985
		% of Group	100.0%	100.0%	100.0%

A total number of 985 pupils had teacher assessment levels attributed to them by their teachers. 4 students did not, and so are not included in the teacher assessment data.

7.4 Test scores

Test A and B had maximum scores of 81 and 77 respectively. Table 6 shows the mean raw scores and percentage scores for each of the two tests across both modes.

Table 6: Mean test scores

Test	Medium	Group	Mean Scores	Mean %	N
A	Paper	1	57.63	71.15	479
	Computer	2	47.20	58.28	510
	Total	All	52.25	64.51	989
B	Computer	1	51.88	67.38	479
	Paper	2	45.38	58.93	510
	Total	All	48.53	63.02	989

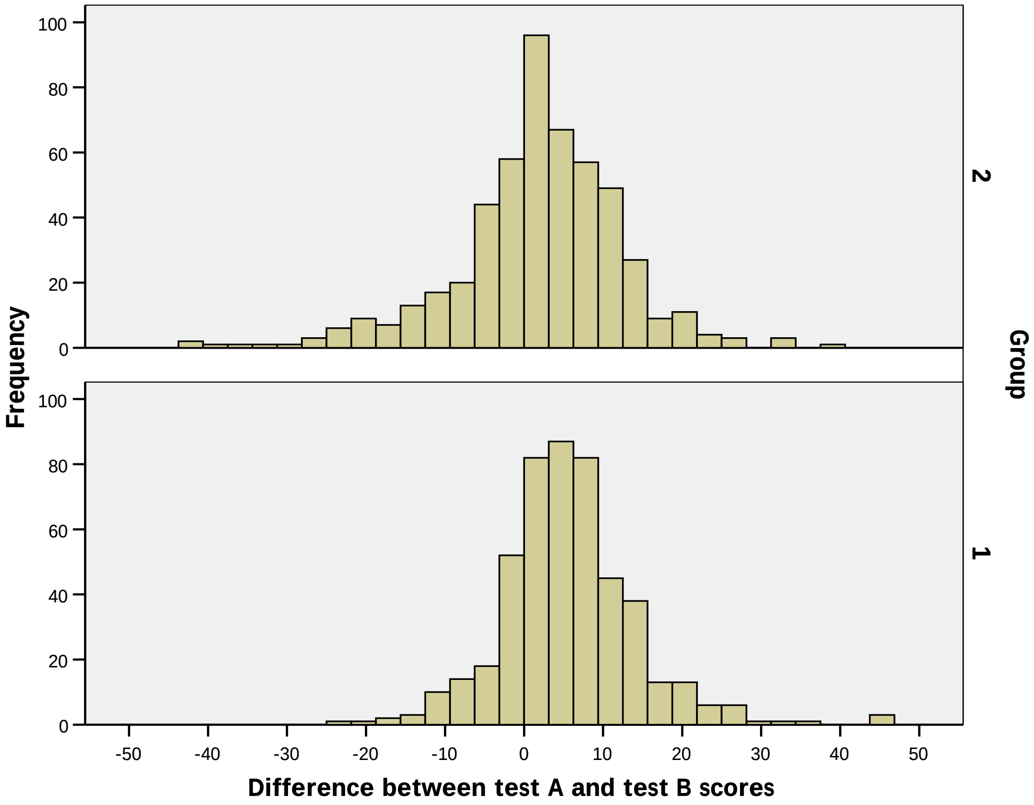
The figures in Table 6 above show that Group 1 achieved higher means on the paper-based and computer tests than Group 2. The Mann-Whitney and Kolmogorov-Smirnov tests described earlier established that Group 1 had a significantly higher ability profile than Group 2 in terms of their teacher assessed national curriculum levels.

Figure 6 below shows the difference in mean marks scored by Groups 1 and 2 taking paper and on-screen tests.

The Group 2 chart (students taking Test A on computer and Test B on paper) shows that there is a slightly negative difference and distribution between the two mean marks. This shows that students scored slightly higher on their computer test than their paper test. The mean difference was 1.83 marks in favour of computer.

The Group 1 chart (pupils taking Test A on paper and Test B on computer) showed a positive difference and distribution between the two mean marks. This shows that students scored higher on their paper test than on their computer test. The mean difference was 5.75 marks in favour of paper

Figure 6: Graphs to show the difference in mean marks scored by Groups 1 and 2, taking paper and on-screen tests.



There does not appear to be any correlation between Teacher Assessment level and the mean difference between papers. This is shown in Table 7 below, which shows the mean difference of scores between modes according to the teacher assessed national curriculum level. The disparity between Groups 1 and 2 is in part due to the difference in ability of the groups. As very few students were L2, this data can largely be ignored. Levels 3-7 show a variety of mean differences, with no underlying traits.

Table 7: The difference in mean marks scored across the ability ranges of Groups 1 and 2 taking paper and on-screen tests

	Difference in mean marks across modes by student ability levels, based on Teacher Assessment (TA) Levels						
	2	3	4	5	6	7	Total
Group 1	8.6	2.4	2.9	7.3	5.7	5.7	5.7
Group 2	4.3	2.1	3.2	3.0	.3	-1.3	1.8

As there was a difference between the ability between the two groups, it was necessary to carry out more detailed analysis of modal differences. This analysis was to investigate whether there was any difference between test modes once imbalances between the groups had been taken into account. This was carried out as follows: Table 8 below shows the mean difference in performance between Test A and Test B for each of the 2 groups.

Table 8: Score differences between Test A and Test B

Student Group	Tests Taken	Mean mark difference between modes	Std. Deviation
1	Test A: Paper Test B: Computer	5.75 better on paper	8.59
2	Test A: Computer Test B: Paper	1.83 better on computer	10.63
Whole sample		3.73 better on paper	9.89

It can be seen that Group 1 students scored an average of 5.75 marks higher on the paper-based test. Group 2 students scored an average of 1.83 marks better on the computer-based test.

The equation below was then used to calculate the average mode difference:

A represents the average score for Test A

B represents the average score for Test B

Group 1 $A - (B + C) = 5.75$

Group 2 $(A + C) - B = 1.83$

$\Rightarrow 2C = -3.92$

$\mathbf{C = -1.96}$

C is the calculated difference in performance between modes, showing approximately a 2 mark difference in favour of the paper-based versions.

7.5 Reliability Measures of the Tests

Chapters 2 and 5 outlined the uses of internal reliability measures in tests. They are an indication of how well items within tests correlate with each other and with total student scores assuming that the test is measuring a particular construct. They operate on the premise that there will be consistency of performance across a test by students. Therefore there will be a predictable gradation of student performance across the ability range.

The internal reliability indicator I used in this study was Cronbach's alpha co-efficient, and the results for the Test versions A and B are shown below in Table 9.

Table 9: Internal Reliability Measures of Tests A and B

Test A

Test Mode	Cronbach's Alpha	N of Items
Paper	.919	65
Computer	.927	65

Test B

Test mode	Cronbach's Alpha	N of Items
Paper	.930	60
Computer	.930	60

Cronbachs Alpha co-efficients of Test A in paper and computer modes both showed very high level of Internal Reliability. The high number of items in the tests ensured reliable and secure values.

Cronbachs Alpha Co-efficients of Test B in paper and computer modes also showed very high levels of reliability, and as can be seen, identical values were achieved on the paper and computer versions. Although there were 5 fewer items in the Test B versions, 60 items in the tests ensured a reliable and secure measurement.

Overall, both versions of tests, in paper based and computer based modes demonstrated very high Cronbach's Alpha scores and any one of them would be operable in a high stakes assessments as evidenced in Chapter 2 from the DIIA (2003).

7.6 Standard Error of Measurement of Tests A and B (SEM)

Using the standard deviation from the tests and the associated Cronbach Alpha co-efficient, the SEM was calculated for each of the tests, in each mode. The results are shown in Table 10 below.

Table 10: Standard Error of Measurement (SEM) for Tests A and B.

Test Version	Test Group	Mode	Cronbach's alpha Co-efficient	Standard Deviation (SD)	Standard Error of Measurement (SEM)
A	1	Paper	0.919	13.20	3.76
A	2	Computer	0.927	16.48	4.45
B	2	Paper	0.930	15.99	4.23
B	1	Computer	0.930	13.34	3.53

The SEM for each test version was consistent for each student group who had taken a matched pair of paper and on-screen tests. The SEM figure estimates the potential error of the mean scores of student relative to a theoretical true mean, therefore the smaller this figure, the less the theoretical error of student scores on the test.

The following section of this chapter will show the performance across Tests A and B at item level.

Table 11 below shows the summary statistics of Test A. The performance of the items in the paper and on-screen versions are shown alongside each other. The facility and discrimination values for all items in both modes are shown. In addition to this information, significant differences in performance across the modes are also shown, and highly significant differences (DIFs) are indicated, together with the mode that these DIFs favoured.

Within Table 11, the comparative facilities and discrimination values for the same items across across modes are shown. Facility value is essentially a performance indicator, showing the percentage of students getting an item or marks within items correct. Discrimination values indicate the correlation between student performance on a particular item to performance across the test as a whole, therefore the amount of correlated differentiation.

7.7 Item statistics for Test A

Table 11: Item Statistics for Test A, showing Facility, Discrimination and DIFs values

Although significant differences can be attributed to values p< 0.05 (a 5% probability of results occurring by chance, I have taken this significance level down to p< 0.005 (a 0.05% probability of results occurring by chance).

Item Number	Number on Paper Test	Number on Computer Test	Items showing significant difference at **0.005** in named mode	Significance values for these items p>0.005	Significance values for all items	Facility values for Paper version	Facility values for Computer version	Discrimination values for Paper version	Discrimination values for Computer version
1	1a	T1_1AA			0.065	0.99	0.95	0.154	0.364
2	2b	T1_1BB			0.168	0.95	0.88	0.280	0.479
3	1c1	T1_1CC			0.095	0.93	0.83	0.305	0.521
4	1c2	T1_1DD			0.062	0.89	0.77	0.236	0.487
5	1d	T1_2DA			0.012	0.49	0.46	0.245	0.374
6	2a	T1_3AA			0.190	0.44	0.38	0.373	0.398
7	2b	T1_4BA			0.126	0.81	0.67	0.390	0.403
8	2c	T1_4CB			0.016	0.97	0.91	0.354	0.41
9	2d	T1_5DA			0.208	0.95	0.89	0.383	0.344
10	3a	T1_6AA			0.522	0.83	0.71	0.239	0.295
11	3b	T1_6BB	Paper	0.000	0.000	0.87	0.61	0.316	0.242
12	3c	T1_7CA			0.646	0.74	0.62	0.298	0.295
13	3d	T1_7DB			0.014	0.57	0.5	0.285	0.251
14	3e	T1_7EC			0.112	0.76	0.6	0.360	0.365
15	3f	T1_7FD			0.006	0.93	0.83	0.395	0.404
16	4a1	T1_8AA	Computer	0.000	0.000	0.15	0.19	0.045	0.026
17	4a2	T1_8BB	Paper	0.001	0.001	0.89	0.74	0.413	0.477
18	4a3	T1_8CC			0.025	0.95	0.88	0.292	0.363
19	4a4	T1_8DD	Paper	0.002	0.002	0.89	0.75	0.483	0.579
20	4a5	T1_8EE	Paper	0.000	0.000	0.95	0.85	0.325	0.421
21	4a6	T1_8FF			0.889	0.82	0.71	0.338	0.336
22	5a	T1_9AA			0.006	0.80	0.62	0.464	0.562
23	5b	T1_9BB			0.011	0.73	0.54	0.444	0.496
24	5c	T1_9CC			0.168	0.94	0.88	0.426	0.433
25	5d	T1_10DA	Computer	0.000	0.000	0.47	0.75	0.316	0.186
26	6a	T1_11AA			0.374	0.42	0.36	0.482	0.609
27	6b	T1_12BA			0.008	0.86	0.73	0.472	0.456
28	6c	T1_12CB			0.159	0.83	0.79	0.444	0.462
29	6d	T1_12DC			0.091	0.88	0.86	0.305	0.431
30	7a	T1_13AA			0.019	0.82	0.66	0.460	0.546
31	8a	T1_14AA	Paper	0.003	0.003	0.93	0.81	0.464	0.454
32	8b	T1_14BB	Paper	0.002	0.002	0.88	0.74	0.430	0.412
33	8c	T1_14CC	Paper	0.001	0.001	0.95	0.84	0.451	0.411
34	8d	T1_15DA			0.337	0.88	0.85	0.327	0.319
35	8e	T1_15EB			0.059	0.68	0.50	0.351	0.309
36	8f	T1_15FC			0.044	0.70	0.52	0.356	0.347
37	8g	T1_15GD			0.050	0.70	0.53	0.394	0.381
38	8h	T1_16HA			0.201	0.76	0.68	0.392	0.401
39	8i	T1_16IB			0.050	0.41	0.25	0.324	0.196
40	9a	T1_17AA			0.246	0.36	0.27	0.617	0.603
41	9b	T1_18BA	Paper	0.000	0.000	0.73	0.4	0.562	0.517
42	9c	T1_18CB			0.190	0.47	0.31	0.535	0.518

43	9d	T1_18DC			0.952	0.31	0.22	0.478	0.502
44	9e	T1_18ED			0.007	0.33	0.17	0.469	0.428
45	9f	T1_18FE			0.353	0.31	0.20	0.464	0.485
46	9g	T1_18GF	Paper	0.005	0.005	0.32	0.15	0.462	0.459
47	10a	T1_19AA			0.436	0.65	0.50	0.497	0.413
48	10b	T1_19BB	Paper	0.000	0.000	0.87	0.66	0.529	0.542
49	10c	T1_19CC			0.063	0.65	0.47	0.445	0.396
50	10d	T1_20DA			0.005	0.76	0.57	0.571	0.504
51	10e	T1_20EB			0.010	0.73	0.53	0.614	0.54
52	10f	T1_21FA	Paper	0.000	0.000	0.88	0.68	0.541	0.547
53	10g	T1_21GB			0.005	0.71	0.50	0.500	0.574
54	11a	T1_22AA			0.303	0.60	0.45	0.411	0.381
55	11b	T1_23BA			0.719	0.56	0.39	0.754	0.668
56	12a	T1_24AA			0.009	0.38	0.27	0.594	0.54
57	12b	T1_25BA	Computer	0.000	0.000	0.35	0.33	0.458	0.518
58	13a	T1_26AA	Paper	0.000	0.000	0.74	0.45	0.705	0.671
59	13b	T1_26BB	Paper	0.000	0.000	0.78	0.53	0.667	0.638
60	13c	T1_26CC			0.007	0.81	0.63	0.573	0.544
61	13d	T1_26DD	Paper	0.000	0.000	0.59	0.35	0.558	0.57
62	13e	T1_27EA	Paper	0.000	0.000	0.67	0.40	0.649	0.606
63	13f	T1_27FB			0.276	0.52	0.36	0.526	0.338
64	13g	T1_28GA			0.787	0.53	0.40	0.43	0.477
65	13h	T1_28HB			0.353	0.17	0.10	0.228	0.15

Table 11 also shows any significant levels of difference between the performance of items in paper and on-screen modes. For my research purposes, I concentrated on very highly significant differences, ones where $p < 0.005$. This are classified as DIF items (differential item functioning). These DIF items are also indicated in Table 11, together with the mode in which they occurred.

Table 12 shows all the DIF items from Test A together. Their facility and discrimination values in both modes are shown, and also their level of DIF in terms of a p value.
Table 12 also includes information on the types of item showing DIFs in test A. The science subject area, the stimuli and response mechanism for each of the items are coded. The legends for these codes are shown in Table 13.

Table 12: Test A. Collated data showing differential item performance at p<0.005

Paper Question Number No.	Computer Question Number.	Computer Question Type: Response, subject and stimulus	Paper Q. Facility	Computer Q. Facility	Discrimination on Paper version	Discrimination on Computer version	Significant Difference in performance of named mode p<0.005
3b	6BB	DD C V	0.87	0.61	0.316	0.242	0.000 P
4a1	8AA	DL S D	0.15	0.19	0.045	0.026	0.000 C
4a2	8BB	DL S D	0.89	0.74	0.413	0.477	0.001 P
4a4	8DD	DL S D	0.89	0.75	0.483	0.579	0.002 P
4a5	8EE	DL S D	0.95	0.85	0.325	0.421	0.000 P
5d	10DA	DR P D	0.47	0.75	0.316	0.186	0.000 C
8a	14AA	DL P CP	0.93	0.81	0.464	0.454	0.003 P
8b	14BB	DL P CP	0.88	0.74	0.430	0.412	0.002 P
8c	14CC	DL P CP	0.95	0.84	0.451	0.411	0.001 P
9b	18BA	DD B D	0.73	0.40	0.562	0.517	0.000 P
9g	18GF	DD B I	0.32	0.15	0.462	0.459	0.005 P
10b	19BB	DL B O	0.87	0.66	0.529	0.542	0.000 P
10f	21FA	OR1 B I	0.88	0.68	0.541	0.547	0.000 P
12b	25BA	DD C CP	0.35	0.33	0.458	0.518	0.000 C
13a	26AA	DL P I	0.74	0.45	0.705	0.671	0.000 P
13b	26BB	DL P I	0.78	0.53	0.667	0.638	0.000 P
13d	26DD	OR2 P I	0.59	0.35	0. 558	0.570	0.000 P
13e	27EA	DD P I	0.67	0.40	0.649	0.606	0.000 P

Table 13: Item Codes for Tables 12 and 15

Category	Type	Code
Response type	Drag and drop	DD
	Drop down list	DL
	Draw	DR
	Open response (Numeric)	OR
	Open response (Single word)	OR1
	Open response (Extended writing)	OR2
	Tick box	TB
Subject	Biology	B
	Chemistry	C
	Physics	P
	Science1	S
Stimulus	Colour photo/ drawing	CP
	Diagram/ drawing	D
	Diagram and Information box	DI
	Information box	I
	Interactive diagram	ID
	No stimulus	O
	Table	T
	Video	V

7.8 Item statistics for Test B

Table 14: Item Statistics for Test B, showing facility and discrimination values

Item Number	Number on Paper Test	Number on Computer Test	Items showing significant difference at 0.005 in named mode	Significance values for these items at p<0.005	Significance values for all items	Facility values for Paper version	Facility values for Computer version	Discrimination values for Paper version	Discrimination values for Computer version
1	1a	T2_1AA			0.093	0.68	0.80	0.471	0.466
2	1b	T2_1BA			0.920	0.30	0.43	0.265	0.218
3	1c	T2_2CB			0.363	0.88	0.94	0.389	0.335
4	1d	T2_2DC	Computer	0.005	0.005	0.90	0.97	0.416	0.313
5	1e	T2_3EA			0.025	0.77	0.87	0.457	0.372
6	1f	T2_3FB			0.436	0.75	0.84	0.350	0.301
7	2a	T2_4AA			0.156	0.68	0.76	0.451	0.253
8	2b	T2_5BA			0.424	0.76	0.84	0.406	0.423
9	2c	T2_5CB			0.960	0.79	0.86	0.441	0.356
10	2d	T2_5DC			0.165	0.54	0.66	0.457	0.375
11	2e	T2_6EA			0.401	0.30	0.43	0.332	0.325
12	3a	T2_7AA			0.430	0.60	0.68	0.362	0.319
13	3b	T2_8BA			0.168	0.55	0.69	0.352	0.266
14	3c	T2_9CA	Computer	0.000	0.000	0.75	0.89	0.495	0.38
15	3d	T2_10DA			0.549	0.50	0.64	0.246	0.32
16	3e	T2_11EA	Paper	0.000	0.000	0.67	0.64	0.209	0.269
17	4a	T2_12AA	Paper	0.000	0.000	0.69	0.70	0.484	0.387
18	4b	T2_12BB			0.289	0.68	0.79	0.434	0.393
19	4c	T2_13CA			0.119	0.78	0.82	0.449	0.455
20	4c	T2_13DB	Paper	0.000	0.000	0.68	0.57	0.301	0.178
21	4d	T2_13EC			0.596	0.25	0.34	0.333	0.336
22	5a	T2_14AA			0.795	0.76	0.82	0.521	0.421
23	5b	T2_14BB	Paper	0.000	0.000	0.73	0.72	0.52	0.507
24	5c	T2_15CA	Computer	0.000	0.000	0.54	0.76	0.312	0.39
25	5d	T2_16DA			0.131	0.92	0.92	0.352	0.431
26	5e	T2_16EB			0.097	0.42	0.48	0.324	0.4
27	5f	T2_16FC			0.082	0.52	0.69	0.328	0.36
28	6a	T2_17AA	Paper	0.000	0.000	0.62	0.23	0.239	-0.004
29	7a	T2_18AA			0.787	0.61	0.69	0.254	0.364
30	7b	T2_18BB	Paper	0.000	0.000	0.51	0.51	0.21	0.229
31	7c	T2_19CA			0.466	0.58	0.68	0.358	0.403
32	7d	T2_19DB			0.589	0.38	0.50	0.345	0.377
33	7e	T2_20EA			0.080	0.68	0.80	0.588	0.516
34	7f	T2_20FB	Paper	0.000	0.000	0.87	0.84	0.419	0.561
35	8a	T2_21AA			0.010	0.48	0.64	0.617	0.603
36	8b	T2_22BA			0.108	0.78	0.82	0.473	0.474
37	8c	T2_22CB			0.011	0.75	0.77	0.469	0.528
38	8d	T2_23DA			0.230	0.61	0.74	0.585	0.667
39	8e	T2_23EB			0.667	0.42	0.54	0.473	0.516
40	8f	T2_23FC			0.379	0.19	0.28	0.219	0.279
41	9a	T2_24AA			0.042	0.83	0.85	0.525	0.645
42	9b	T2_24BB			0.757	0.68	0.78	0.527	0.561
43	9c	T2_25CA			0.749	0.62	0.71	0.582	0.544
44	9d	T2_25DB	Paper	0.003	0.003	0.63	0.70	0.63	0.664
45	9e	T2_25EC			0.019	0.59	0.66	0.654	0.637

46	9f	T2_25FD			0.430	0.76	0.81	0.567	0.636
47	10a	T2_26AA	Paper	0.000	0.000	0.69	0.77	0.709	0.656
48	11a	T2_27AA			0.212	0.38	0.52	0.506	0.559
49	11b	T2_27BB			0.873	0.41	0.50	0.528	0.538
50	11c	T2_27CC	Paper	0.000	0.000	0.41	0.34	0.435	0.416
51	11d	T2_28DA			0.276	0.67	0.77	0.665	0.714
52	11e	T2_29EA			0.780	0.46	0.60	0.543	0.644
53	12a	T2_30AA			0.171	0.26	0.31	0.193	0.197
54	12b	T2_30BB			0.424	0.55	0.65	0.436	0.472
55	12c	T2_31CA			0.682	0.22	0.28	0.344	0.393
56	12d	T2_31DB			0.337	0.59	0.67	0.544	0.616
57	12e	T2_31EC			0.038	0.62	0.67	0.55	0.595
58	13a	T2_32AA	Paper	0.000	0.000	0.45	0.41	0.533	0.461
59	13b	T2_32BB	Paper	0.000	0.000	0.41	0.43	0.526	0.474
60	13c	T2_32CC	Paper	0.000	0.000	0.16	0.15	0.375	0.342

Table 14 above shows the summary statistics of Test B. The performance of the items in the paper and on-screen versions are shown alongside each other. The facility and discrimination values for all items in both modes are shown. In addition to this information, significant differences in performance across modes are also shown, and highly significant differences (DIFs) are indicated, together with the mode that these DIFs favoured.

Within Table 14, the comparative facilities and discrimination values for the same items across across modes are shown. Facility value is essentially a performance indicator, showing the percentage of students getting an item or marks within items correct. Discrimination values indicate the correlation between student performance on a particular item to performance across the test as a whole, therefore the amount of correlated differentiation.

Table 14 also shows any significant levels of difference between the performance of items in paper and on-screen modes. For my research purposes, I concentrated on very highly significant differences, ones where p <0.005. This are classified as DIF items (differential item functioning). These DIF items are also indicated in Table 14.

Table 15 shows all the DIF items from Test B together. Their facility and discrimination values in both modes are shown, and also their level of DIF in terms of a p value.
Table 15 also includes information on the types of item showing DIFs in test B. The science subject area, the stimuli and response mechanism for each of the items are coded. The legends for these codes are shown in Table 13.

Table 15: Test B. Statistics showing differential item performance at P< 0.005 Sig level.

Paper Question Number No.	Computer Question Number.	Computer Question Type: Response, subject and stimulus	Paper Q. Facility	Computer Q. Facility	Discrimination on Paper version	Discrimination on Computer version	Significant Difference in performance of named mode. P<0.005
1d	2DC	OR1 B D	0.9	0.97	0.416	0.313	0.005 C
3c	9CA	DL C CP	0.75	0.89	0.495	0.380	0.000 C
3e	11EA	DD C T	0.67	0.64	0.209	0.269	0.000 P
4a	12AA	DD C CP	0.69	0.70	0.484	0.387	0.000 P
4c	13DB	DL C T	0.68	0.57	0.301	0.178	0.000 P
5b	14BB	TB P T	0.73	0.72	0.520	0.507	0.000 P
5c	15CA	DR P T	0.54	0.76	0.312	0.390	0.000 C
6a	17AA	TB P ID	0.62	0.23	0.239	-0.004	0.000 P
7b	18BB	TB P D	0.51	0.51	0.210	0.229	0.000 P
7f	20FB	TB P D	0.87	0.84	0.419	0.561	0.000 P
9d	25DB	TB B O	0.63	0.70	0.630	0.664	0.003 P
10a	26AA	DD C CP	0.69	0.77	0.709	0.656	0.000 P
11c	27CC	OR C I	0.41	0.34	0.435	0.416	0.000 P
13a	32AA	OR P T	0.45	0.41	0.533	0.461	0.000 P
13b	32BB	OR2 P T	0.41	0.43	0.526	0.474	0.000 P
13c	32CC	OR2 P T	0.16	0.15	0.375	0.342	0.000 P

The following section of this chapter will explore the Rasch analyses of Tests A and B as a whole and then look at the identified items within the tests that had highly significant DIF performances across modes.

For each DIF item, the paper and onscreen facility and discrimination values will be shown, alongside the level and mode of DIF. Screen shots of what these items looked like in paper and on-screen modes will also be shown together with a brief description of the items and their performances. Rasch item characteristic curves will also be shown for all the DIF items, visually showing how these items performed across the ability range of the student groups.

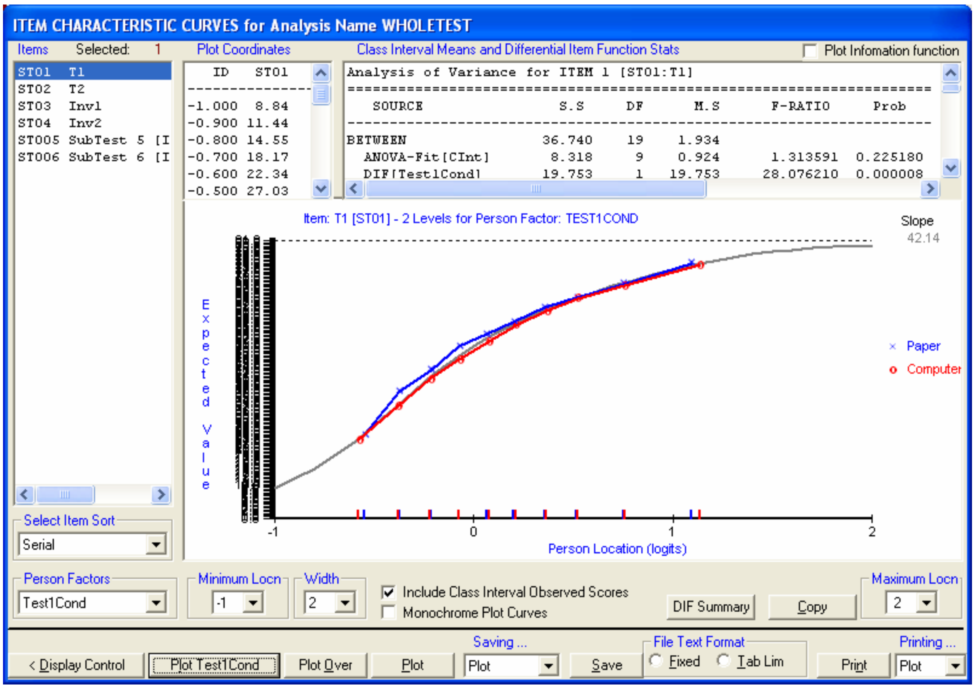
7.9 Overall performance and relationship to ability for Test A

For each of the two tests, Rasch analysis was used to represent performance on tests as whole and individual item characteristic curves.

The rounded DIF figure of P = 0.000 for Test A is shown in Figure 7 below. This indicates a highly significant differential performance (DIF) between students taking this test on paper (the blue line) and computer (the red line) in favour of the paper test.

In the case of Test A, as shown below in Figure 7, the item characteristic curves show an increased difference in performance (the Y axis) at the lower end of the ability range (the X axis) between the paper and computer modes.

Figure 7: Rasch item characteristic curves for the whole of Test A in paper and computer modes



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

```
ST01  T1
ST02  T2
ST03  Inv1
ST04  Inv2
ST005 SubTest 5 [I
ST006 SubTest 6 [I
```

```
  ID    ST01
--------------
-1.000  8.84
-0.900 11.44
-0.800 14.55
-0.700 18.17
-0.600 22.34
-0.500 27.03
```

```
Analysis of Variance for ITEM 1 [ST01:T1]
==================================================================
  SOURCE               S.S     DF      M.S      F-RATIO     Prob
------------------------------------------------------------------
BETWEEN              36.740    19    1.934
  ANOVA-Fit[CInt]     8.318     9    0.924     1.313591   0.225180
  DIF[Test1Cond]     19.753     1   19.753    28.076210   0.000008
```

Item: T1 [ST01] - 2 Levels for Person Factor: TEST1COND

Slope 42.14

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

□ Monochrome Plot Curves

Maximum Locn: 2

As described, there was a significant difference between the total test scores in paper and computer modes of Test A, and the items shown in Table 12 on page 122, show all the items within Test A which had the highest level of significant differences (significance where P< 0.005).

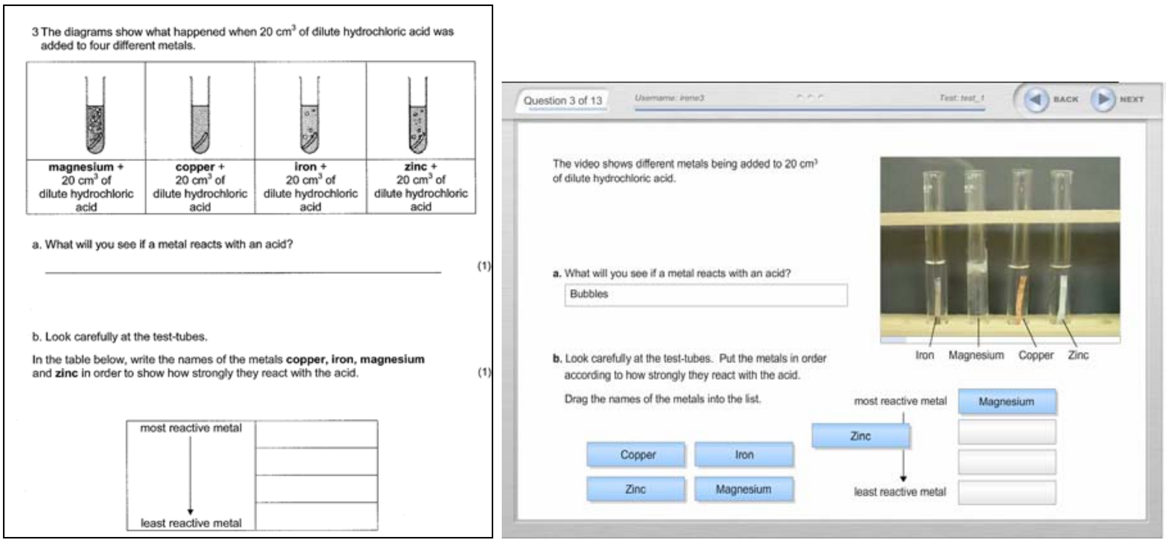
Discussion of the differences between modes will be carried out in Chapter 9. This section will indicate the nature of the differences for each DIF item within Tests A and B.

7.10 DIF items within Test A

Table 16: Performance data on Q3b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
3b	6BB	DD C V	0.87	0.61	0.316	0.242	0.000 P

Figure 8: Screenshots of Q3b in paper and computer modes



3 The diagrams show what happened when 20 cm³ of dilute hydrochloric acid was added to four different metals.

magnesium + 20 cm³ of dilute hydrochloric acid	**copper** + 20 cm³ of dilute hydrochloric acid	**iron** + 20 cm³ of dilute hydrochloric acid	**zinc** + 20 cm³ of dilute hydrochloric acid

a. What will you see if a metal reacts with an acid?

(1)

b. Look carefully at the test-tubes.

In the table below, write the names of the metals **copper, iron, magnesium** and **zinc** in order to show how strongly they react with the acid. (1)

most reactive metal	
least reactive metal	

Question 3 of 13

Username: irene3

Test: test_1

The video shows different metals being added to 20 cm³ of dilute hydrochloric acid.

a. What will you see if a metal reacts with an acid?

Bubbles

b. Look carefully at the test-tubes. Put the metals in order according to how strongly they react with the acid.

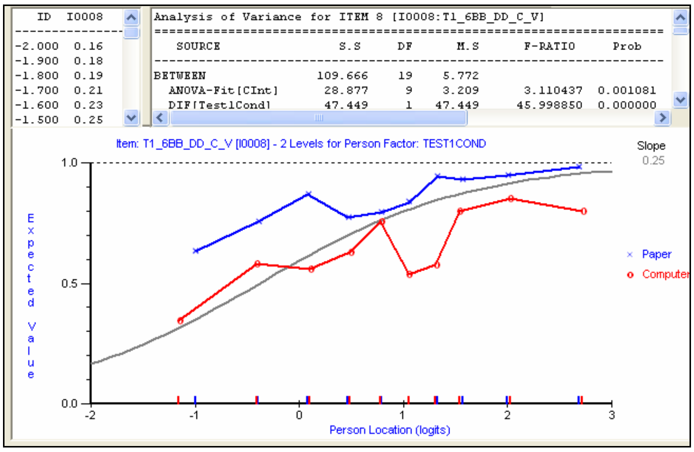
Drag the names of the metals into the list.

most reactive metal: Magnesium

least reactive metal

This item was targeted at a low level of difficulty. It included a video sequence of a chemistry experiment in the computer mode and a 2D experimental diagram in the paper version. Using either the diagram or the video sequence, students had to place the metals in order of reactivity. Facility values were high for the paper version (0.87), as expected, however they were much lower in the computer version (0.61). The discrimination value was not particularly high on paper, which was not surprising for an easy item, however the paper version performed more effectively in discriminating between higher and lower performing students than the computer version. The Rasch item characteristic curves for this item shown in Figure 9 showed a fairly consistent performance difference across the ability ranges of students in favour of the paper version.

Figure 9: Rasch item characteristic curves for Q3b in paper and computer modes



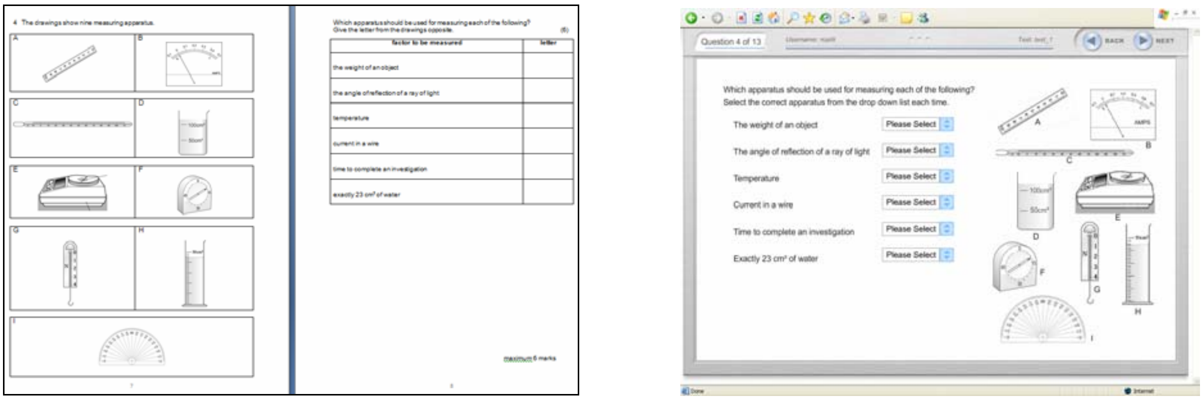
```
  ID   I0008
----------------
-2.000  0.16
-1.900  0.18
-1.800  0.19
-1.700  0.21
-1.600  0.23
-1.500  0.25
```

```
Analysis of Variance for ITEM 8 [I0008:T1_6BB_DD_C_V]
==========================================================================
   SOURCE                S.S      DF       M.S       F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                109.666    19     5.772
  ANOVA-Fit[CInt]       28.877     9     3.209      3.110437  0.001081
  DIF[Test1Cond]        47.449     1    47.449     45.998850  0.000000
```

Table 17: Performance data on Q4a1

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
4a1	8AA	DL S D	0.15	0.19	0.045	0.026	0.000 C

Figure 10: Screenshots of Q4a1 in paper and computer modes



4 The drawings show nine measuring apparatus.

Which apparatus should be used for measuring each of the following?
Give the letter from the drawings opposite. (6)

factor to be measured	letter
the weight of an object	
the angle of reflection of a ray of light	
temperature	
current in a wire	
time to complete an investigation	
exactly 23 cm³ of water	

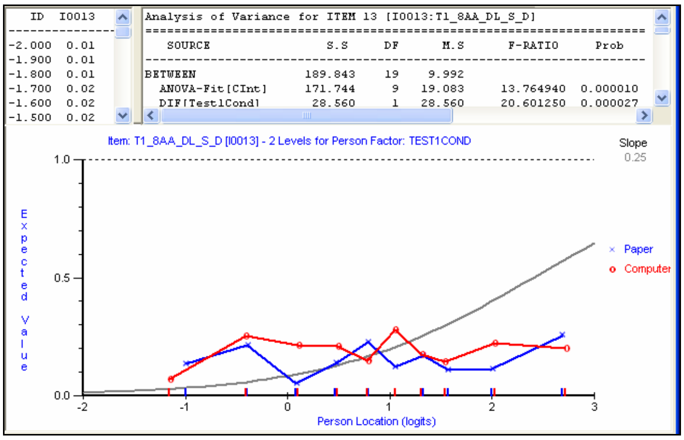
maximum 6 marks

Which apparatus should be used for measuring each of the following?
Select the correct apparatus from the drop down list each time.

The weight of an object	Please Select
The angle of reflection of a ray of light	Please Select
Temperature	Please Select
Current in a wire	Please Select
Time to complete an investigation	Please Select
Exactly 23 cm³ of water	Please Select

This item was targeted at a medium level of difficulty. It involved students choosing appropriate scientific measuring apparatus. The diagrams used in both modes were the same, therefore the only difference was in the answering mechanisms. On paper students had to write in a letter, on computer the letter was chosen from a drop down list. This item had very low facilities in both modes and therefore also low discrimination values. Although this item showed a DIF in favour of computer, the Rasch item characteristic curves in Figure 11 showed inconsistent performance differences across the student ability range

Figure 11: Rasch item characteristic curves for Q4a1 in paper and computer modes



```
  ID   I0013
----------------
-2.000   0.01
-1.900   0.01
-1.800   0.01
-1.700   0.02
-1.600   0.02
-1.500   0.02
```

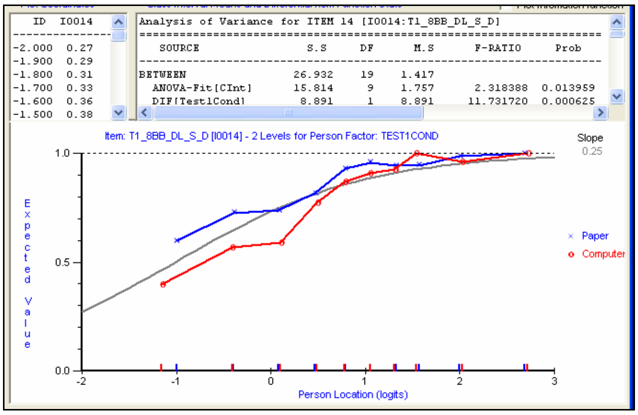
```
Analysis of Variance for ITEM 13 [I0013:T1_8AA_DL_S_D]
===========================================================================
  SOURCE                S.S      DF       M.S       F-RATIO     Prob
---------------------------------------------------------------------------
BETWEEN               189.843    19     9.992
  ANOVA-Fit[CInt]     171.744     9    19.083     13.764940  0.000010
  DIF[Test1Cond]       28.560     1    28.560     20.601250  0.000027
```

Table 18: Performance data on Q42

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
4a2	8BB	DL S D	0.89	0.74	0.413	0.477	0.001 P

This question was the second item within Question 4, choosing appropriate scientific measuring apparatus and was also targeted at a medium level of difficulty. The only difference in modes was the answering mechanism, on paper students writing in a letter and on computer, students choosing a letter from a drop down list. The facility on paper (0.89) was considerably higher than on computer (0.74). This item demonstrated good discrimination values in both modes, however the Rasch item characteristic curves shown in Figure 12 below shows greater modal performance differences at the lower end of the ability range, far less so as the ability of students increased.

Figure 12: Rasch item characteristic curves for Q4a2 in paper and computer modes



```
   ID   I0014
---------------
-2.000   0.27
-1.900   0.29
-1.800   0.31
-1.700   0.33
-1.600   0.36
-1.500   0.38
```

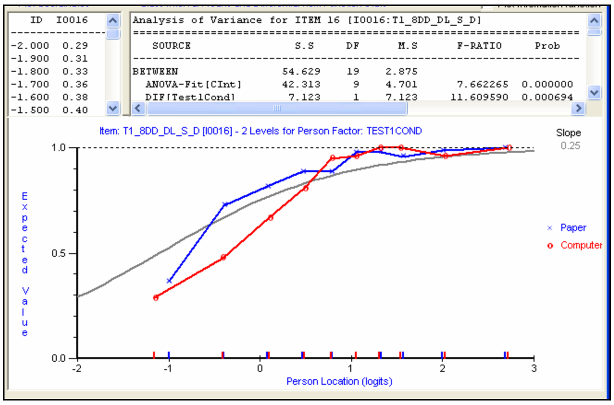
```
Analysis of Variance for ITEM 14 [I0014:T1_8BB_DL_S_D]
=====================================================================
   SOURCE               S.S      DF      M.S      F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN               26.932     19    1.417
  ANOVA-Fit[CInt]     15.814      9    1.757     2.318388   0.013959
  DIF[Test1Cond]       8.891      1    8.891    11.731720   0.000625
```

Table 19: Performance data on Q4a4

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
4a4	8DD	DL S D	0.89	0.75	0.483	0.579	0.002 P

This item was the fourth item within question 4; another item targeting a medium level of difficulty scientific apparatus choice. The facility value on paper (0.89) was significantly better than on computer (0.75), although both modes discriminated well. The Rasch item characteristic curves shown in Figure 13 below shows greater modal differences at the lower end of the ability range, and far less so as the ability of students increased.

Figure 13: Rasch item characteristic curves for Q4a4 in paper and computer modes



```
   ID   I0016
---------------
-2.000   0.29
-1.900   0.31
-1.800   0.33
-1.700   0.36
-1.600   0.38
-1.500   0.40
```

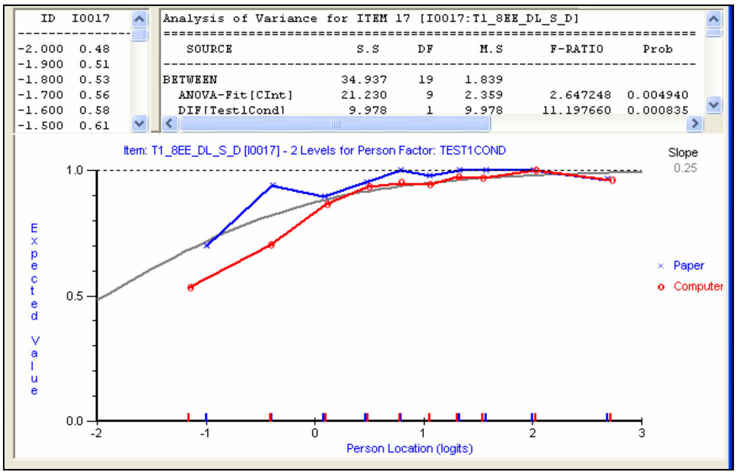
```
Analysis of Variance for ITEM 16 [I0016:T1_8DD_DL_S_D]
=====================================================================
   SOURCE               S.S      DF      M.S      F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN               54.629     19    2.875
  ANOVA-Fit[CInt]     42.313      9    4.701     7.662265   0.000000
  DIF[Test1Cond]       7.123      1    7.123    11.609590   0.000694
```

Table 20: Performance data on Q4a5

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
4a5	8EE	DL S D	0.95	0.85	0.325	0.421	0.000 P

This item was the last nested in Question 4, targeted at a low level of difficulty. The facility value on paper was very high on paper (0.95) not unexpectedly for an easy item, however the facility on computer was significantly poorer (0.85). The discrimination values were good in both modes. The Rasch item characteristic curves shown in Figure 14 below shows greater modal differences at the lower end of the ability range, and far less so as the ability of students increased.

Figure 14: Rasch item characteristic curves for Q4a5 in paper and computer modes



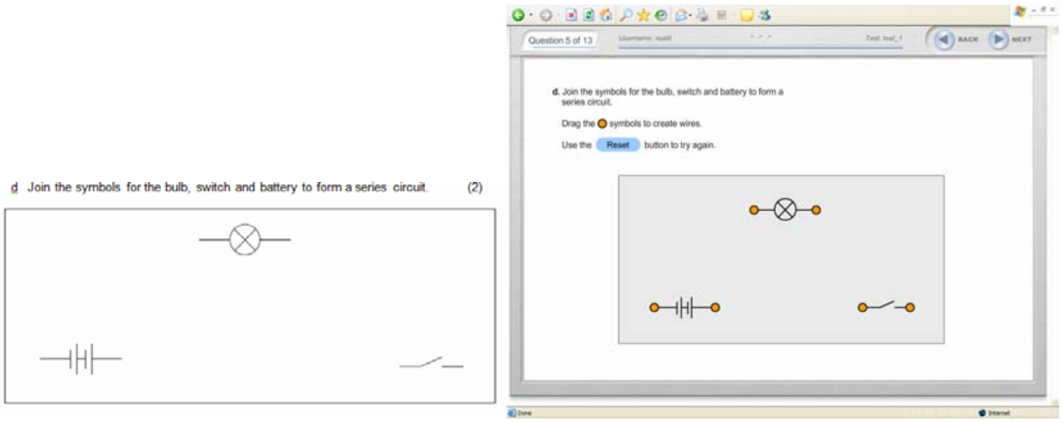
```
  ID   I0017
---------------
-2.000   0.48
-1.900   0.51
-1.800   0.53
-1.700   0.56
-1.600   0.58
-1.500   0.61
```

```
Analysis of Variance for ITEM 17 [I0017:T1_8EE_DL_S_D]
=======================================================================
   SOURCE                  S.S     DF      M.S       F-RATIO     Prob
-----------------------------------------------------------------------
BETWEEN                  34.937    19    1.839
  ANOVA-Fit[CInt]        21.230     9    2.359      2.647248  0.004940
  DIF[Test1Cond]          9.978     1    9.978     11.197660  0.000835
```

Table 21: Performance data on Q5d

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
5d	10DA	DR P D	0.47	0.75	0.316	0.186	0.000 C

Figure 15: Screenshots of Q5d in paper and computer modes



d Join the symbols for the bulb, switch and battery to form a series circuit. (2)

Question 5 of 13

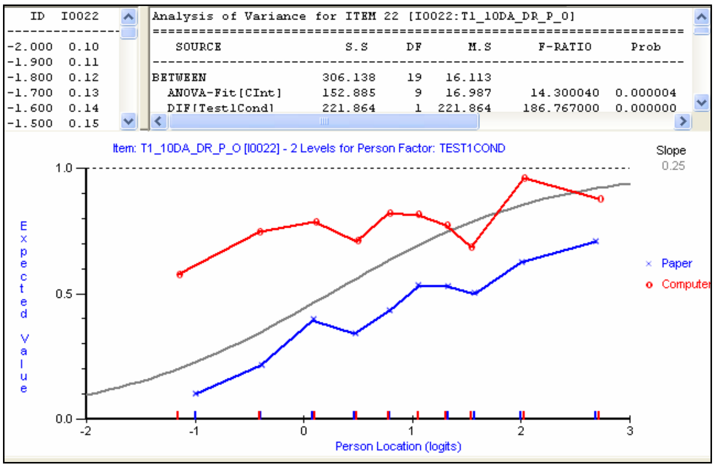
d. Join the symbols for the bulb, switch and battery to form a series circuit.

Drag the ● symbols to create wires.

Use the Reset button to try again.

This item was targeted at a low level of difficulty, involving students drawing a series circuit diagram on paper, and using a mouse to draw series circuit connections on-screen. The computer facility value was reasonably high on computer (0.75) and disappointingly low on paper (0.47). The item was however far more discriminating on paper than on computer. The Rasch item characteristic curves shown below in Figure 16 showed a fairly consistent performance difference across the ability ranges of students in favour of the computer version.

Figure 16: Rasch item characteristic curves for Q5d in paper and computer modes



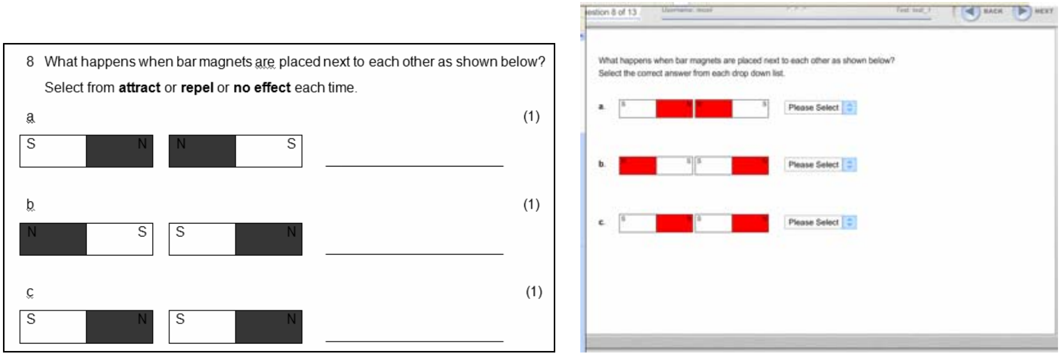
```
   ID   I0022
----------------
-2.000   0.10
-1.900   0.11
-1.800   0.12
-1.700   0.13
-1.600   0.14
-1.500   0.15
```

```
Analysis of Variance for ITEM 22 [I0022:T1_10DA_DR_P_O]
=====================================================================
   SOURCE               S.S      DF       M.S      F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN               306.138    19    16.113
  ANOVA-Fit[CInt]     152.885     9    16.987    14.300040  0.000004
  DIF[Test1Cond]      221.864     1   221.864   186.767000  0.000000
```

Table 22: Performance data on Q8a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
8a	14AA	DL P CP	0.93	0.81	0.464	0.454	0.003 P

Figure 17: Screenshots of Q8a in paper and computer modes



8 What happens when bar magnets are placed next to each other as shown below?

Select from **attract** or **repel** or **no effect** each time.

a (1)

____

b (1)

____

c (1)

____

What happens when bar magnets are placed next to each other as shown below?
Select the correct answer from each drop down list.

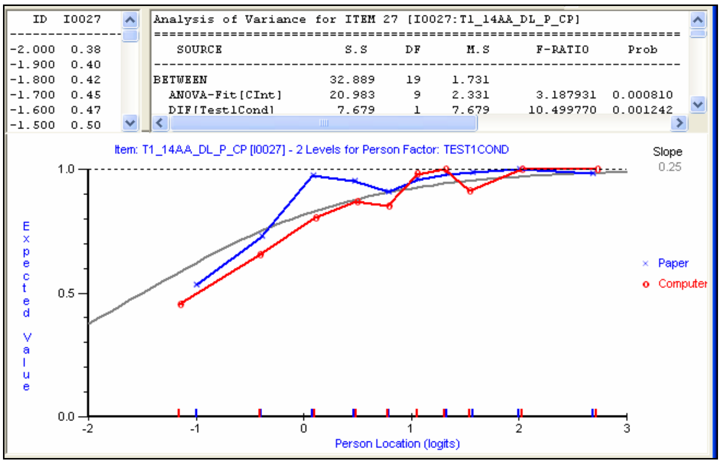
a. Please Select

b. Please Select

c. Please Select

This item, together with the following two, were targeted at a low level of difficulty and involved students deciding what would happen when particular poles of magnets were placed near each other. On paper, the diagrams were in black and white and students had to write their answer onto the paper. On computer the diagrams were in colour and students had to choose their response from a drop down list. The facility value on paper was significantly higher than on computer, although this item discriminated well in both modes. The Rasch item characteristic curves for this item shown below in Figure 18 showed a fairly consistently pattern of performance across the ability range favouring paper.

Figure 18: Rasch item characteristic curves for Q8a in paper and computer modes



```
  ID   I0027
---------------
-2.000  0.38
-1.900  0.40
-1.800  0.42
-1.700  0.45
-1.600  0.47
-1.500  0.50
```

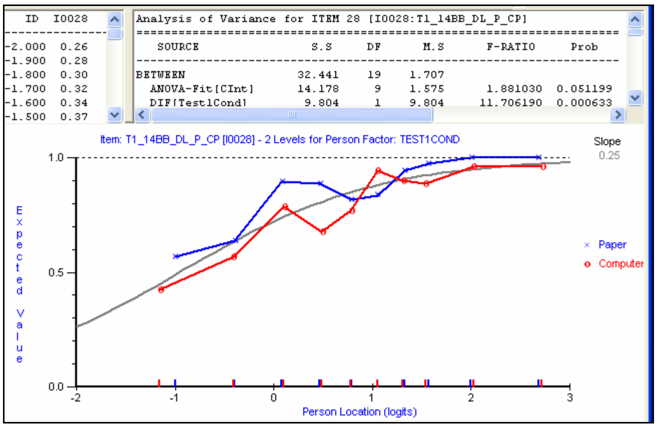
```
Analysis of Variance for ITEM 27 [I0027:T1_14AA_DL_P_CP]
=====================================================================
   SOURCE                 S.S      DF      M.S       F-RATIO    Prob
---------------------------------------------------------------------
BETWEEN                 32.889     19     1.731
  ANOVA-Fit[CInt]       20.983      9     2.331     3.187931  0.000810
  DIF[Test1Cond]         7.679      1     7.679    10.499770  0.001242
```

Table 23: Performance data on Q8b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
8b	14BB	DL P CP	0.88	0.74	0.430	0.412	0.002 P

The second magnet item in Question 8 was also targeted at a low level of difficulty and performed similarly to 8a. The facility value for paper was significantly higher (0.88) than on computer (0.74), although this item discriminated well in both modes. The Rasch item characteristic curves for this item shown below in Figure 19 showed a fairly consistently pattern of performance across the ability range favouring paper.

Figure 19: Rasch item characteristic curves for Q8b in paper and computer modes



```
  ID   I0028
---------------
-2.000   0.26
-1.900   0.28
-1.800   0.30
-1.700   0.32
-1.600   0.34
-1.500   0.37
```

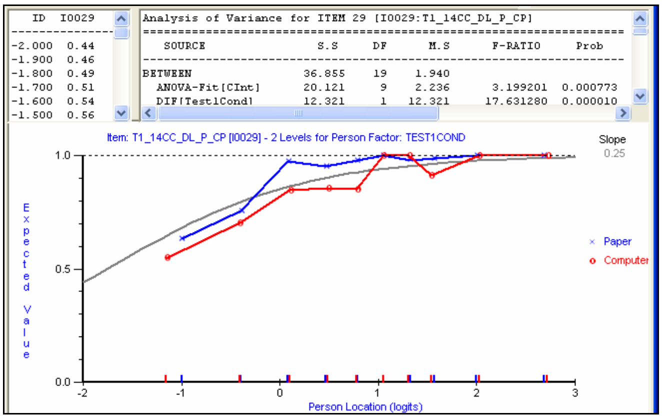
```
Analysis of Variance for ITEM 28 [I0028:T1_14BB_DL_P_CP]
==============================================================================
   SOURCE                  S.S      DF       M.S        F-RATIO      Prob
------------------------------------------------------------------------------
BETWEEN                  32.441     19     1.707
  ANOVA-Fit[CInt]        14.178      9     1.575       1.881030   0.051199
  DIF[Test1Cond]          9.804      1     9.804      11.706190   0.000633
```

Table 24: Performance data on Q8c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
8c	14CC	DL P CP	0.95	0.84	0.451	0.411	0.001 P

The third magnet item in Question 8 performed similarly to the previous two linked items. The facility value on paper (0.95) was significantly higher on computer (0.84), both discrimination values were good and the Rasch item characteristic curves shown below in Figure 20 show a fairly consistently pattern of performance across the ability range favouring paper

Figure 20: Rasch item characteristic curves for Q8c in paper and computer modes



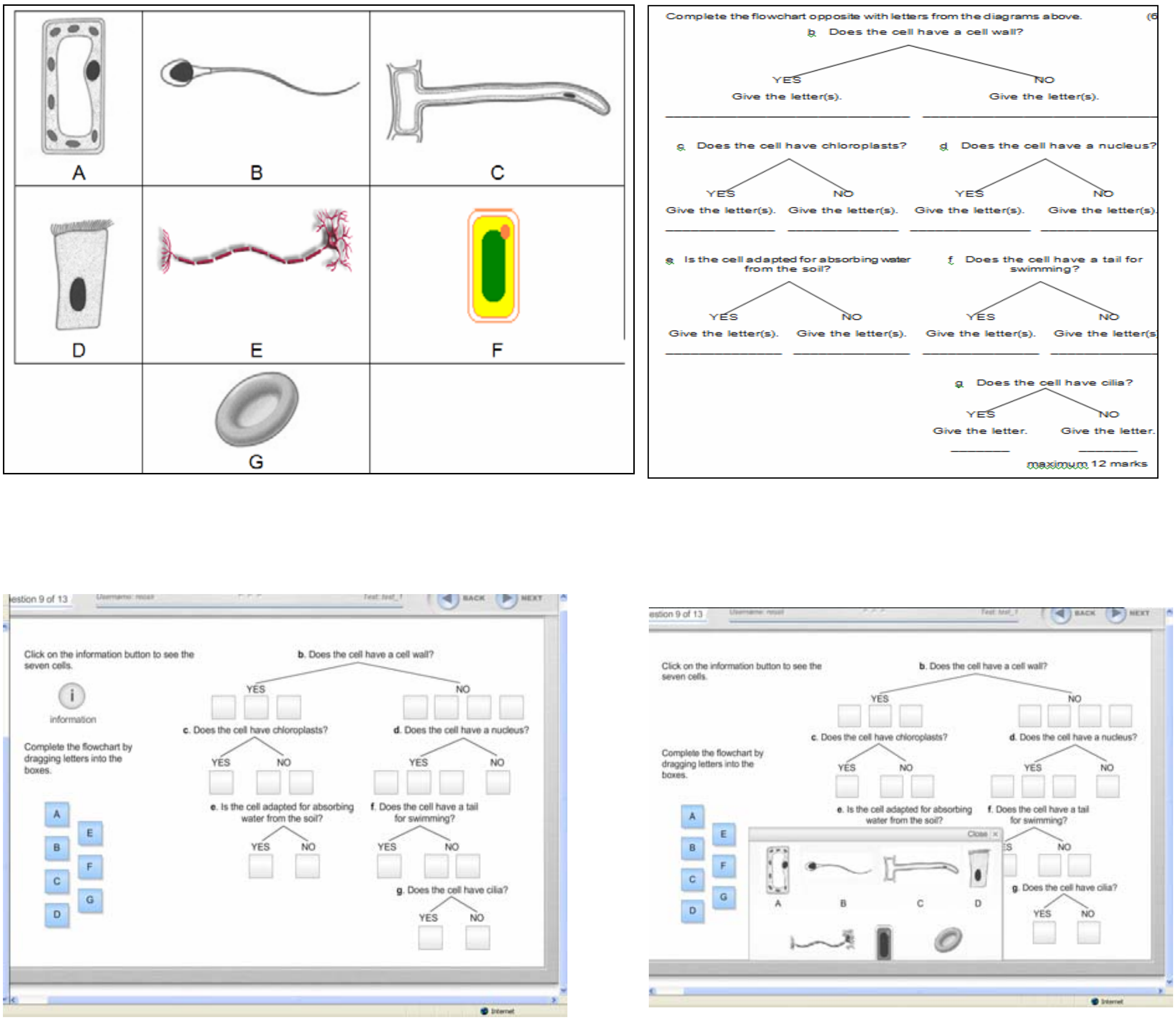
```
  ID   I0029
---------------
-2.000   0.44
-1.900   0.46
-1.800   0.49
-1.700   0.51
-1.600   0.54
-1.500   0.56
```

```
Analysis of Variance for ITEM 29 [I0029:T1_14CC_DL_P_CP]
==============================================================================
   SOURCE                  S.S      DF       M.S        F-RATIO      Prob
------------------------------------------------------------------------------
BETWEEN                  36.855     19     1.940
  ANOVA-Fit[CInt]        20.121      9     2.236       3.199201   0.000773
  DIF[Test1Cond]         12.321      1    12.321      17.631280   0.000010
```

Table 25: Performance data on Q9b

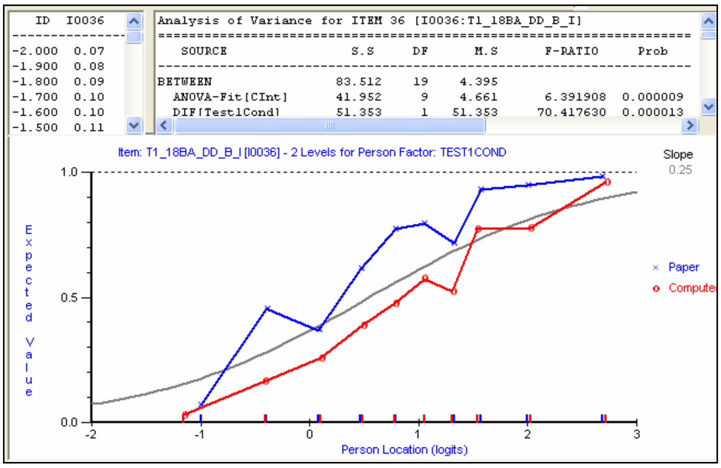
Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
9b	18BA	DD B D	0.73	0.40	0.562	0.517	0.000 P

Figure 21: Screenshots of Q9b and g in paper and computer mode


A
B
C
D
E
F
G
Complete the flowchart opposite with letters from the diagrams above.
(6
b. Does the cell have a cell wall?
YES
Give the letter(s).
NO
Give the letter(s).
c. Does the cell have chloroplasts?
YES
Give the letter(s).
NO
Give the letter(s).
d. Does the cell have a nucleus?
YES
Give the letter(s).
NO
Give the letter(s).
e. Is the cell adapted for absorbing water from the soil?
YES
Give the letter(s).
NO
Give the letter(s).
f. Does the cell have a tail for swimming?
YES
Give the letter(s).
NO
Give the letter(s)
g. Does the cell have cilia?
YES
Give the letter.
NO
Give the letter.
maximum 12 marks
Question 9 of 13
BACK
NEXT
Click on the information button to see the seven cells.
information
Complete the flowchart by dragging letters into the boxes.
Close
Internet

This item (9b) and the following one shown here (9g) were targeted at a medium and high levels of difficulty respectively and involved students using information to identify biological cells. On paper the diagramatical information and question were presented on a double page spread, whereas on computer the information was accessed using an information box mechanism, which could be opened up, moved around the screen or minimised as required. The paper facility (0.73) was significantly higher than on computer (0.40), although both modes discriminated well. The Rasch item characteristic curves shown below in Figure 22 showed that the lowest ability students found this item challenging in both modes, however, from this point students across the ability range performed much better on paper.

Figure 22: Rasch item characteristic curves for Q9b in paper and computer modes



```
  ID   I0036
----------------
-2.000   0.07
-1.900   0.08
-1.800   0.09
-1.700   0.10
-1.600   0.10
-1.500   0.11
```

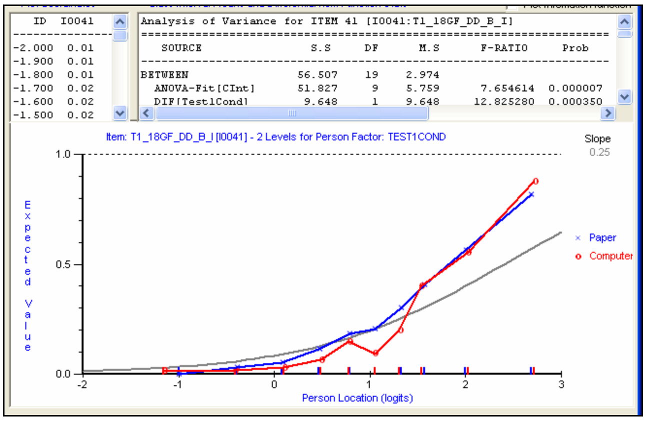
```
Analysis of Variance for ITEM 36 [I0036:T1_18BA_DD_B_I]
=====================================================================
   SOURCE                   S.S      DF       M.S        F-RATIO      Prob
---------------------------------------------------------------------
BETWEEN                   83.512     19     4.395
  ANOVA-Fit[CInt]         41.952      9     4.661       6.391908   0.000009
  DIF[Test1Cond]          51.353      1    51.353      70.417630   0.000013
```

Table 26: Performance data on Q9g

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
9g	18GF	DD B I	0.32	0.15	0.462	0.459	0.005 P

This item was the last part of the cell identification classification question and targeted at a high level of difficulty. The item was challenging, however the facility value on paper (0.32) was much higher than on computer (0.15). Both modes discriminated well. The Rasch item characteristic curves shown below in Figure 23 shows a fairly consistent pattern of performance across the student ability range favouring paper.

Figure 23: Rasch item characteristic curves for Q9g in paper and computer modes



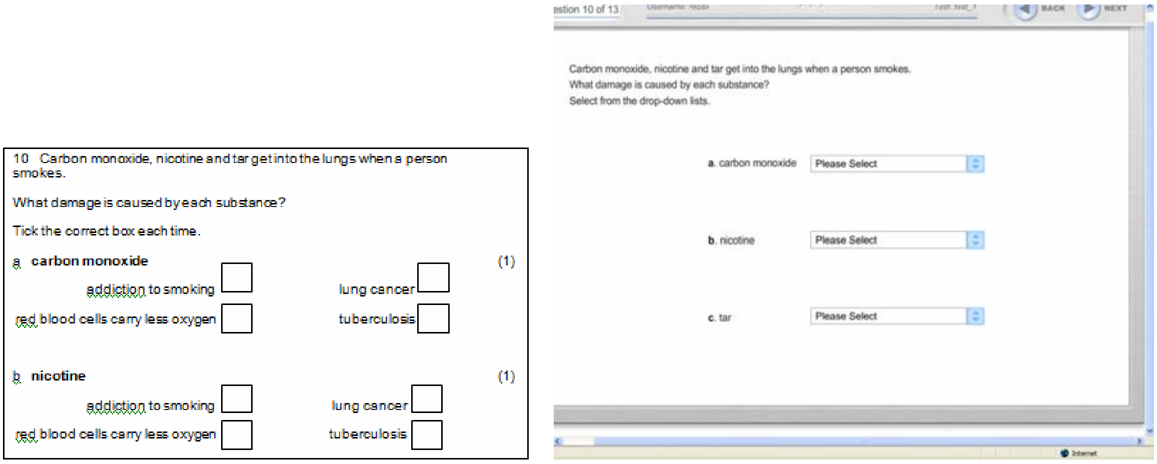
```
  ID   I0041
----------------
-2.000   0.01
-1.900   0.01
-1.800   0.01
-1.700   0.02
-1.600   0.02
-1.500   0.02
```

```
Analysis of Variance for ITEM 41 [I0041:T1_18GF_DD_B_I]
=====================================================================
   SOURCE                   S.S      DF       M.S        F-RATIO      Prob
---------------------------------------------------------------------
BETWEEN                   56.507     19     2.974
  ANOVA-Fit[CInt]         51.827      9     5.759       7.654614   0.000007
  DIF[Test1Cond]           9.648      1     9.648      12.825280   0.000350
```

Table 27: Performance data on Q10b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
10b	19BB	DD P	0.87	0.66	0.529	0.542	0.000 P

Figure 24: Screenshots of Q10b in paper and computer mode



10 Carbon monoxide, nicotine and tar get into the lungs when a person smokes.

What damage is caused by each substance?

Tick the correct box each time.

a carbon monoxide (1)

addiction to smoking □ lung cancer □

red blood cells carry less oxygen □ tuberculosis □

b nicotine (1)

addiction to smoking □ lung cancer □

red blood cells carry less oxygen □ tuberculosis □

Carbon monoxide, nicotine and tar get into the lungs when a person smokes.
What damage is caused by each substance?
Select from the drop-down lists.

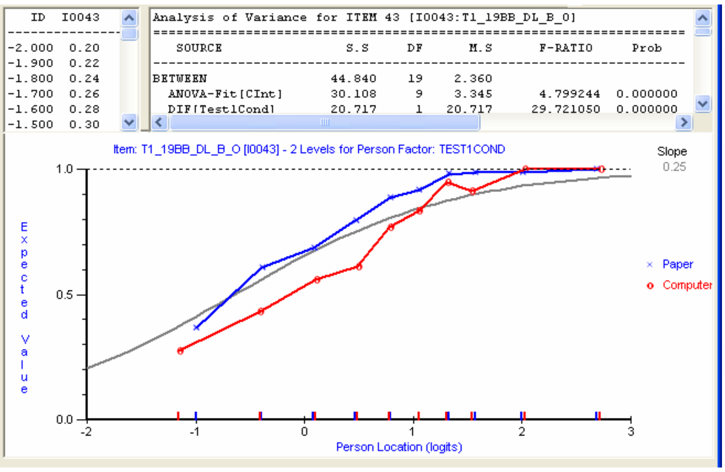
a. carbon monoxide Please Select

b. nicotine Please Select

c. tar Please Select

This item was targeted at a medium level of difficulty and involved students selecting an answer from four options about the effects of nicotine. The only difference in mode was the answering mechanism where on paper students ticked a box and on computer they chose an option from a drop down list. The facility value on paper (0.87) was significantly higher than on computer (0.66). The discrimination values in both modes were very good. The Rasch item characteristic curves shown below in Figure 25 showed a fairly consistent pattern of performance difference across the student ability range in favour of paper.

Figure 25: Rasch item characteristic curves for Q10b in paper and computer modes



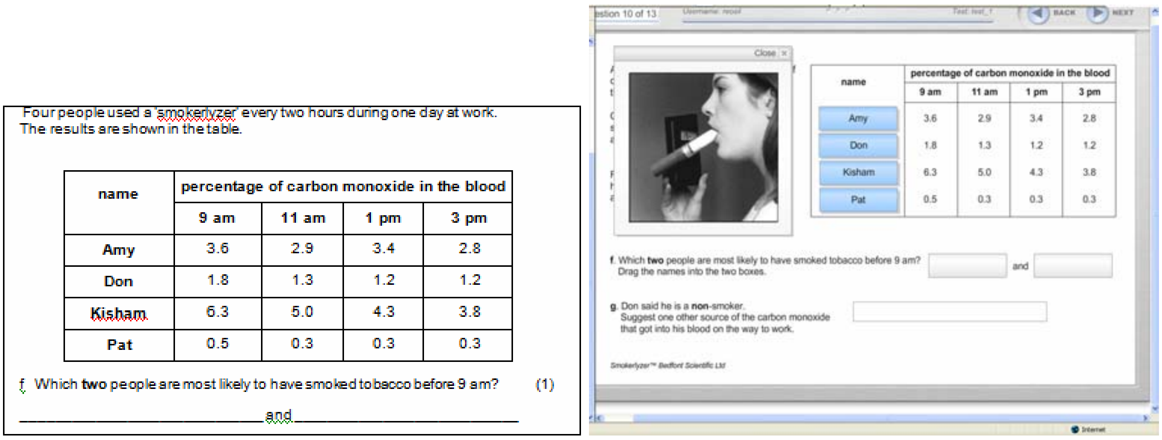
```
  ID   I0043
---------------
-2.000  0.20
-1.900  0.22
-1.800  0.24
-1.700  0.26
-1.600  0.28
-1.500  0.30
```

```
Analysis of Variance for ITEM 43 [I0043:T1_19BB_DL_B_0]
=========================================================================
   SOURCE                 S.S      DF      M.S       F-RATIO     Prob
-------------------------------------------------------------------------
BETWEEN                 44.840     19     2.360
  ANOVA-Fit[CInt]       30.108      9     3.345      4.799244  0.000000
  DIF[Test1Cond]        20.717      1    20.717     29.721050  0.000000
```

Table 28: Performance data on Q10f

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
10f	21FA	OR1 B I	0.88	0.68	0.541	0.547	0.000 P

Figure 26: Screenshots of Q10f in paper and computer mode



Four people used a 'smokerlyzer' every two hours during one day at work.
The results are shown in the table.

name	percentage of carbon monoxide in the blood			
	9 am	11 am	1 pm	3 pm
Amy	3.6	2.9	3.4	2.8
Don	1.8	1.3	1.2	1.2
Kisham	6.3	5.0	4.3	3.8
Pat	0.5	0.3	0.3	0.3

f Which **two** people are most likely to have smoked tobacco before 9 am? (1)

________________________ and ________________________

name	percentage of carbon monoxide in the blood			
	9 am	11 am	1 pm	3 pm
Amy	3.6	2.9	3.4	2.8
Don	1.8	1.3	1.2	1.2
Kisham	6.3	5.0	4.3	3.8
Pat	0.5	0.3	0.3	0.3

f. Which **two** people are most likely to have smoked tobacco before 9 am?
Drag the names into the two boxes.

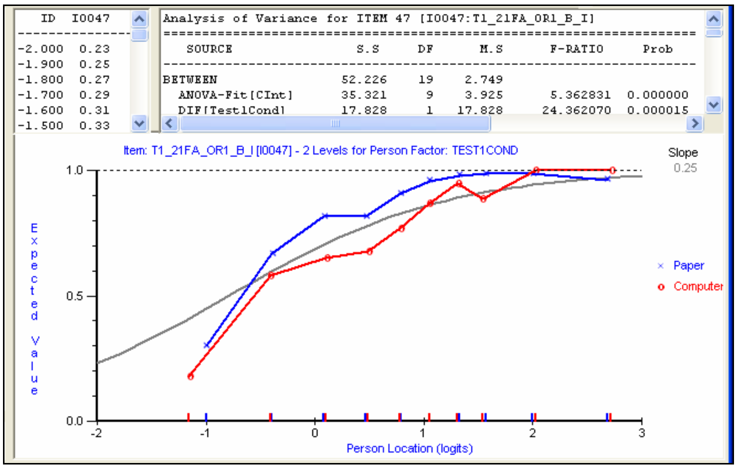
and

g. Don said he is a **non**-smoker.
Suggest one other source of the carbon monoxide that got into his blood on the way to work.

Smokerlyzer™ Bedfont Scientific Ltd

This item was targeted at a medium level of difficulty and involved students interpreting information from a given table. The same table was shown in paper and computer versions. The only difference was that on paper students had write two names from the table, on computer they dragged two names from the table. The facility value on paper was significantly higher (0.88) than on computer (0.68), although both modes discriminated very well. The Rasch item characteristic curves shown below in Figure 25 showed a fairly consistent difference in performance across the student ability range favouring paper.

Figure 27: Rasch item characteristic curves for Q10f in paper and computer modes



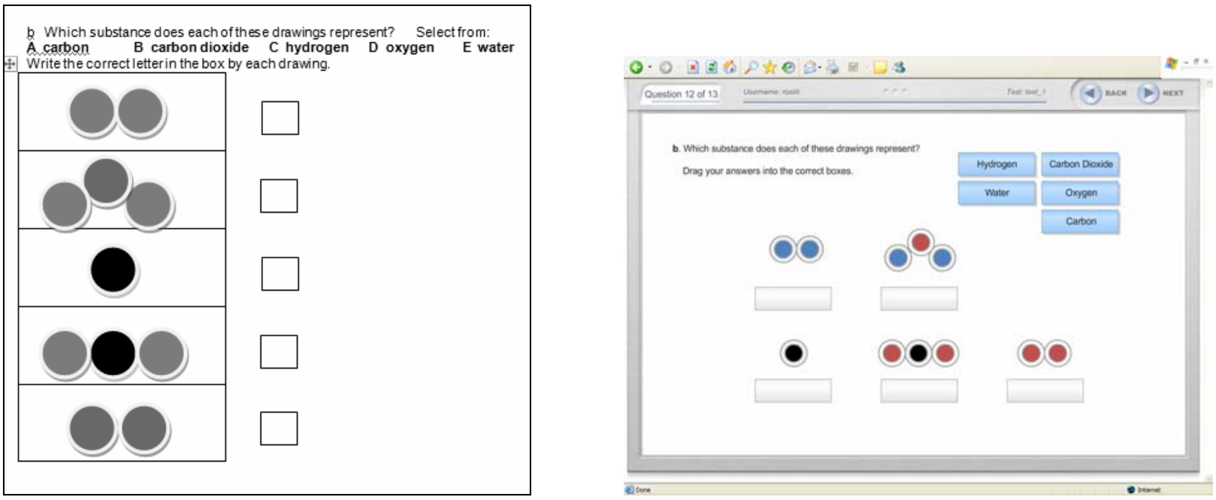
```
  ID   I0047
---------------
-2.000  0.23
-1.900  0.25
-1.800  0.27
-1.700  0.29
-1.600  0.31
-1.500  0.33
```

```
Analysis of Variance for ITEM 47 [I0047:T1_21FA_OR1_B_I]
============================================================================
  SOURCE                  S.S       DF       M.S        F-RATIO      Prob
----------------------------------------------------------------------------
BETWEEN                 52.226     19     2.749
  ANOVA-Fit[CInt]       35.321      9     3.925       5.362831   0.000000
  DIF[Test1Cond]        17.828      1    17.828      24.362070   0.000015
```

Table 29: Performance data on Q12b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
12b	25BA	DD C CP	0.35	0.33	0.458	0.518	0.000 C

Figure 28: Screenshots of Q12b in paper and computer mode



b Which substance does each of these drawings represent? Select from:
A carbon B carbon dioxide C hydrogen D oxygen E water
Write the correct letter in the box by each drawing.

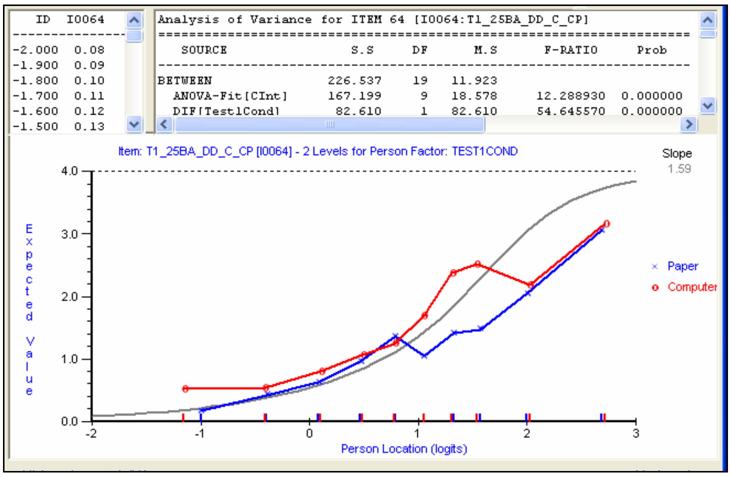
Question 12 of 13

b. Which substance does each of these drawings represent?

Drag your answers into the correct boxes.

This item was targeted at a high level of difficulty and carried 4 marks. The item involved students identifying chemical substances. On paper, diagramatic information was presented in black and white images whereas on computer the diagrams were shown in colour. The difference in response mode was that on paper students had to write letters into five boxes and on computer students dragged five responses name boxes under the correct chemical substance. The overall facility values across modes were similar, however the Rasch item characteristic curves shown below in Figure 29 show the computer versions performing much better for the more able students.

Figure 29: Rasch item characteristic curves for Q10f in paper and computer modes



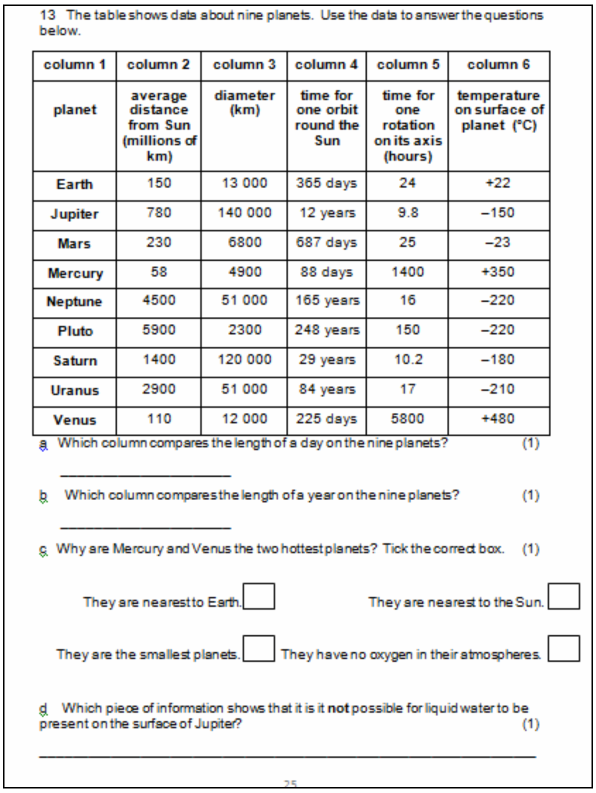
```
  ID   I0064
---------------
-2.000  0.08
-1.900  0.09
-1.800  0.10
-1.700  0.11
-1.600  0.12
-1.500  0.13
```

```
Analysis of Variance for ITEM 64 [I0064:T1_25BA_DD_C_CP]
=========================================================================
   SOURCE                S.S      DF      M.S       F-RATIO     Prob
-------------------------------------------------------------------------
BETWEEN                226.537    19   11.923
  ANOVA-Fit[CInt]      167.199     9   18.578     12.288930   0.000000
  DIF[Test1Cond]        82.610     1   82.610     54.645570   0.000000
```

Table 30: Performance data on Q13a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13a	26AA	DL P I	0.74	0.45	0.705	0.671	0.000 P

Figure 30: Screenshots of Q13a, b and d in paper and computer mode



13 The table shows data about nine planets. Use the data to answer the questions below.

column 1	column 2	column 3	column 4	column 5	column 6
planet	average distance from Sun (millions of km)	diameter (km)	time for one orbit round the Sun	time for one rotation on its axis (hours)	temperature on surface of planet (°C)
Earth	150	13 000	365 days	24	+22
Jupiter	780	140 000	12 years	9.8	–150
Mars	230	6800	687 days	25	–23
Mercury	58	4900	88 days	1400	+350
Neptune	4500	51 000	165 years	16	–220
Pluto	5900	2300	248 years	150	–220
Saturn	1400	120 000	29 years	10.2	–180
Uranus	2900	51 000	84 years	17	–210
Venus	110	12 000	225 days	5800	+480

a Which column compares the length of a day on the nine planets? (1)

____

b Which column compares the length of a year on the nine planets? (1)

____

c Why are Mercury and Venus the two hottest planets? Tick the correct box. (1)

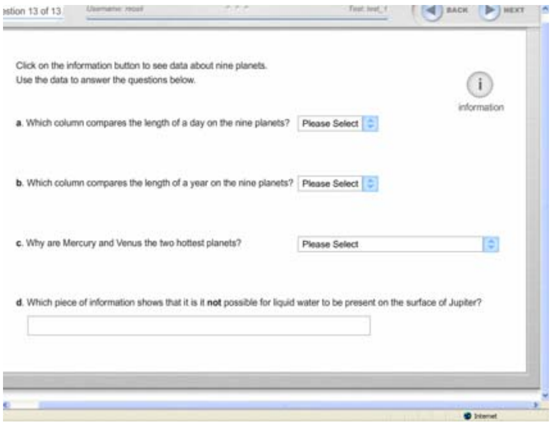
They are nearest to Earth. □ They are nearest to the Sun. □

They are the smallest planets. □ They have no oxygen in their atmospheres. □

d Which piece of information shows that it is it **not** possible for liquid water to be present on the surface of Jupiter? (1)

____

25



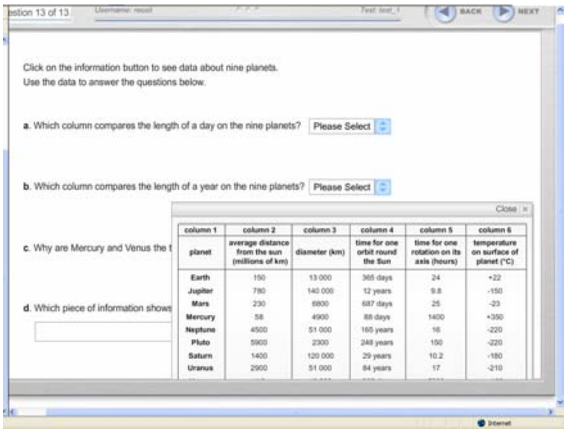
Click on the information button to see data about nine planets.
Use the data to answer the questions below.

a. Which column compares the length of a day on the nine planets? Please Select

b. Which column compares the length of a year on the nine planets? Please Select

c. Why are Mercury and Venus the two hottest planets? Please Select

d. Which piece of information shows that it is it **not** possible for liquid water to be present on the surface of Jupiter?



Click on the information button to see data about nine planets.
Use the data to answer the questions below.

a. Which column compares the length of a day on the nine planets? Please Select

b. Which column compares the length of a year on the nine planets? Please Select

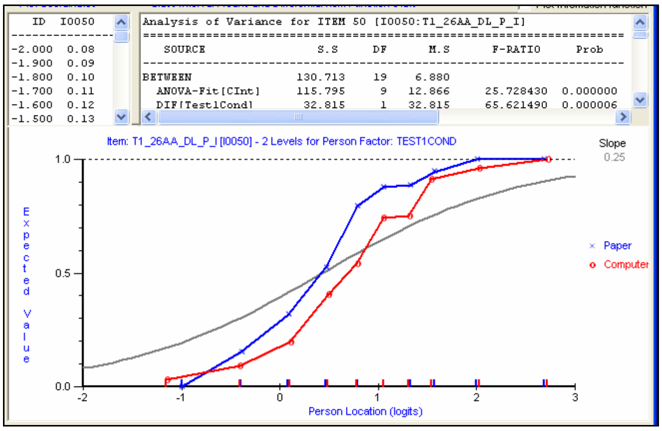
c. Why are Mercury and Venus the t

d. Which piece of information shows

column 1	column 2	column 3	column 4	column 5	column 6
planet	average distance from the sun (millions of km)	diameter (km)	time for one orbit round the Sun	time for one rotation on its axis (hours)	temperature on surface of planet (°C)
Earth	150	13 000	365 days	24	+22
Jupiter	780	140 000	12 years	9.8	-150
Mars	230	6800	687 days	25	-23
Mercury	58	4900	88 days	1400	+350
Neptune	4500	51 000	165 years	16	-220
Pluto	5900	2300	248 years	150	-220
Saturn	1400	120 000	29 years	10.2	-180
Uranus	2900	51 000	84 years	17	-210

This item, together with the following three items, were targeted at a high level of difficulty and involved students analysing astronomical numeric information. On paper the diagramatical information and question were presented on a double page spread, whereas on computer the information was accessed using a pop up information box mechanism, which could be opened up, moved around the screen or minimised as required. The answering mechanism on paper was a written selection of a column and on computer a drop down list choice from four options. The paper facility value for this item (0.74) was significantly higher than the computer version (0.45). The discrimination values for both modes were very high, which is not unusual for difficult items. The Rasch item characteristic curves shown below in Figure 31 showed a fairly consistent difference in performance across the student ability range favouring paper.

Figure 31: Rasch item characteristic curves for Q13a in paper and computer modes



```
  ID   I0050
----------------
-2.000   0.08
-1.900   0.09
-1.800   0.10
-1.700   0.11
-1.600   0.12
-1.500   0.13
```

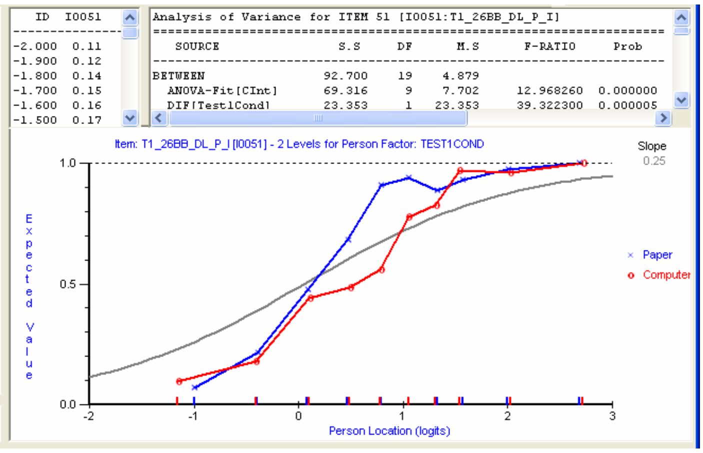
```
Analysis of Variance for ITEM 50 [I0050:T1_26AA_DL_P_I]
==========================================================================
  SOURCE                   S.S      DF       M.S       F-RATIO      Prob
--------------------------------------------------------------------------
BETWEEN                  130.713    19     6.880
  ANOVA-Fit[CInt]        115.795     9    12.866      25.728430  0.000000
  DIF[Test1Cond]          32.815     1    32.815      65.621490  0.000006
```

Table 31: Performance data on Q13b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13b	26BB	DL P I	0.78	0.53	0.667	0.638	0.000 P

This item 13b, the second part of the astronomical data question performed similarly to the first item, 13a. The stimuli and response mechanisms were also the same as in 13a. The facility value on paper (0.78) was significantly higher than on computer (0.53), with both modes discriminated very well. The Rasch item characteristic curves shown below in Figure 32 showed a fairly consistent difference in performance across the student ability range favouring paper.

Figure 32: Rasch item characteristic curves for Q13b in paper and computer modes



```
  ID   I0051
----------------
-2.000   0.11
-1.900   0.12
-1.800   0.14
-1.700   0.15
-1.600   0.16
-1.500   0.17
```

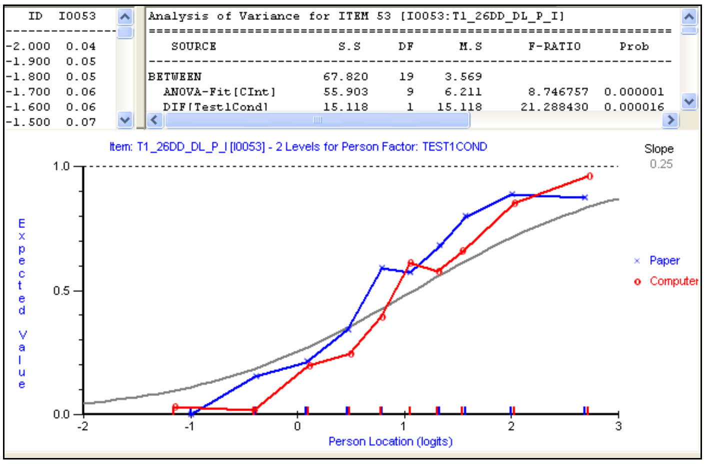
```
Analysis of Variance for ITEM 51 [I0051:T1_26BB_DL_P_I]
==========================================================================
  SOURCE                   S.S      DF       M.S       F-RATIO      Prob
--------------------------------------------------------------------------
BETWEEN                   92.700    19     4.879
  ANOVA-Fit[CInt]         69.316     9     7.702      12.968260  0.000000
  DIF[Test1Cond]          23.353     1    23.353      39.322300  0.000005
```

Table 32: Performance data on Q13d

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13d	26DD	OR2 P I	0.59	0.35	0. 558	0.570	0.000 P

13d was targeted at a high level of difficulty and involved students writing an open response, using information from a table on a double page spread on paper and a pop up information box on the computer version. The facility value on paper (0.59) was significantly higher than on computer (0.35), although the discriminations were very good in both modes. The Rasch item characteristic curves shown below in Figure 33 showed a fairly consistent difference in performance across the student ability range in favour of paper

Figure 33: Rasch item characteristic curves for Q13bin paper and computer modes



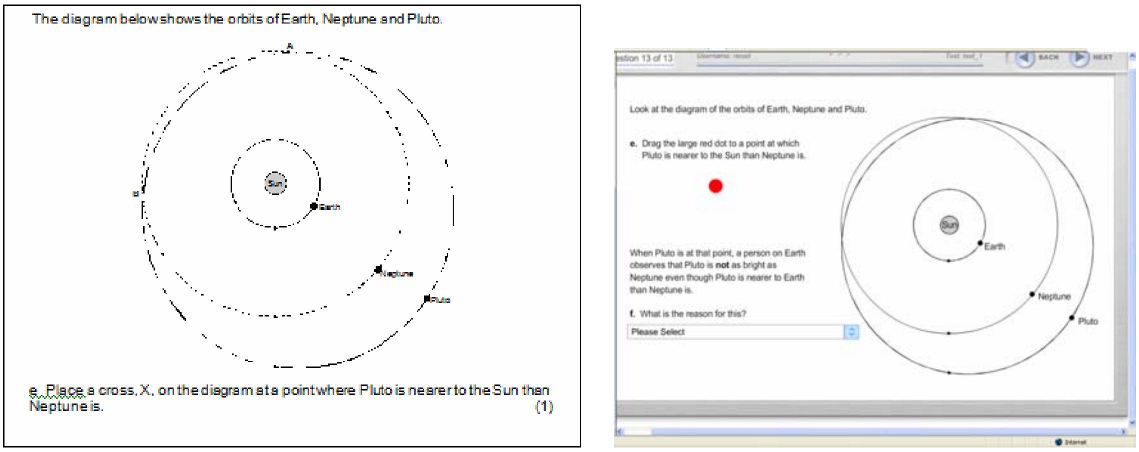
```
  ID   I0053
----------------
-2.000  0.04
-1.900  0.05
-1.800  0.05
-1.700  0.06
-1.600  0.06
-1.500  0.07
```

```
Analysis of Variance for ITEM 53 [I0053:T1_26DD_DL_P_I]
==========================================================================
  SOURCE                S.S     DF      M.S       F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                67.820   19     3.569
  ANOVA-Fit[CInt]      55.903    9     6.211     8.746757   0.000001
  DIF[Test1Cond]       15.118    1    15.118    21.288430   0.000016
```

Table 33: Performance data on Q13e

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13e	27EA	DD P I	0.67	0.40	0.649	0.606	0.000 P

Figure 34: Screenshots of Q13e in paper and computer mode



The diagram below shows the orbits of Earth, Neptune and Pluto.

e. Place a cross, X, on the diagram at a point where Pluto is nearer to the Sun than Neptune is. (1)

Look at the diagram of the orbits of Earth, Neptune and Pluto.

e. Drag the large red dot to a point at which Pluto is nearer to the Sun than Neptune is.

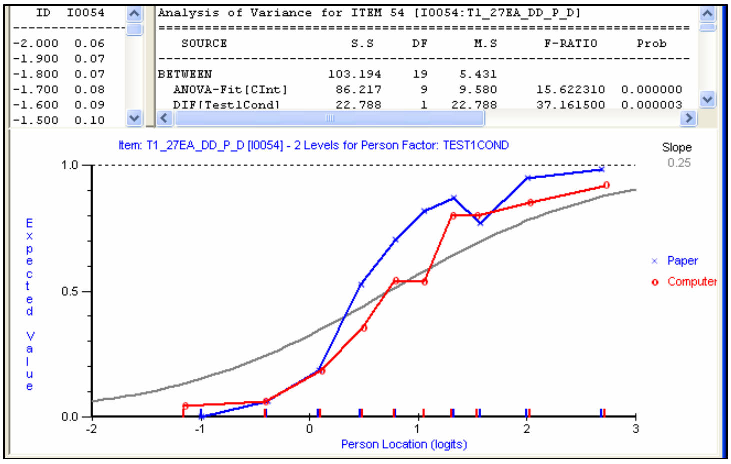
When Pluto is at that point, a person on Earth observes that Pluto is **not** as bright as Neptune even though Pluto is nearer to Earth than Neptune is.

f. What is the reason for this?

Please Select

This item was targeted at a high of difficulty and involved students identifying the position of a planet at particular point of an orbit. On paper, they had to draw where they thought the Pluto would be at a position of its orbit, whereas on computer, they had to drag a red dot to their chosen position. The facility value on paper (0.67) was significantly higher than on computer (0.40) although the discriminations in both modes were very good. The Rasch item characteristic curves shown below in Figure 35 showed a fairly consistent difference in performance across the student ability range in favour of paper

Figure 35: Rasch item characteristic curves for Q13e in paper and computer modes



```
  ID   I0054
---------------
-2.000   0.06
-1.900   0.07
-1.800   0.07
-1.700   0.08
-1.600   0.09
-1.500   0.10
```

```
Analysis of Variance for ITEM 54 [I0054:T1_27EA_DD_P_D]
======================================================================
   SOURCE                    S.S       DF       M.S        F-RATIO      Prob
----------------------------------------------------------------------
BETWEEN                    103.194     19     5.431
  ANOVA-Fit[CInt]           86.217      9     9.580      15.622310   0.000000
  DIF[Test1Cond]            22.788      1    22.788      37.161500   0.000003
```

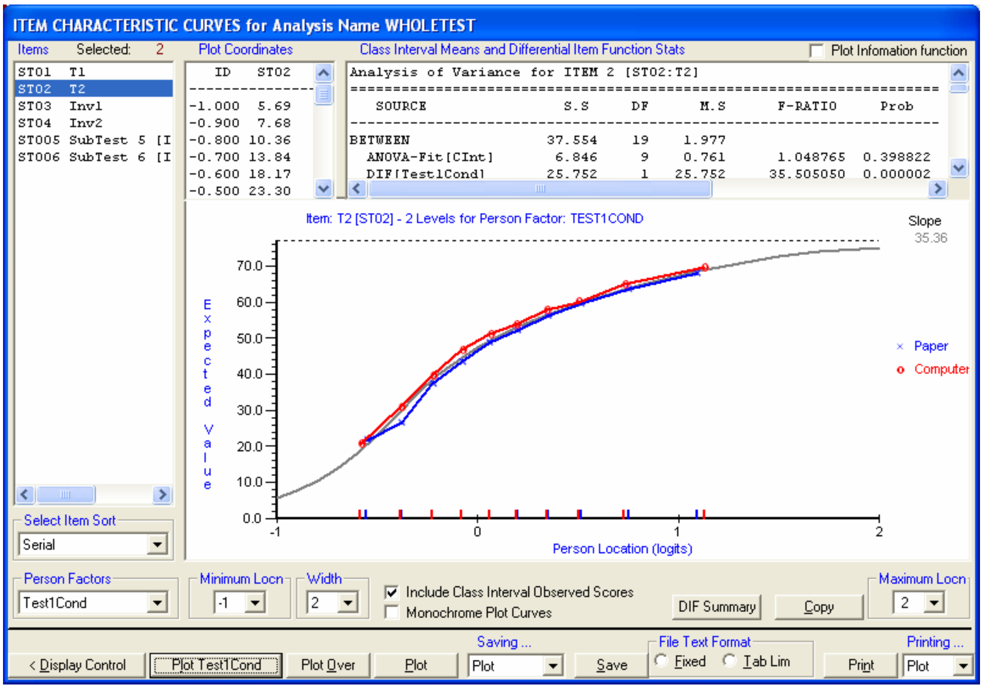
7.11 Overall performance and relationship to ability for Test B

For each of the two tests, Rasch analysis was used to represent performance on tests as whole and individual item characteristic curves.

The rounded DIF figure of p = 0.000 for Test B is shown in Figure 36 below. This indicates a highly significant differential performance (DIF) between students taking this test on paper (the blue line) and computer (the red line) in favour of the computer test.

In the case of Test B, as shown in Figure 36, the item characteristic curves show an fairly consistent difference in performance (the Y axis) across the student ability range (the X axis) between paper and computer modes.

Figure 36: Rasch item characteristic curves for the whole of Test B in paper and computer modes



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

Items Selected: 2

```
ST01   T1
ST02   T2
ST03   Inv1
ST04   Inv2
ST005  SubTest 5 [I
ST006  SubTest 6 [I
```

Plot Coordinates

```
   ID     ST02
---------------
-1.000   5.69
-0.900   7.68
-0.800  10.36
-0.700  13.84
-0.600  18.17
-0.500  23.30
```

Class Interval Means and Differential Item Function Stats

```
Analysis of Variance for ITEM 2 [ST02:T2]
========================================================================
   SOURCE                S.S      DF       M.S       F-RATIO      Prob
------------------------------------------------------------------------
BETWEEN                37.554     19     1.977
  ANOVA-Fit[CInt]       6.846      9     0.761      1.048765   0.398822
  DIF[Test1Cond]       25.752      1    25.752     35.505050   0.000002
```

Select Item Sort: Serial

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

□ Monochrome Plot Curves

Maximum Locn: 2

As described, there was a significant difference between the total test scores in paper and computer modes of Test B, and the items shown in Table 15 on page 125, show all the items within Test B which had the highest level of significant differences (significance where P< 0.005).

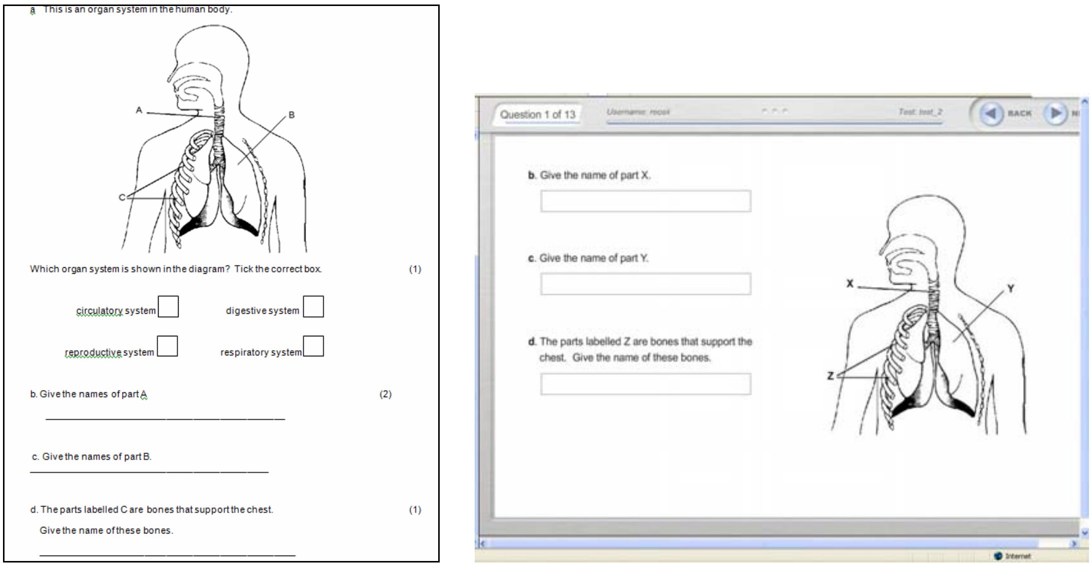
Discussion of the differences between modes will be carried out in the Analysis chapter. The following section will indicate the nature of the differences for each DIF item within Tests B.

7.12 DIF items within Test B

Table 34: Performance data on Q1d

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
1d	2DC	OR1 B D	0.90	0.97	0.416	0.313	0.005 C

Figure 37: Screenshots of Q1d in paper and computer mode



a This is an organ system in the human body.

Which organ system is shown in the diagram? Tick the correct box. (1)

circulatory system □ digestive system □

reproductive system □ respiratory system □

b. Give the names of part A (2)

____

c. Give the names of part B.

____

d. The parts labelled C are bones that support the chest. (1)

Give the name of these bones.

____

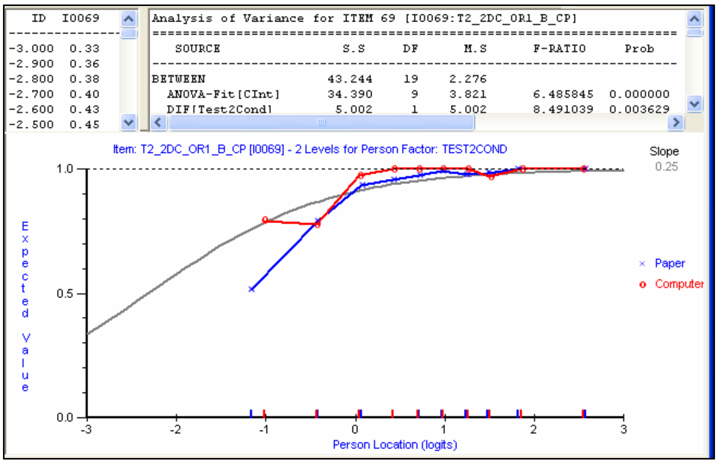
b. Give the name of part X.

c. Give the name of part Y.

d. The parts labelled Z are bones that support the chest. Give the name of these bones.

This item was targeted at a low level of difficulty, requiring students to name particular bones in the body using open response. The facility values in both modes were both high, as expected for this relatively easy item, however there was a DIF in favour of computer. Although the facility values were high, this item also achieved good discrimination values. The Rasch item characteristic curves shown below in Figure 38 showed a fairly consistent difference in performance across the student ability range in favour of computer.

Figure 38: Rasch item characteristic curves for Q1d in paper and computer modes



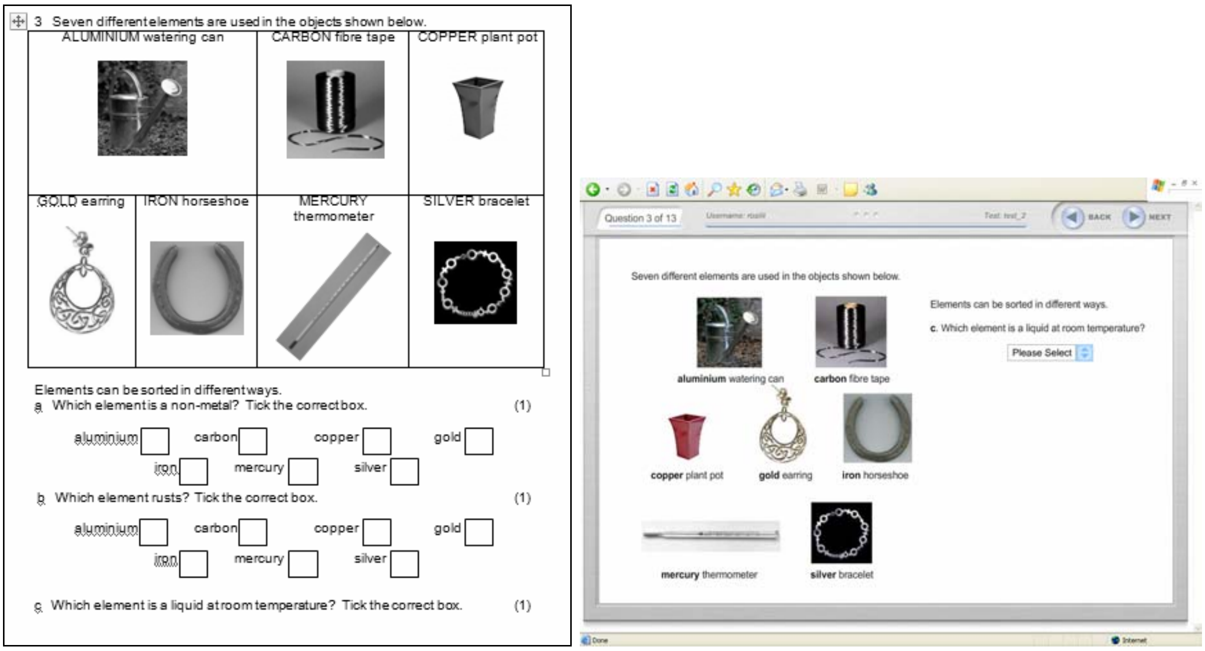
```
  ID   I0069
---------------
-3.000  0.33
-2.900  0.36
-2.800  0.38
-2.700  0.40
-2.600  0.43
-2.500  0.45
```

```
Analysis of Variance for ITEM 69 [I0069:T2_2DC_OR1_B_CP]
=====================================================================
   SOURCE                S.S     DF      M.S       F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN                43.244    19     2.276
  ANOVA-Fit[CInt]      34.390     9     3.821      6.485845  0.000000
  DIF[Test2Cond]        5.002     1     5.002      8.491039  0.003629
```

Table 35: Performance data on Q3c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
3c	9CA	DL C CP	0.75	0.89	0.495	0.380	0.000 C

Figure 39: Screenshots of Q3c in paper and computer mode



3 Seven different elements are used in the objects shown below.

ALUMINIUM watering can		CARBON fibre tape	COPPER plant pot
GOLD earring	IRON horseshoe	MERCURY thermometer	SILVER bracelet

Elements can be sorted in different ways.

a Which element is a non-metal? Tick the correct box. (1)

aluminium □ carbon □ copper □ gold □
iron □ mercury □ silver □

b Which element rusts? Tick the correct box. (1)

aluminium □ carbon □ copper □ gold □
iron □ mercury □ silver □

c Which element is a liquid at room temperature? Tick the correct box. (1)

Seven different elements are used in the objects shown below.

aluminium watering can
carbon fibre tape
copper plant pot
gold earring
iron horseshoe
mercury thermometer
silver bracelet

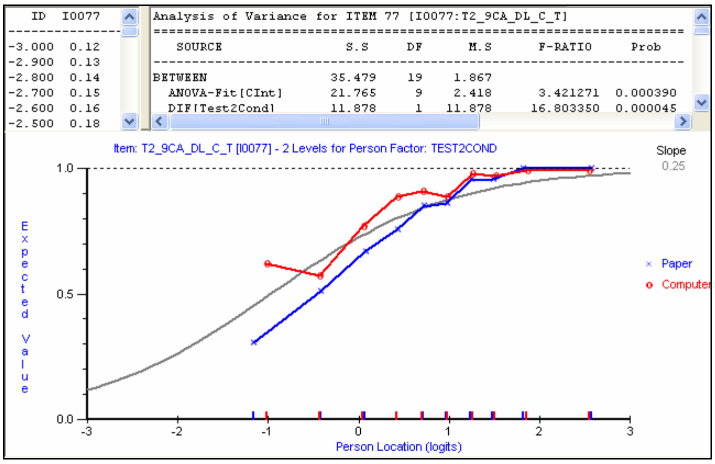
Elements can be sorted in different ways.

c. Which element is a liquid at room temperature?

Please Select

This item was targeted at a low level of difficulty and involved students selecting the name of a metal liquid at room temperature. The images of metals on paper were in black and white and in colour on computer. On paper students ticked a box whereas on computer students chose an option from a drop down. The facility value on computer (0.89) was significantly higher than on paper (0.75). The discrimination values in both modes were good. The Rasch item characteristic curves shown in figure 40 below showed a fairly consistent difference in performance across the student ability range in favour of computer.

Figure 40: Rasch item characteristic curves for Q3c in paper and computer modes



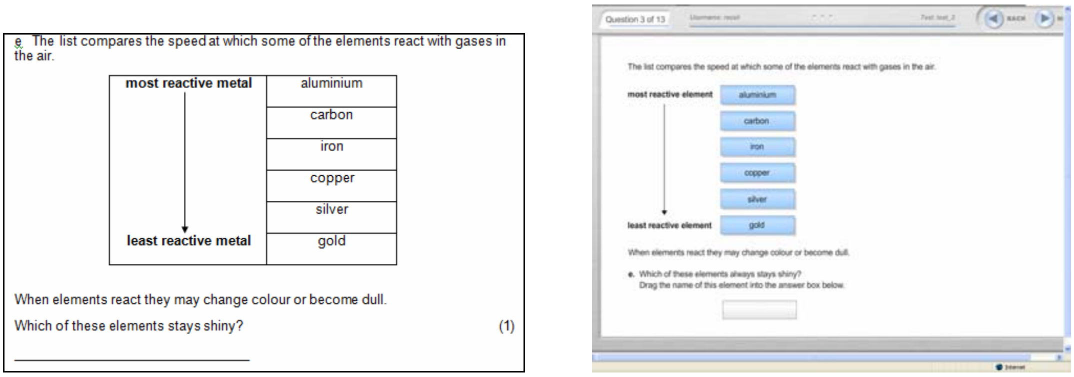
```
  ID   I0077
---------------
-3.000  0.12
-2.900  0.13
-2.800  0.14
-2.700  0.15
-2.600  0.16
-2.500  0.18
```

```
Analysis of Variance for ITEM 77 [I0077:T2_9CA_DL_C_T]
==========================================================================
  SOURCE                  S.S      DF       M.S       F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                 35.479     19     1.867
  ANOVA-Fit[CInt]       21.765      9     2.418      3.421271  0.000390
  DIF[Test2Cond]        11.878      1    11.878     16.803350  0.000045
```

Table 36: Performance data on Q3e

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
3e	11EA	DD C T	0.67	0.64	0.209	0.269	0.000 P

Figure 41: Screenshots of Q3e in paper and computer mode



e The list compares the speed at which some of the elements react with gases in the air.

most reactive metal	aluminium
	carbon
	iron
	copper
	silver
least reactive metal	gold

When elements react they may change colour or become dull.

Which of these elements stays shiny? (1)

____

The list compares the speed at which some of the elements react with gases in the air.

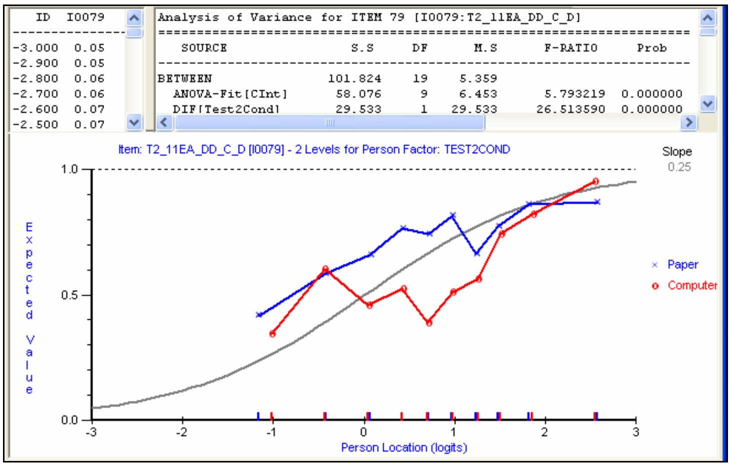
most reactive element	aluminium
	carbon
	iron
	copper
	silver
least reactive element	gold

When elements react they may change colour or become dull.

e. Which of these elements always stays shiny?
Drag the name of this element into the answer box below.

This item was targeted at a low level of difficulty and involved students selecting a metal that stays shiny from a list. On paper students had to write their answer whereas on computer they dragged their answer from a list. The overall facilities in both modes were similar, although not particularly good, and similarly the discrimination values for this item were poor. The Rasch item characteristic curves shown below in Figure 42 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 42: Rasch item characteristic curves for Q3e in paper and computer modes



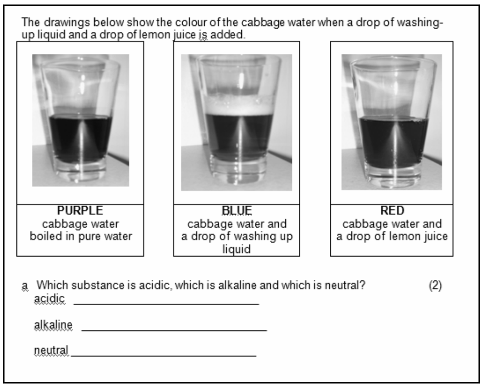
```
   ID   I0079
-----------------
-3.000   0.05
-2.900   0.05
-2.800   0.06
-2.700   0.06
-2.600   0.07
-2.500   0.07
```

```
Analysis of Variance for ITEM 79 [I0079:T2_11EA_DD_C_D]
=======================================================================
   SOURCE                S.S      DF      M.S       F-RATIO     Prob
-----------------------------------------------------------------------
BETWEEN               101.824     19     5.359
  ANOVA-Fit[CInt]      58.076      9     6.453      5.793219  0.000000
  DIF[Test2Cond]       29.533      1    29.533     26.513590  0.000000
```

Table 37: Performance data on Q4a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
4a	12AA	DD C CP	0.69	0.70	0.484	0.387	0.000 P

Figure 43: Screenshots of Q4a in paper and computer mode



The drawings below show the colour of the cabbage water when a drop of washing-up liquid and a drop of lemon juice is added.

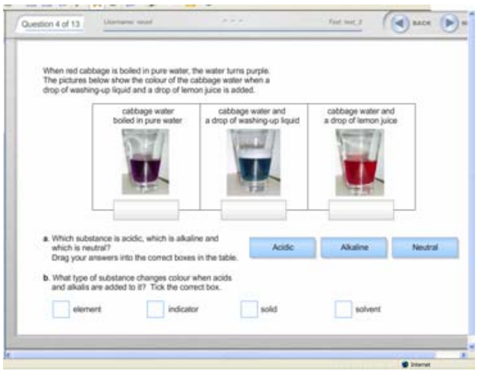
PURPLE cabbage water boiled in pure water	**BLUE** cabbage water and a drop of washing up liquid	**RED** cabbage water and a drop of lemon juice

a Which substance is acidic, which is alkaline and which is neutral? (2)

acidic ____________________________

alkaline ____________________________

neutral ____________________________



When red cabbage is boiled in pure water, the water turns purple. The pictures below show the colour of the cabbage water when a drop of washing-up liquid and a drop of lemon juice is added.

cabbage water boiled in pure water	cabbage water and a drop of washing-up liquid	cabbage water and a drop of lemon juice

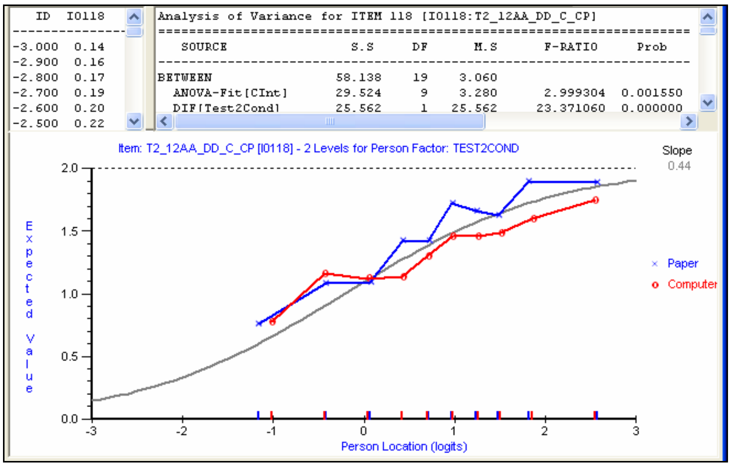
a. Which substance is acidic, which is alkaline and which is neutral? Drag your answers into the correct boxes in the table.

b. What type of substance changes colour when acids and alkalis are added to it? Tick the correct box.

□ element □ indicator □ solid □ solvent

This item was targeted at a medium level of difficulty and involved students identifying acidic, neutral or alkaline substances. The paper diagrams were in black and white and therefore indicated the colour of the indicator whereas the computer version was in colour and students had to use the colour information. The overall facility values were similar and reasonably good, as were the discrimination values. The Rasch item characteristic curves shown below in Figure 44 showed a fairly consistent difference in performance across the student ability range in favour of paper

Figure 44: Rasch item characteristic curves for Q4a in paper and computer modes



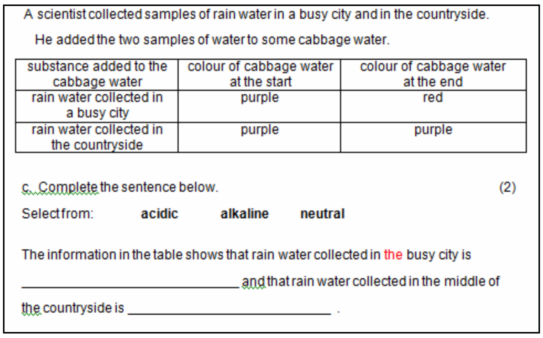
```
   ID   I0118
---------------
-3.000   0.14
-2.900   0.16
-2.800   0.17
-2.700   0.19
-2.600   0.20
-2.500   0.22
```

```
Analysis of Variance for ITEM 118 [I0118:T2_12AA_DD_C_CP]
==============================================================================
   SOURCE                  S.S      DF       M.S       F-RATIO      Prob
------------------------------------------------------------------------------
BETWEEN                  58.138     19     3.060
  ANOVA-Fit[CInt]        29.524      9     3.280      2.999304   0.001550
  DIF[Test2Cond]         25.562      1    25.562     23.371060   0.000000
```

Table 38: Performance data on Q4a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
4c	13DB	DL C T	0.68	0.57	0.301	0.178	0.000 P

Figure 45: Screenshots of Q4a in paper and computer mode



A scientist collected samples of rain water in a busy city and in the countryside.

He added the two samples of water to some cabbage water.

substance added to the cabbage water	colour of cabbage water at the start	colour of cabbage water at the end
rain water collected in a busy city	purple	red
rain water collected in the countryside	purple	purple

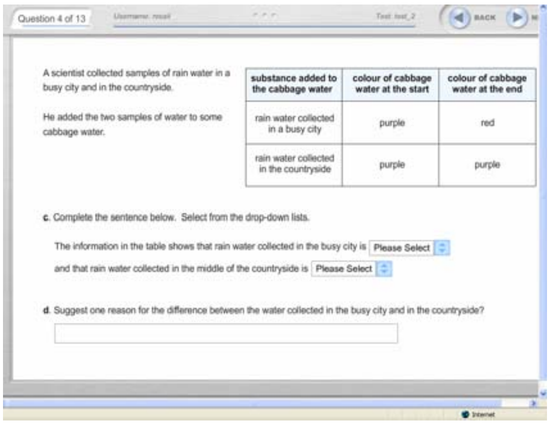
c. Complete the sentence below. (2)

Select from: **acidic** **alkaline** **neutral**

The information in the table shows that rain water collected in the busy city is

______________________ and that rain water collected in the middle of

the countryside is ______________________ .



A scientist collected samples of rain water in a busy city and in the countryside.

He added the two samples of water to some cabbage water.

substance added to the cabbage water	colour of cabbage water at the start	colour of cabbage water at the end
rain water collected in a busy city	purple	red
rain water collected in the countryside	purple	purple

c. Complete the sentence below. Select from the drop-down lists.

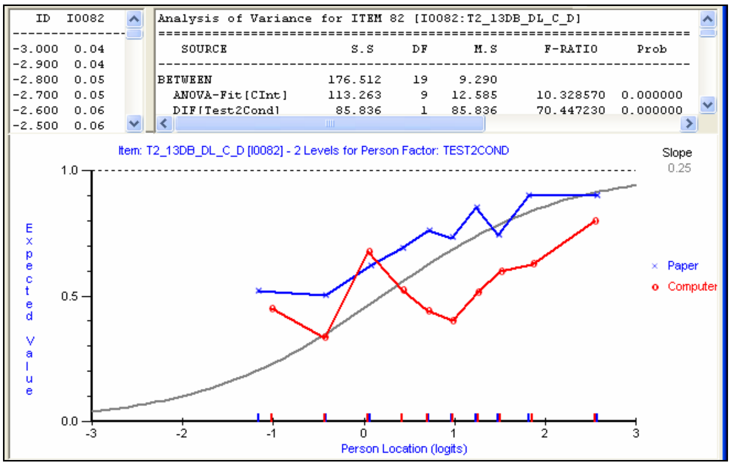
The information in the table shows that rain water collected in the busy city is Please Select

and that rain water collected in the middle of the countryside is Please Select

d. Suggest one reason for the difference between the water collected in the busy city and in the countryside?

This item was targeted at a low level of difficulty and involved students having to use information in a table to categorise acidity. Tables of information were given in both modes, the only difference being that on paper students wrote their answer, whereas on computer they used a drop down list. The facility value was significantly higher on paper (0.68) than on computer (0.57) as were the discrimination values. The Rasch item characteristic curves shown below in Figure 46 showed a consistent difference in performance across the student ability range in favour of paper.

Figure 46: Rasch item characteristic curves for Q4a in paper and computer modes



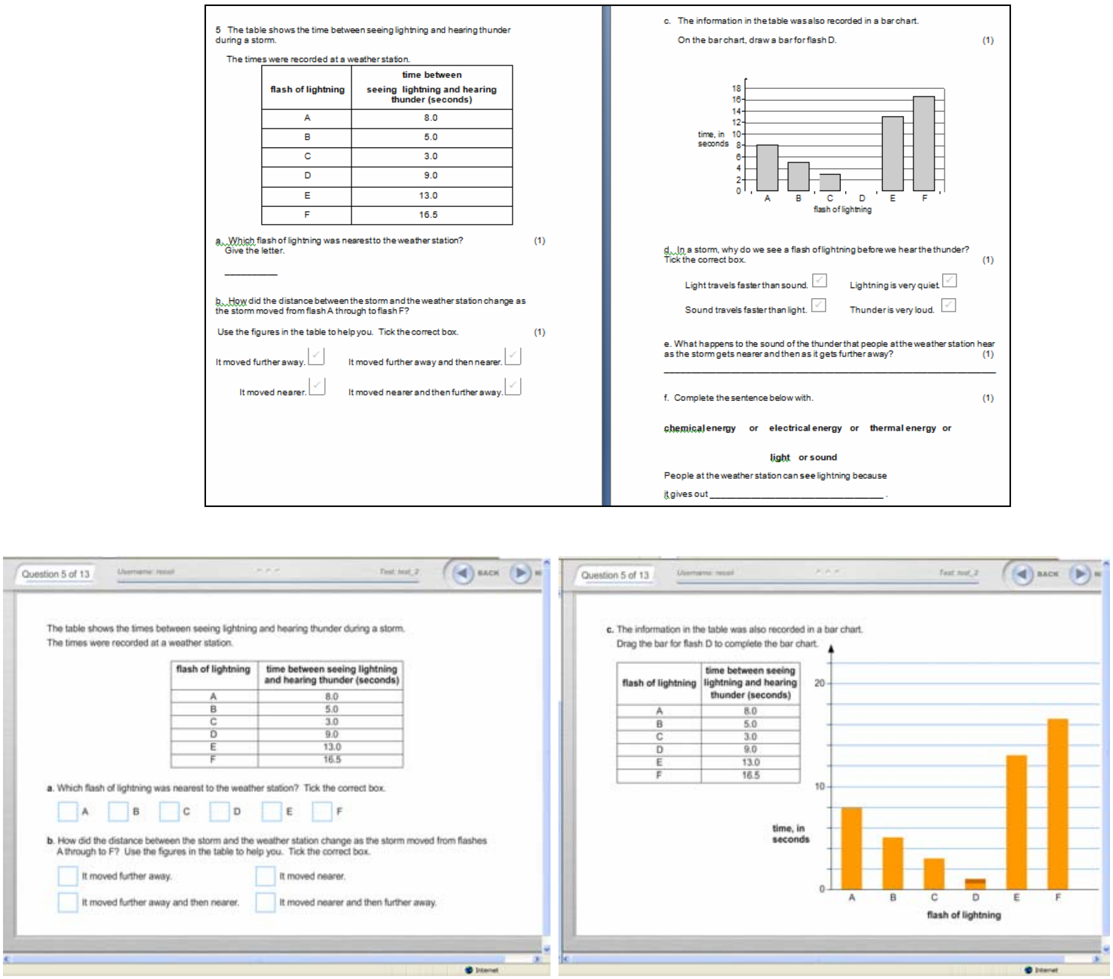
```
  ID   I0082
---------------
-3.000  0.04
-2.900  0.04
-2.800  0.05
-2.700  0.05
-2.600  0.06
-2.500  0.06
```

```
Analysis of Variance for ITEM 82 [I0082:T2_13DB_DL_C_D]
=====================================================================
   SOURCE                S.S      DF       M.S       F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN               176.512     19     9.290
  ANOVA-Fit[CInt]     113.263      9    12.585     10.328570  0.000000
  DIF[Test2Cond]       85.836      1    85.836     70.447230  0.000000
```

Table 39: Performance data on Q5b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
5b	14BB	TB P T	0.73	0.72	0.52	0.507	0.000 P

Figure 47: Screenshots of Q5b and c in paper and computer mode



5 The table shows the time between seeing lightning and hearing thunder during a storm.

The times were recorded at a weather station.

flash of lightning	time between seeing lightning and hearing thunder (seconds)
A	8.0
B	5.0
C	3.0
D	9.0
E	13.0
F	16.5

a. Which flash of lightning was nearest to the weather station? (1)
Give the letter.

____

b. How did the distance between the storm and the weather station change as the storm moved from flash A through to flash F?

Use the figures in the table to help you. Tick the correct box. (1)

It moved further away. ☐ It moved further away and then nearer. ☐

It moved nearer. ☐ It moved nearer and then further away. ☐

c. The information in the table was also recorded in a bar chart.

On the bar chart, draw a bar for flash D. (1)

d. In a storm, why do we see a flash of lightning before we hear the thunder? (1)
Tick the correct box.

Light travels faster than sound. ☐ Lightning is very quiet. ☐

Sound travels faster than light. ☐ Thunder is very loud. ☐

e. What happens to the sound of the thunder that people at the weather station hear as the storm gets nearer and then as it gets further away? (1)

____

f. Complete the sentence below with. (1)

chemical energy or **electrical energy** or **thermal energy** or

light or **sound**

People at the weather station can see lightning because

it gives out ______________________________.

The table shows the times between seeing lightning and hearing thunder during a storm.
The times were recorded at a weather station.

flash of lightning	time between seeing lightning and hearing thunder (seconds)
A	8.0
B	5.0
C	3.0
D	9.0
E	13.0
F	16.5

a. Which flash of lightning was nearest to the weather station? Tick the correct box.

☐ A ☐ B ☐ C ☐ D ☐ E ☐ F

b. How did the distance between the storm and the weather station change as the storm moved from flashes A through to F? Use the figures in the table to help you. Tick the correct box.

☐ It moved further away. ☐ It moved nearer.

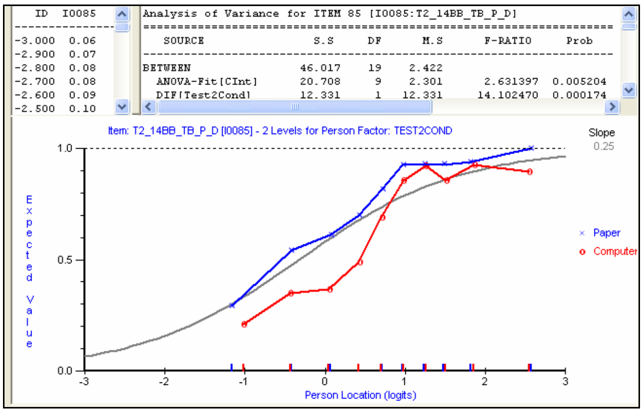
☐ It moved further away and then nearer. ☐ It moved nearer and then further away.

c. The information in the table was also recorded in a bar chart.
Drag the bar for flash D to complete the bar chart.

flash of lightning	time between seeing lightning and hearing thunder (seconds)
A	8.0
B	5.0
C	3.0
D	9.0
E	13.0
F	16.5

This item was targeted at a medium level of difficulty involving students describing the movement of a storm. On paper students ticked a box whereas on computer they used an on-screen tick box function. The overall facilities in both modes were similar and good as were the discrimination values. Although the overall facilities for this item were similar, the Rasch item discrimination curves shown below in Figure 48 showed a consistent difference in performance across the student ability range in favour of paper.

Figure 48: Rasch item characteristic curves for Q5b in paper and computer modes



```
   ID   I0085
---------------
-3.000   0.06
-2.900   0.07
-2.800   0.08
-2.700   0.08
-2.600   0.09
-2.500   0.10
```

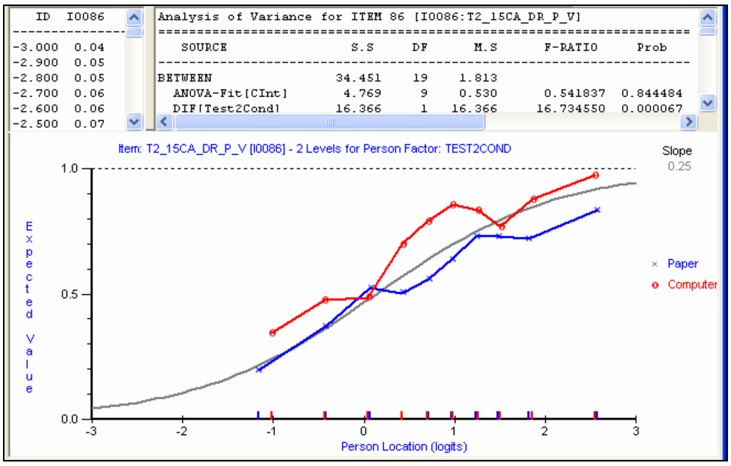
```
Analysis of Variance for ITEM 85 [I0085:T2_14BB_TB_P_D]
======================================================================
   SOURCE                 S.S      DF      M.S       F-RATIO     Prob
----------------------------------------------------------------------
BETWEEN                 46.017     19     2.422
  ANOVA-Fit[CInt]       20.708      9     2.301      2.631397  0.005204
  DIF[Test2Cond]        12.331      1    12.331     14.102470  0.000174
```

Table 40: Performance data on Q5c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
5c	15CA	DR P T	0.54	0.76	0.312	0.390	0.000 C

This item was another item within Question 5. It was targeted at a low level of difficulty and involved students having to complete a bar chart. On paper students had to draw in a bar whereas on computer they had to drag a bar into place. The facility value on computer (0.76)was significantly higher than on paper (0.54), with both modes achieving reasonable discrimination values. The Rasch item characteristic curves shown below in Figure 49 showed a fairly consistent difference in performance across the student ability range in favour of computer.

Figure 49: Rasch item characteristic curves for Q5b in paper and computer modes



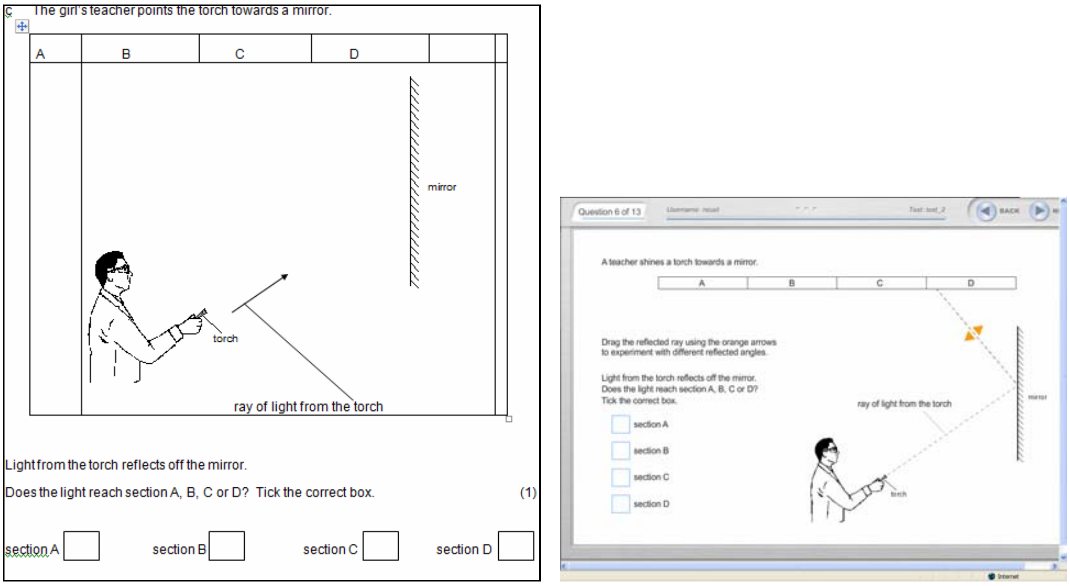
```
   ID   I0086
---------------
-3.000   0.04
-2.900   0.05
-2.800   0.05
-2.700   0.06
-2.600   0.06
-2.500   0.07
```

```
Analysis of Variance for ITEM 86 [I0086:T2_15CA_DR_P_V]
======================================================================
   SOURCE                 S.S      DF      M.S       F-RATIO     Prob
----------------------------------------------------------------------
BETWEEN                 34.451     19     1.813
  ANOVA-Fit[CInt]        4.769      9     0.530      0.541837  0.844484
  DIF[Test2Cond]        16.366      1    16.366     16.734550  0.000067
```

Table 41: Performance data on Q6a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
6a	17AA	TB P ID	0.62	0.23	0.239	-0.004	0.000 P

Figure 50: Screenshots of Q6a in paper and computer mode



c The girl's teacher points the torch towards a mirror.

Light from the torch reflects off the mirror.

Does the light reach section A, B, C or D? Tick the correct box. (1)

section A □ section B □ section C □ section D □

Question 6 of 13

A teacher shines a torch towards a mirror.

Drag the reflected ray using the orange arrows to experiment with different reflected angles.

Light from the torch reflects off the mirror.
Does the light reach section A, B, C or D?
Tick the correct box.

□ section A

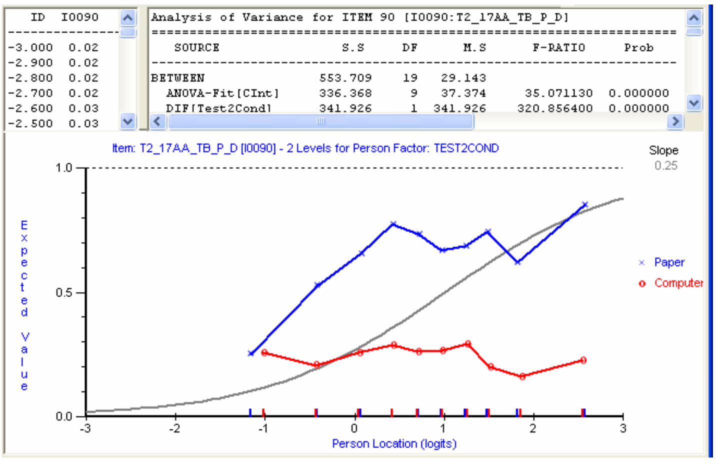
□ section B

□ section C

□ section D

This item was targeted at a medium level of difficulty and involved students selecting a refecting angle. On paper students could use a ruler or protractor to estimate the angle whereas on computer they could use an on screen tool to manipulate a ray and estimate at which point it would reflect. The facility value on paper was significantly higher than on computer. This was the only item within either test that showed a negative discrimination, in this case on the computer version. The Rasch item characteristic curves shown below in Figure 51 show a considerable difference in performance across the student ability range in favour of paper.

Figure 51: Rasch item characteristic curves for Q6a in paper and computer modes



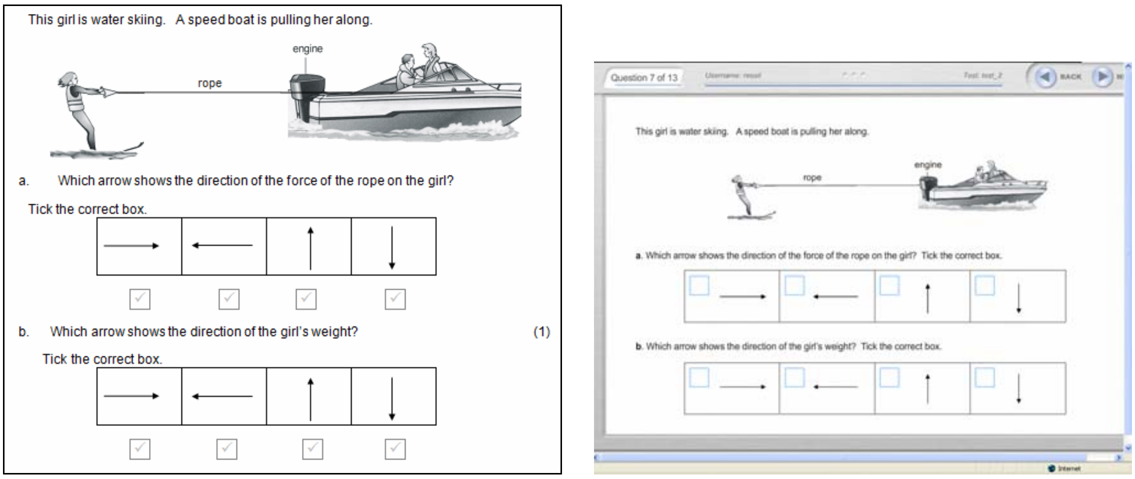
```
  ID   I0090
---------------
-3.000  0.02
-2.900  0.02
-2.800  0.02
-2.700  0.02
-2.600  0.03
-2.500  0.03
```

```
Analysis of Variance for ITEM 90 [I0090:T2_17AA_TB_P_D]
=======================================================================
   SOURCE                S.S      DF      M.S       F-RATIO      Prob
-----------------------------------------------------------------------
BETWEEN                553.709    19   29.143
  ANOVA-Fit[CInt]      336.368     9   37.374     35.071130  0.000000
  DIF[Test2Cond]       341.926     1  341.926    320.856400  0.000000
```

Table 42: Performance data on Q7b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
7b	18BB	TB P D	0.51	0.51	0.21	0.229	0.000 P

Figure 52: Screenshots of Q7b in paper and computer mode



This girl is water skiing. A speed boat is pulling her along.

a. Which arrow shows the direction of the force of the rope on the girl?

Tick the correct box.

b. Which arrow shows the direction of the girl's weight? (1)

Tick the correct box.

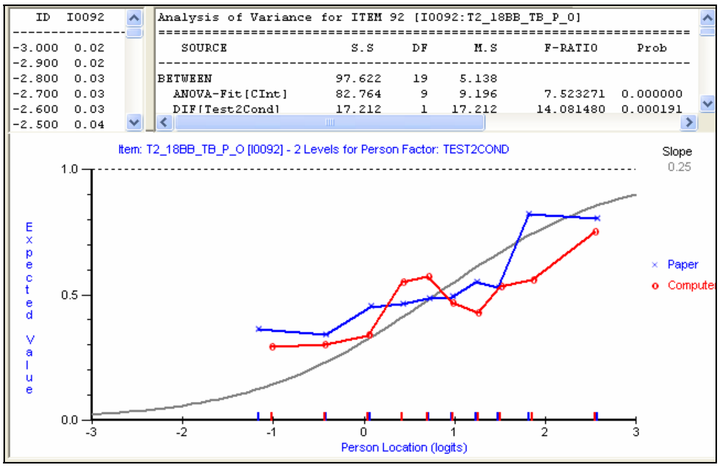
This girl is water skiing. A speed boat is pulling her along.

a. Which arrow shows the direction of the force of the rope on the girl? Tick the correct box.

b. Which arrow shows the direction of the girl's weight? Tick the correct box.

This item was targeted at a medium level of difficulty and involved students choosing a direction of force. The images were the same in both modes, the only difference being the answering mechanism. On paper students ticked a box, on screen they clicked a box. The overall facility value of this item was the same in both modes, as were the discrimination values. The item characteristic curves shown below in Figure 53 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 53: Rasch item characteristic curves for Q7b in paper and computer modes



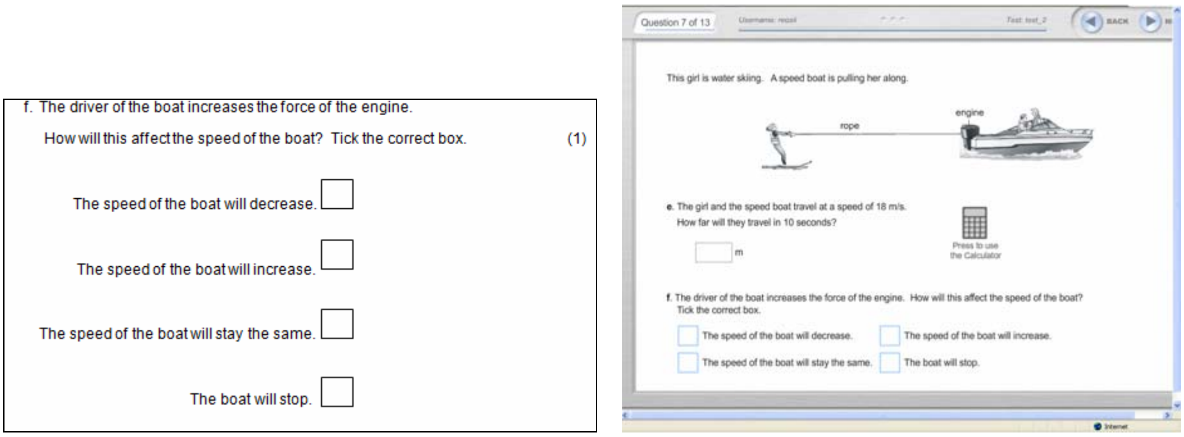
```
  ID   I0092
----------------
-3.000  0.02
-2.900  0.02
-2.800  0.03
-2.700  0.03
-2.600  0.03
-2.500  0.04
```

```
Analysis of Variance for ITEM 92 [I0092:T2_18BB_TB_P_O]
============================================================================
   SOURCE                 S.S      DF      M.S        F-RATIO      Prob
----------------------------------------------------------------------------
BETWEEN                 97.622     19     5.138
  ANOVA-Fit[CInt]       82.764      9     9.196       7.523271   0.000000
  DIF[Test2Cond]        17.212      1    17.212      14.081480   0.000191
```

Table 43: Performance data on Q7f

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
7f	20FB	TB P D	0.87	0.84	0.419	0.561	0.000 P

Figure 54: Screenshots of Q7f in paper and computer mode



f. The driver of the boat increases the force of the engine.

How will this affect the speed of the boat? Tick the correct box. (1)

The speed of the boat will decrease. □

The speed of the boat will increase. □

The speed of the boat will stay the same. □

The boat will stop. □

Question 7 of 13

This girl is water skiing. A speed boat is pulling her along.

e. The girl and the speed boat travel at a speed of 18 m/s.
How far will they travel in 10 seconds?

□ m

Press to use the Calculator

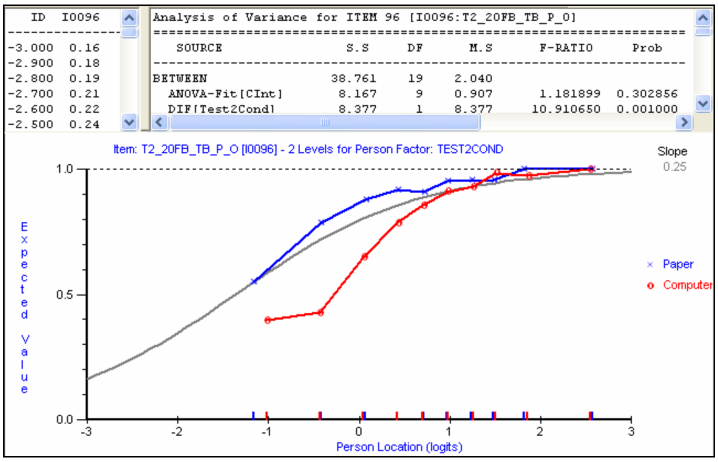
f. The driver of the boat increases the force of the engine. How will this affect the speed of the boat?
Tick the correct box.

□ The speed of the boat will decrease. □ The speed of the boat will increase.

□ The speed of the boat will stay the same. □ The boat will stop.

This item was targeted at a low level of difficulty and involved students deciding the effect an increased force would have on speed. The diagrams used in both modes were the same and the only difference was that on paper students ticked a box and on computer they clicked a box. Although the overall facilities in each mode were high and similar, as were the discrimination values, the Rasch item characteristic curves shown below in Figure 55 showed a difference in performance across the ability range in favour of paper, and in particular at lower ability levels.

Figure 55: Rasch item characteristic curves for Q7f in paper and computer modes



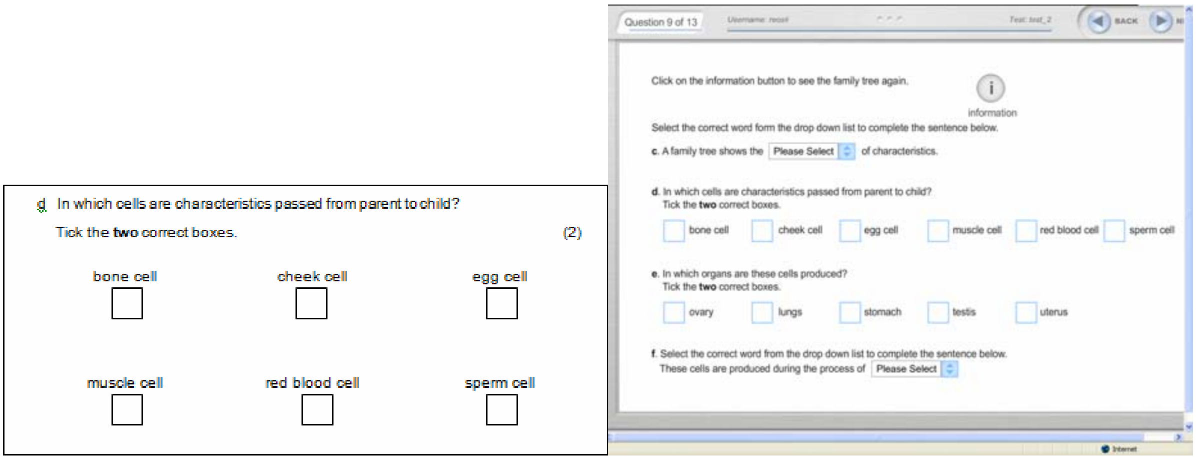
```
  ID   I0096
---------------
-3.000  0.16
-2.900  0.18
-2.800  0.19
-2.700  0.21
-2.600  0.22
-2.500  0.24
```

```
Analysis of Variance for ITEM 96 [I0096:T2_20FB_TB_P_O]
=====================================================================
   SOURCE                S.S      DF       M.S       F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN                38.761     19     2.040
  ANOVA-Fit[CInt]       8.167      9     0.907      1.181899   0.302856
  DIF[Test2Cond]        8.377      1     8.377     10.910650   0.001000
```

Table 44: Performance data on Q9d

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
9d	25DB	TB B O	0.63	0.70	0.630	0.664	0.003 P

Figure 56: Screenshots of Q9d in paper and computer mode



d In which cells are characteristics passed from parent to child?

Tick the **two** correct boxes. (2)

bone cell □ cheek cell □ egg cell □

muscle cell □ red blood cell □ sperm cell □

Question 9 of 13

Click on the information button to see the family tree again.

Select the correct word form the drop down list to complete the sentence below.

c. A family tree shows the Please Select of characteristics.

d. In which cells are characteristics passed from parent to child?
Tick the **two** correct boxes.

□ bone cell □ cheek cell □ egg cell □ muscle cell □ red blood cell □ sperm cell

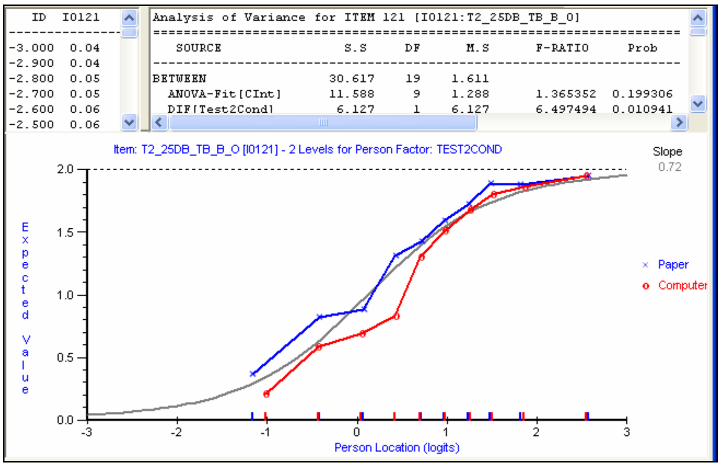
e. In which organs are these cells produced?
Tick the **two** correct boxes.

□ ovary □ lungs □ stomach □ testis □ uterus

f. Select the correct word from the drop down list to complete the sentence below.
These cells are produced during the process of Please Select

This item was targeted at a medium level of difficulty and involved students selecting two cells involved with inheritance. There was no stimulus for this question, and the response mechanisms were similar, students ticking two boxes on paper and clicking two boxes on computer. The overall facility values for this item were good; computer (0.70) higher than paper (0.63) and the discrimination values in both modes were very good. Although the overall facility for computer was higher than on paper, the Rasch item characteristic curve shown below in Figure 57 showed a fairly consistent difference in performance across the ability range in favour of paper.

Figure 57: Rasch item characteristic curves for Q7f in paper and computer modes



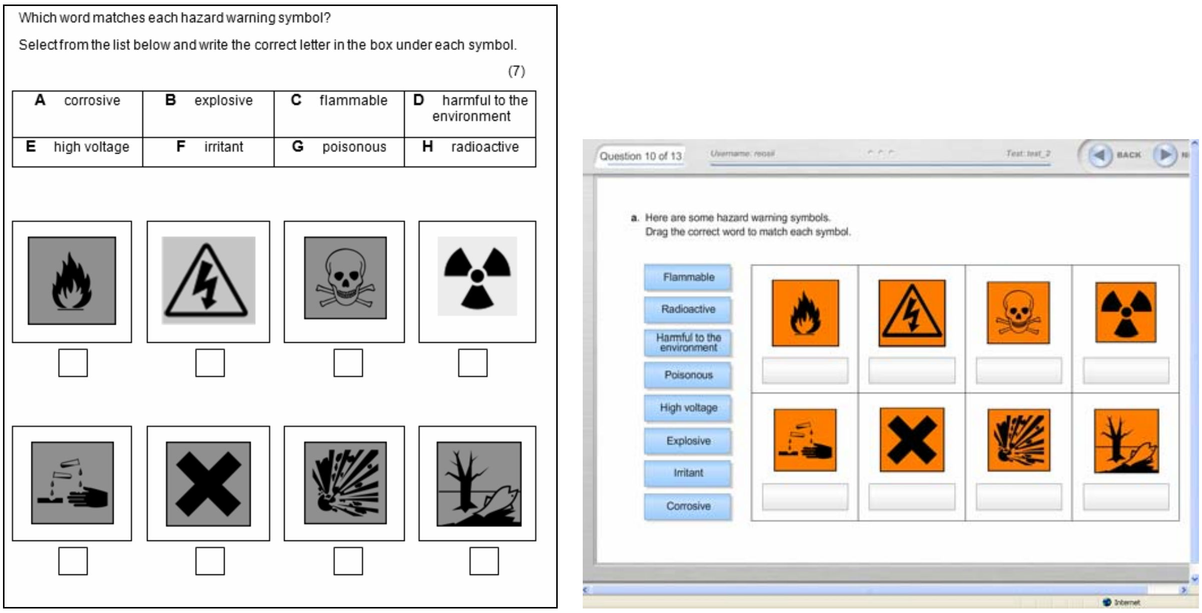
```
  ID   I0121
---------------
-3.000  0.04
-2.900  0.04
-2.800  0.05
-2.700  0.05
-2.600  0.06
-2.500  0.06
```

```
Analysis of Variance for ITEM 121 [I0121:T2_25DB_TB_B_O]
=====================================================================
   SOURCE                S.S      DF      M.S        F-RATIO    Prob
---------------------------------------------------------------------
BETWEEN                30.617     19     1.611
  ANOVA-Fit[CInt]      11.588      9     1.288      1.365352  0.199306
  DIF[Test2Cond]        6.127      1     6.127      6.497494  0.010941
```

Table 45: Performance data on Q10a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
10a	26AA	DD C CP	0.69	0.77	0.709	0.656	0.000 P

Figure 58: Screenshots of Q10a in paper and computer mode



Which word matches each hazard warning symbol?

Select from the list below and write the correct letter in the box under each symbol.

(7)

A corrosive	**B** explosive	**C** flammable	**D** harmful to the environment
E high voltage	**F** irritant	**G** poisonous	**H** radioactive

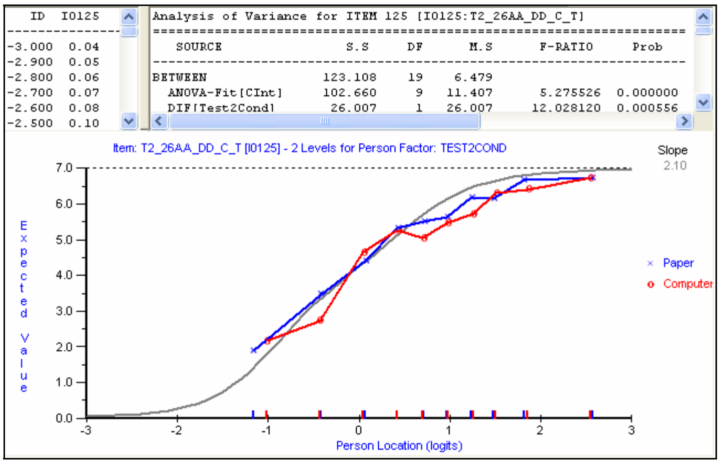
□ □ □ □

□ □ □ □

a. Here are some hazard warning symbols.
Drag the correct word to match each symbol.

This was a multi-mark question targeted at a medium level of difficulty. It involved students choosing scientific hazard symbols. On paper the images were in black and white and students wrote appropriate letters into boxes under the symbols. On screen the symbols were in colour, and the answering mechanisms were drag and drop. The overall facility value was higher on computer than on paper, however the Rasch item characteristic curves shown below in Figure 59 showed a fairly showed a fairly consistent difference in performance across the ability range in favour of paper.

Figure 59: Rasch item characteristic curves for Q10a in paper and computer modes



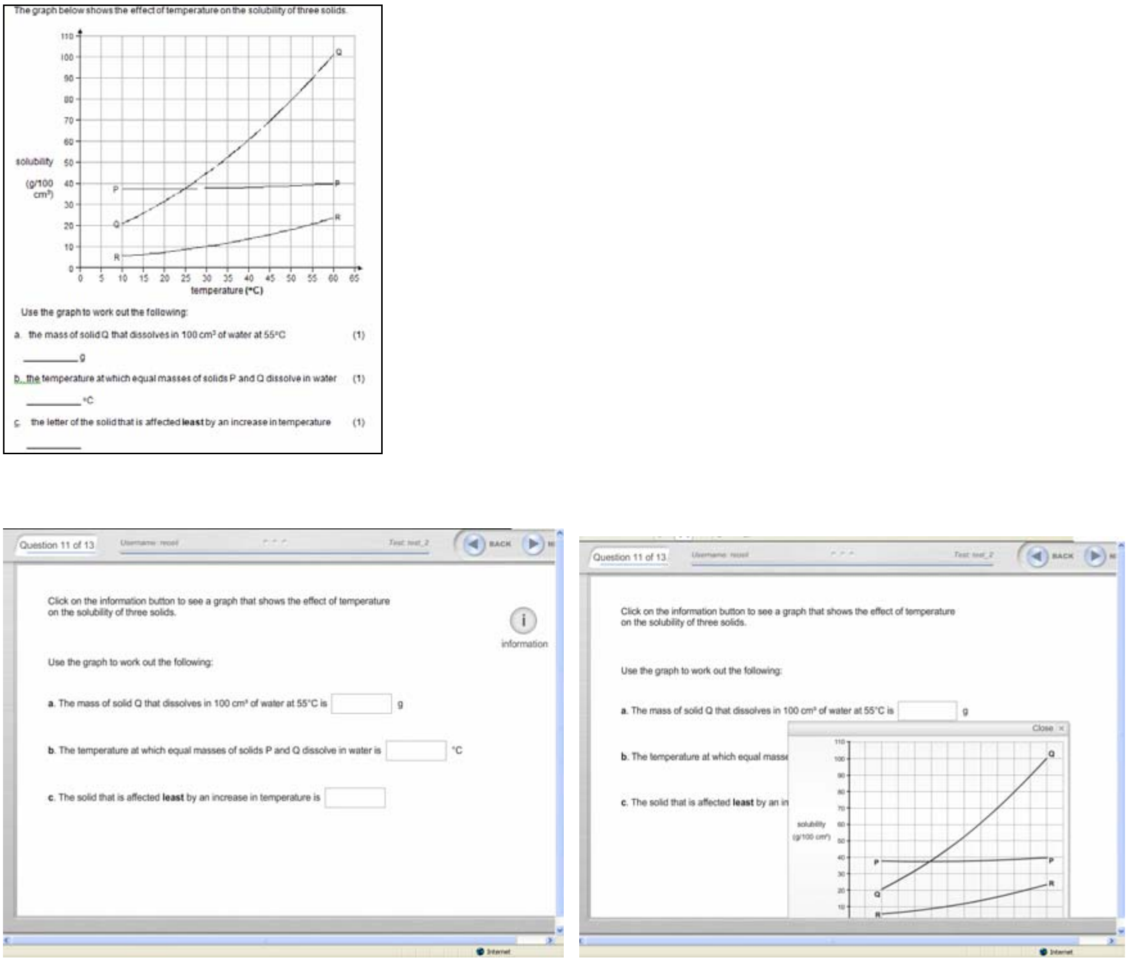
```
  ID   I0125
----------------
-3.000  0.04
-2.900  0.05
-2.800  0.06
-2.700  0.07
-2.600  0.08
-2.500  0.10
```

```
Analysis of Variance for ITEM 125 [I0125:T2_26AA_DD_C_T]
==========================================================================
   SOURCE              S.S      DF      M.S       F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN              123.108    19    6.479
  ANOVA-Fit[CInt]    102.660     9   11.407      5.275526  0.000000
  DIF[Test2Cond]      26.007     1   26.007     12.028120  0.000556
```

Table 46: Performance data on Q11c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
11c	27CC	OR C I	0.41	0.34	0.435	0.416	0.000 P

Figure 60: Screenshots of Q11c in paper and computer mode



The graph below shows the effect of temperature on the solubility of three solids.

Use the graph to work out the following:

a. the mass of solid Q that dissolves in 100 cm³ of water at 55°C (1)

________g

b. the temperature at which equal masses of solids P and Q dissolve in water (1)

________°C

c. the letter of the solid that is affected **least** by an increase in temperature (1)

____

Click on the information button to see a graph that shows the effect of temperature on the solubility of three solids.

Use the graph to work out the following:

a. The mass of solid Q that dissolves in 100 cm³ of water at 55°C is □ g

b. The temperature at which equal masses of solids P and Q dissolve in water is □ °C

c. The solid that is affected **least** by an increase in temperature is □

Click on the information button to see a graph that shows the effect of temperature on the solubility of three solids.

Use the graph to work out the following:

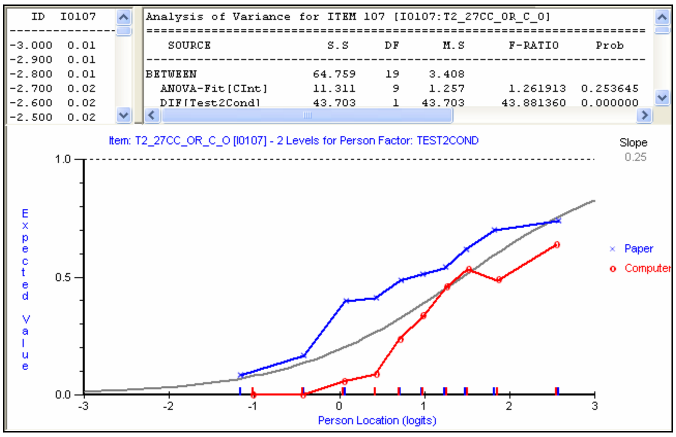
a. The mass of solid Q that dissolves in 100 cm³ of water at 55°C is □ g

b. The temperature at which equal masse

c. The solid that is affected **least** by an in

This item was targeted at a medium level of difficulty and involved students providing an open response using information from a graph. On paper the graph containing the question and the question were given together. On screen, this graph was accessed using a pop up information box mechanism, which could be opened up, moved around the screen or minimised as required. The overall facility value on paper (0.41) was moderate, but higher than the computer facility (0.34).The Rasch item characteristic curves shown below in Figure 61 showed a consistent difference in performance across the student ability range in favour of paper

Figure 61: Rasch item characteristic curves for Q11c in paper and computer modes



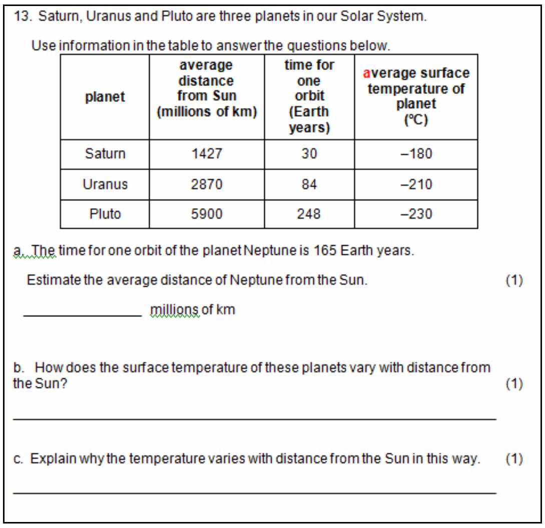
```
  ID   I0107
---------------
-3.000  0.01
-2.900  0.01
-2.800  0.01
-2.700  0.02
-2.600  0.02
-2.500  0.02
```

```
Analysis of Variance for ITEM 107 [I0107:T2_27CC_OR_C_0]
=====================================================================
   SOURCE                 S.S      DF       M.S       F-RATIO      Prob
---------------------------------------------------------------------
BETWEEN                 64.759     19     3.408
  ANOVA-Fit[CInt]       11.311      9     1.257      1.261913   0.253645
  DIF[Test2Cond]        43.703      1    43.703     43.881360   0.000000
```

Table 47: Performance data on Q13a

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13a	32AA	OR P T	0.45	0.41	0.533	0.461	0.000 P

Figure 62: Screenshots of Q13a, b and c in paper and computer mode



13. Saturn, Uranus and Pluto are three planets in our Solar System.

Use information in the table to answer the questions below.

planet	average distance from Sun (millions of km)	time for one orbit (Earth years)	average surface temperature of planet (°C)
Saturn	1427	30	−180
Uranus	2870	84	−210
Pluto	5900	248	−230

a. The time for one orbit of the planet Neptune is 165 Earth years.

Estimate the average distance of Neptune from the Sun. (1)

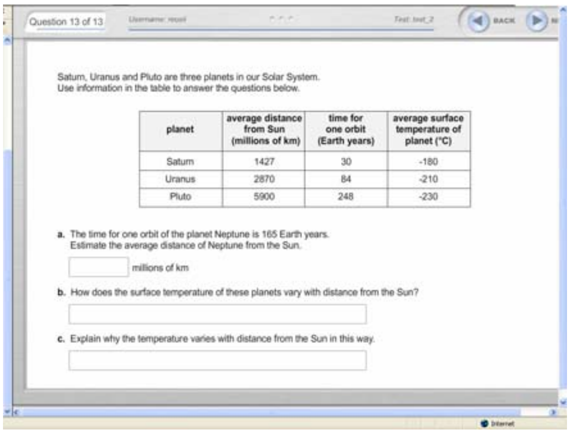
_____________ millions of km

b. How does the surface temperature of these planets vary with distance from the Sun? (1)

____

c. Explain why the temperature varies with distance from the Sun in this way. (1)

____



Saturn, Uranus and Pluto are three planets in our Solar System.
Use information in the table to answer the questions below.

planet	average distance from Sun (millions of km)	time for one orbit (Earth years)	average surface temperature of planet (°C)
Saturn	1427	30	-180
Uranus	2870	84	-210
Pluto	5900	248	-230

a. The time for one orbit of the planet Neptune is 165 Earth years.
Estimate the average distance of Neptune from the Sun.

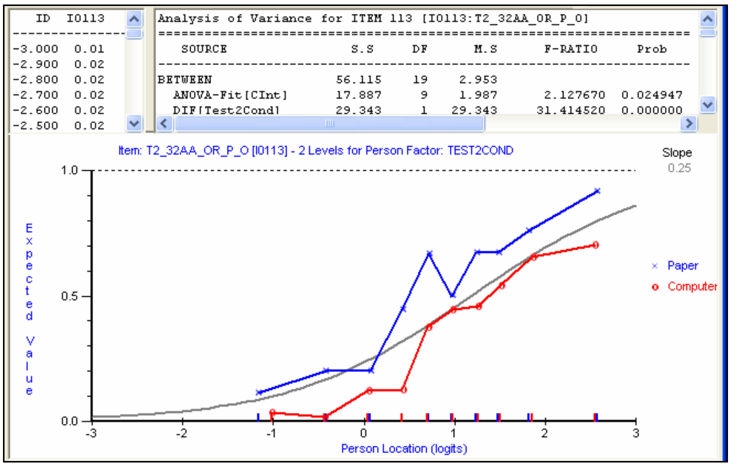
millions of km

b. How does the surface temperature of these planets vary with distance from the Sun?

c. Explain why the temperature varies with distance from the Sun in this way.

This item is the first of three in a question about planetary data and they were all targeted at a high level of difficulty. A data table was given in both modes and was required to answer the items. All three items were open responses, in this particular item a number. The overall facility value for this item was slightly higher on paper (0.45) than on computer (0.41), although the discrimination values in both modes were very good. The Rasch item characteristic curves shown below in Figure 63 showed a consistent difference in performance across the student ability range in favour of paper.

Figure 63: Rasch item characteristic curves for Q7f in paper and computer modes



```
  ID   I0113
---------------
-3.000  0.01
-2.900  0.02
-2.800  0.02
-2.700  0.02
-2.600  0.02
-2.500  0.02
```

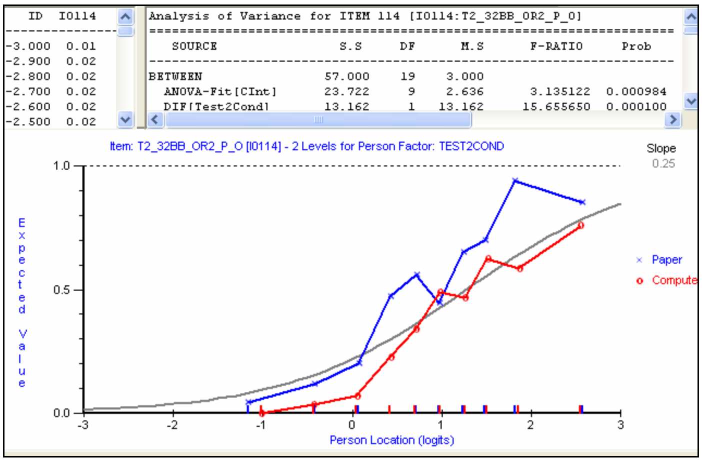
```
Analysis of Variance for ITEM 113 [I0113:T2_32AA_OR_P_O]
==========================================================================
  SOURCE                 S.S     DF      M.S       F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                56.115    19    2.953
  ANOVA-Fit[CInt]      17.887     9    1.987      2.127670  0.024947
  DIF[Test2Cond]       29.343     1   29.343     31.414520  0.000000
```

Table 48: Performance data on Q13b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13b	32BB	OR2 P T	0.41	0.43	0.526	0.474	0.000 P

This was the second item of Question 13 and again targeted at a high level of difficulty. This item required an open response in both modes, based on information given in a table. The overall facility values were similar in both modes achieving reasonable success and the discrimination values were very good. The Rasch item characteristic curves shown below in Figure 64 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 64: Rasch item characteristic curves for Q7f in paper and computer modes



```
  ID   I0114
----------------
-3.000   0.01
-2.900   0.02
-2.800   0.02
-2.700   0.02
-2.600   0.02
-2.500   0.02
```

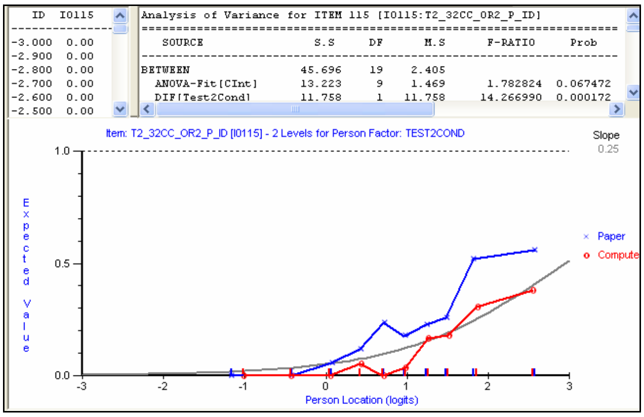
```
Analysis of Variance for ITEM 114 [I0114:T2_32BB_OR2_P_O]
================================================================================
   SOURCE                  S.S      DF      M.S        F-RATIO      Prob
--------------------------------------------------------------------------------
BETWEEN                  57.000     19     3.000
  ANOVA-Fit[CInt]        23.722      9     2.636       3.135122   0.000984
  DIF[Test2Cond]         13.162      1    13.162      15.655650   0.000100
```

Table 49: Performance data on Q13c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
13c	32CC	OR2 P T	0.16	0.15	0.375	0.342	0.000 P

This was the final item on the test, a difficult item requiring students to give an open response explanation for temperature variance with distance from the Sun. The facilities in both modes were low indicating that this item was demanding, however the discrimination values were quite good. The Rasch item characteristic curves shown below in Figure 65 showed a consistent difference in performance across the student ability range in favour of paper.

Figure 65: Rasch item characteristic curves for Q7f in paper and computer modes



```
  ID   I0115
----------------
-3.000   0.00
-2.900   0.00
-2.800   0.00
-2.700   0.00
-2.600   0.00
-2.500   0.00
```

```
Analysis of Variance for ITEM 115 [I0115:T2_32CC_OR2_P_ID]
================================================================================
   SOURCE                  S.S      DF      M.S        F-RATIO      Prob
--------------------------------------------------------------------------------
BETWEEN                  45.696     19     2.405
  ANOVA-Fit[CInt]        13.223      9     1.469       1.782824   0.067472
  DIF[Test2Cond]         11.758      1    11.758      14.266990   0.000172
```

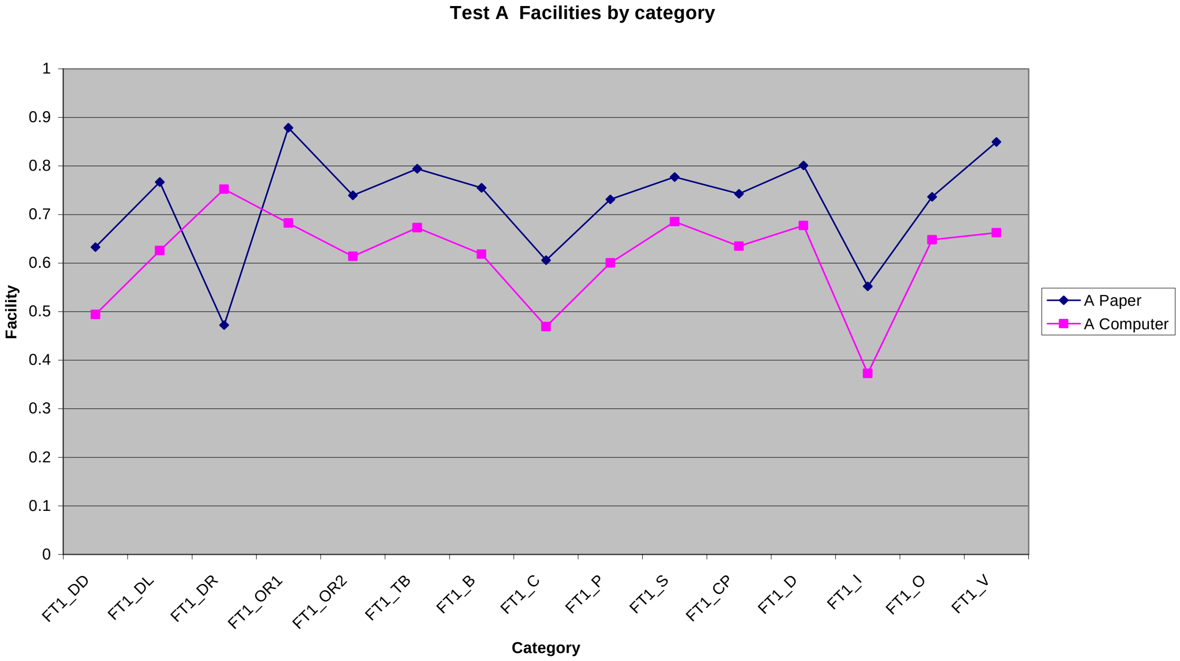
7.13 Performance by item

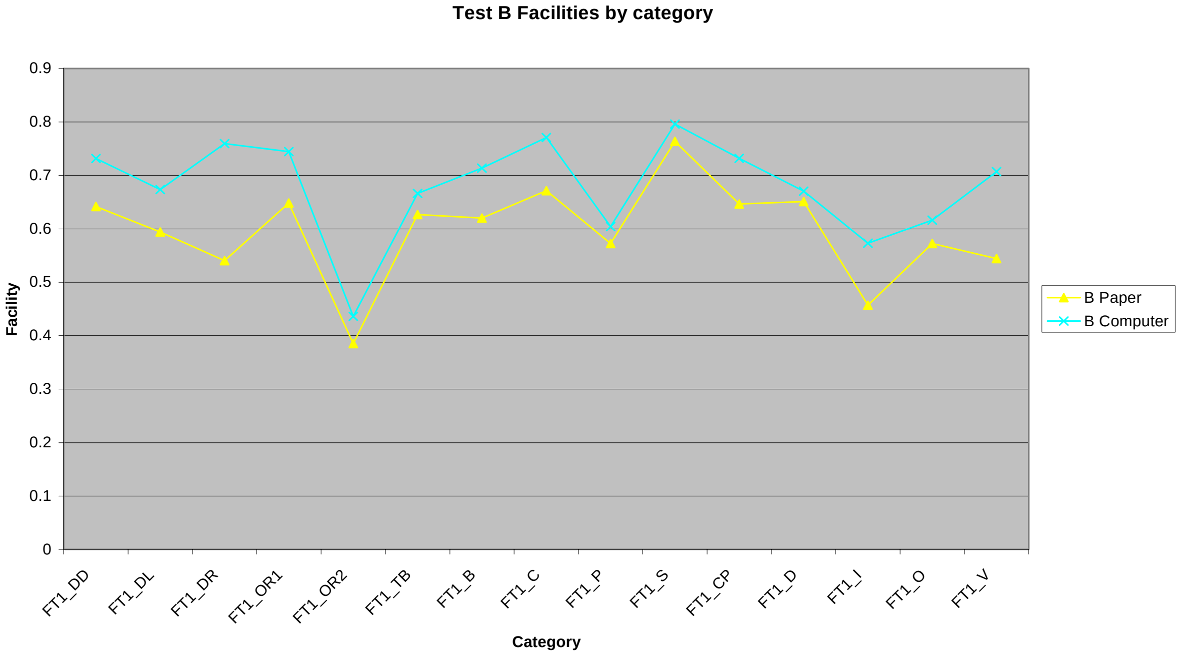
In order to investigate performance of different question types each question was put into one of three categories; the subject they assessed, the onscreen stimuli and the response type. Each category included a variety of types. The average facility for each question in each of the categories was then calculated. These are shown in Table 50 below.

Table 50: Facility Performance of Questions by Response, Subject and Question Stimulus Categories

				Facilities for Test A			Facilities for Test B		Average Facility values	
Category	**Type**	**Code**	**Number of items**	**Test A Paper**	**Test A Computer**	**Number of items**	**Test B paper**	**Test B Computer**	**Test Paper**	**Test Computer**
Response type	Drag and drop	DD	33	0.63	0.49	24	0.64	0.73	0.64	0.61
	Drop down list	DL	26	0.77	0.63	15	0.58	0.67	0.68	0.65
	Draw	DR	1	0.47	0.76	1	0.53	0.76	0.50	0.76
	Open response (Numeric)	OR		a	A	5	0.45	0.52		
	Open response (Single word)	OR1	1	0.88	0.69	7	0.64	0.74	0.76	0.72
	Open response (Extended writing)	OR2	10	0.74	0.61	5	0.38	0.44	0.56	0.53
	Tick box	TB	10	0.79	0.67	20	0.61	0.67	0.7	0.67
Subject	Biology	B	27	0.76	0.62	28	0.61	0.72	0.69	0.67
	Chemistry	C	21	0.61	0.47	31	0.67	0.78	0.64	0.63
	Physics	P	27	0.73	0.6	16	0.56	0.61	0.65	0.61
	Science1	S	6	0.78	0.69	2	0.76	0.8	0.77	0.75
Stimulus	Colour photo/ drawing	CP	22	0.74	0.63	21	0.64	0.73	0.69	0.68
	Diagram/ drawing	D	24	0.8	0.68	18	0.65	0.67	0.73	0.68
	Diagram and Information box	DI	1			2	0.65	0.78		
	Information box	I	22	0.55	0.37	4	0.45	0.58	0.50	0.48
	Interactive diagram	ID				1	0.16	0.15		
	No stimulus	O	11	0.74	0.65	16	0.56	0.62	0.65	0.64
	Table	T		0.69	0.52	12	0.58	0.69	0.64	0.61
	Video	V	2	0.85	0.66	2	0.54	0.71	0.70	0.69

Figure 66: Facilities by category for Tests A and B





The graphs above show the average facility values by Response, Subject and Stimulus of Tests A and B (the codes used are explained in the Table 39). In Test A all categories show consistently better performance on paper than computer except for the Drawing category (there were only 2 questions in this category involving a circuit on computer).

In Test B, all categories showed consistently better performance on computer than paper, although the differences were smaller than for Test A, as described earlier.

7.14 The Investigations

As described earlier in this chapter, Chapters 2 and 5 outlined the uses of internal reliability measures in tests. They are an indication of how well items within tests correlate with each other and with total student scores assuming that the test is measuring a particular construct. They operate on the premise that there will be consistency of performance across a test by students. Therefore there will be a predictable gradation of student performance across the ability range.

7.14.1 Internal reliability of the investigations

The internal reliability indicator I used in this study was Cronbach's Alpha co-efficient, and the results for the Investigations versions A and B are shown below in Table 51.

Table 51: Internal Reliability Measures of the Investigations

Investigation 1

Investigation mode	Cronbach's Alpha	N of Items
Paper	0.677	6
Computer	0.594	7

Investigation 2

Investigation Mode	Cronbach's Alpha	N of Items
Paper	0.664	6
Computer	0.595	7

The Cronbach's alpha co-efficient for both Investigation versions are shown in Table 51 above. It is clear that these values are low. However, the low numbers of measured items in these tests make the calculation of Cronbach's alpha not secure and not really an appropriate measurement of reliability.

7.14.2 The Standard Error of Measurement of the Investigations (SEM)

Table 52: Standard Error of Measurement for the Investigations (SEM)

Investigation Version	Test Group	Mode	Cronbach's alpha	Standard deviation (SD)	Standard error of measurement (SEM)
A	1	Paper	0.677	2.11	1.20
A	2	Computer	0.594	2.17	1.38

B	2	Paper	0.664	2.06	1.19
B	1	Computer	0.595	2.35	1.50

The SEM for each investigation version was consistent for each student group who had taken a matched pair of paper and on-screen tests. The SEM figure estimates the potential error of the mean scores of student relative to a theoretical true mean, therefore the smaller this figure, the less the theoretical error of student scores on the test.

7.15 Investigation Statistics

The following section of this chapter will show the performance across Investigations A and B at whole investigation and item level.

Table 53 below shows the summary statistics of Investigation A. The performance of the items in the paper and on-screen versions are shown alongside each other. The facility and discrimination values for all items in both modes are shown. In addition to this information, significant differences in performance across the modes are also shown, and highly significant differences (DIFs) are indicated, together with the mode that these DIFs favoured.

Within Table 53, the comparative facilities and discrimination values for the same items across modes are shown. Facility value is essentially a performance indicator, showing the percentage of students getting an item or marks within items correct. Discrimination values indicate the correlation between student performance on a particular item to performance across the test as a whole, therefore the amount of correlated differentiation.

Table 53 also shows any significant levels of difference between the performance of items in paper and on-screen modes. For my research purposes, I concentrated on very highly significant differences, ones where $p<0.005$. This are classified as DIF items (differential item functioning).

Table 53: Item Statistics for Investigation A, showing facility and discrimination values

Paper Question Number	Computer Question Number	Items showing significant difference at 0.005 in named mode	DIF Significant differences for these items p< 0.005	Significance values for all items	Facility values for Paper version	Facility values for Computer version	Discrimination For Paper version	Discrimination For Computer version
	I1_BB					0.77		0.299
1a	I1_CC			0.407	0.8	0.73	0.425	0.182

1b	I1_DD			0.129	0.65	0.49	0.378	0.273
1b	I1_EE			0.048	0.68	0.51	0.516	0.379
1b	I1_FF			0.764	0.18	0.11	0.283	0.308
1c	I1_GG	Paper	0.000	0.000	0.47	0.24	0.460	0.370
1d	I1_HH	Paper	0.000	0.000	0.28	0.12	0.280	0.380

Table 54 shows all the DIF items from Investigation A together. Their facility and discrimination values in both modes are shown, and also their level of DIF in terms of a p value <0.005.

Table 54: Investigation A: Statistics showing DIFs (differential item performance) at p<0.005 Sig level.

Paper Q. No.	Computer Q. No	Paper Q. facility	Computer Q. Facility	Diff in Performance p<0.005
1c	GG	0.47	0.24	0.000 P
1d	HH	0.28	0.12	0.000 P

Table 55 below shows the summary statistics of Investigation B. The performance of the items in the paper and on-screen versions are shown alongside each other. The facility and discrimination values for all items in both modes are shown. In addition to this information, significant differences in performance across the modes are also shown, and highly significant differences (DIFs) are indicated, together with the mode that these DIFs favoured.

Within Table 55, the comparative facilities and discrimination values for the same items across modes are shown. Facility value is essentially a performance indicator, showing the percentage of students getting an item or marks within items correct. Discrimination values indicate the correlation between student performance on a particular item to performance across the test as a whole, therefore the amount of correlated differentiation.

Table 55 also shows any significant levels of difference between the performance of items in paper and on-screen modes. For my research purposes, I concentrated on very highly significant differences, ones where p<0.005. This are classified as DIF items (differential item functioning).

Table 55: Item Statistics for Investigation B, showing facility and discrimination values

Paper Question Number	Computer Question Number	Items showing significant difference at 0.005 in named mode	DIF Significant differences for these items P<0.005	Significance values for all items	Facility values for Paper version	Facility values for Computer version	Discrimination for paper version	Discrimination for computer version
	I2_BB					0.71		0.309
1a	I2_CC			0.704	0.78	0.88	0.468	0.308
1b	I2_DD			0.379	0.45	0.62	0.409	0.320
1b	I2_EE	Paper	0.000	0.000	0.54	0.58	0.585	0.389
1b	I2_FF			0.646	0.17	0.3	0.336	0.174
1c	I2_GG	Paper	0.000	0.000	0.39	0.37	0.525	0.451
1d	I2_HH	Paper	0.000	0.000	0.21	0.24	0.279	0.325

Table 56 shows all the DIF items from Investigation B together. Their facility and discrimination values in both modes are shown, and also their level of DIF in terms of a p value <0.005.

Table 56: Investigation B. Statistics showing differential item performance at 0.005 Sig level.

Paper Q. No.	Computer Q. No	Paper Q. facility	Computer Q. Facility	Diff in Performance P<0.005
1b	EE	0.54	0.58	0.000 P
1c	GG	0.39	0.37	0.000 P
1d	HH	0.21	0.21	0.000 P

The following section of this chapter will explore the Rasch analyses of Investigations A and B as a whole and then look at the identified items within the Investigations that had highly significant DIF performances across modes.

For each DIF item, the paper and onscreen facility and discrimination values will be shown, alongside the level and mode of DIF. Screen shots of how these items looked in paper and on-screen modes will also be shown together with a brief description of the items and their performance. Rasch item characteristic curves will also be shown for all the DIF items, visually showing how these items performed across the ability range of the student groups.

7.16 Overall performance and relationship to ability for Investigations A and B

For each of the two investigations, Rasch analysis was used to represent performance on tests as whole and individual item characteristic curves.

The rounded DIF figure of p = 0.000 for Investigation A is shown in Figure 67 below. This indicates a highly significant differential performance (DIF) between students taking this Investigation on paper and computer in favour of the paper Investigation

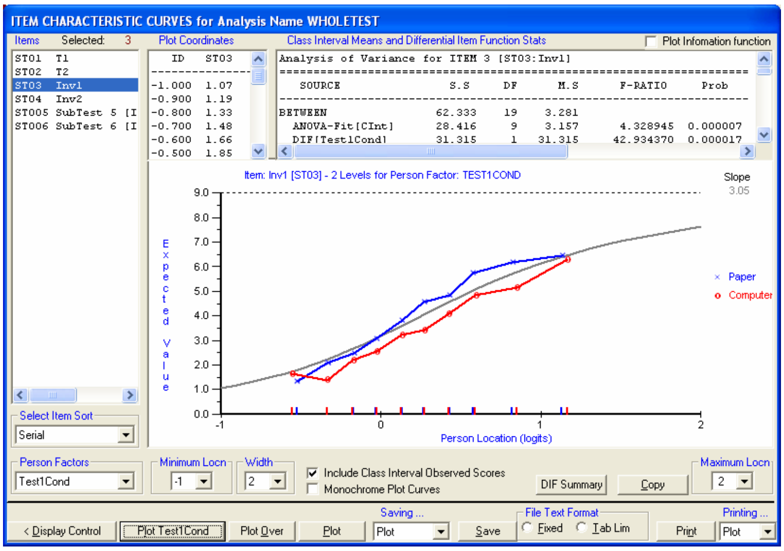
In the case of Investigation A, as shown below in Figure 67, the Rasch item characteristic curves shows a fairly consistent difference in performance (the Y axis) across the the student ability range (the X axis).

For each of the two investigations, Rasch analysis was used to represent performance on the investigations tests as whole and individual item characteristic curves.

Investigation A

The rounded DIF figure of 0.000 for Investigation A (shown in the Rasch chart below as Inv 1) indicates a highly significant differential performance between students taking this investigation on paper and computer in favour of the paper version.

Figure 67: Rasch item characteristic curve for Investigation A in paper and computer modes



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

Items Selected: 3

```
ST01   T1
ST02   T2
ST03   Inv1
ST04   Inv2
ST005  SubTest 5 [I
ST006  SubTest 6 [I
```

Plot Coordinates

```
   ID    ST03
--------------
-1.000  1.07
-0.900  1.19
-0.800  1.33
-0.700  1.48
-0.600  1.66
-0.500  1.85
```

Class Interval Means and Differential Item Function Stats

```
Analysis of Variance for ITEM 3 [ST03:Inv1]
=====================================================================
  SOURCE                  S.S     DF      M.S       F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN                 62.333    19     3.281
  ANOVA-Fit[CInt]       28.416     9     3.157     4.328945  0.000007
  DIF[Test1Cond]        31.315     1    31.315    42.934370  0.000017
```

Select Item Sort: Serial

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

□ Monochrome Plot Curves

Maximum Locn: 2

Saving ...: Plot

File Text Format: ○ Fixed ○ Tab Lim

Printing ...: Plot

The following charts are the individual items characteristic curves within Investigation A, identified in Table 43 showing significant differential performance between paper and computer modes at a level of p< 0.005.

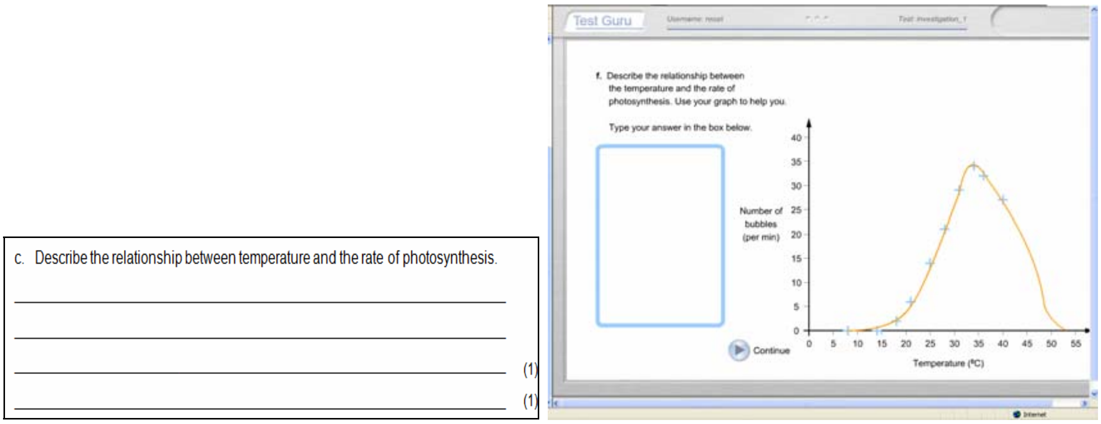
Discussion of the differences between modes will be carried out in the Analysis chapter. This section will indicate the nature of the differences for each DIF item within Investigations A and B.

7.17 DIF items within Investigation A

Table 57: Performance data on Investigation A, Q1c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
1c	GG	OR	0.47	0.24	0.460	0.370	0.000 P

Figure 68: Screenshots of Q1c in paper and computer mode



c. Describe the relationship between temperature and the rate of photosynthesis.

(1)

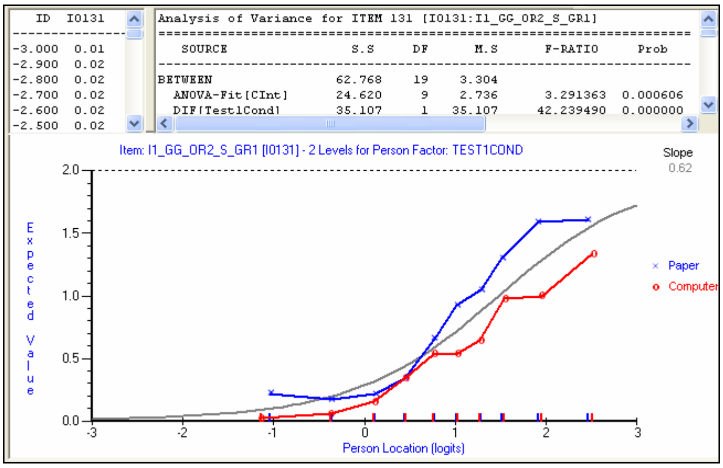
(1)

f. Describe the relationship between the temperature and the rate of photosynthesis. Use your graph to help you.

Type your answer in the box below.

This item was targeted at a medium level of difficulty and involved students providing an open response, describing the relationship of the two variables using their plotted graph. In the paper version students had their plotted graph on the opposite page to this item. The computer version presented their graph alongside the answer space. There was plenty of space provided on paper and on computer for students to write an open response. The overall facility value on paper (0.47) was twice as high as on computer (0.24). The Rasch item characteristic curves shown below in Figure 69 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 69: Rasch item characteristic curve for Q1c in paper and computer modes



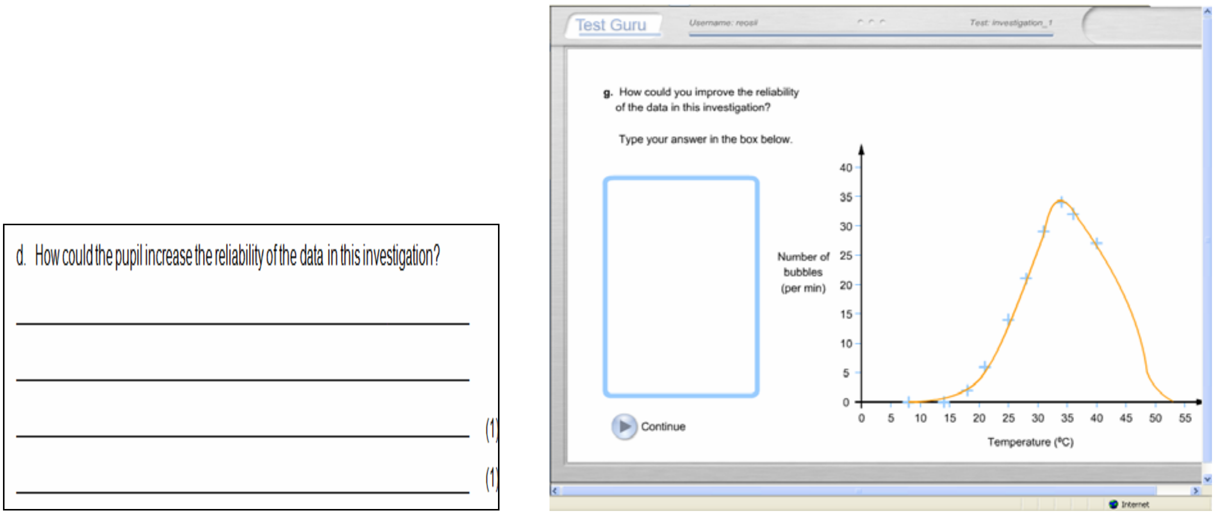
```
  ID   I0131
----------------
-3.000  0.01
-2.900  0.02
-2.800  0.02
-2.700  0.02
-2.600  0.02
-2.500  0.02
```

```
Analysis of Variance for ITEM 131 [I0131:I1_GG_OR2_S_GR1]
======================================================================
   SOURCE                  S.S      DF      M.S        F-RATIO      Prob
----------------------------------------------------------------------
BETWEEN                  62.768     19     3.304
  ANOVA-Fit[CInt]        24.620      9     2.736      3.291363  0.000606
  DIF[Test1Cond]         35.107      1    35.107     42.239490  0.000000
```

Table 58: Performance data on Investigation A, Q1d

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
1d	HH	OR	0.28	0.12	0.280	0.380	0.000P

Figure 70: Screenshots of Q1d in paper and computer mode



d. How could the pupil increase the reliability of the data in this investigation?

(1)

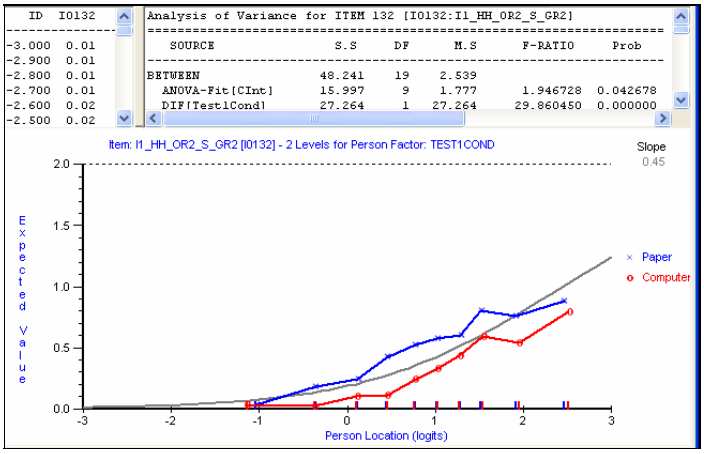
(1)

g. How could you improve the reliability of the data in this investigation?

Type your answer in the box below.

This item was targeted at a high level of difficulty. requiring students to give an open response answer, commenting on how the reliability of the investigation could be improved. The overall paper facility for paper (0.28) was significantly higher than on computer (0.12). The Rasch item characteristic curves shown below in Figure 71 showed a consistent difference in performance across the student ability range in favour of paper.

Figure 71: Rasch item characteristic curve for Q1d in paper and computer modes



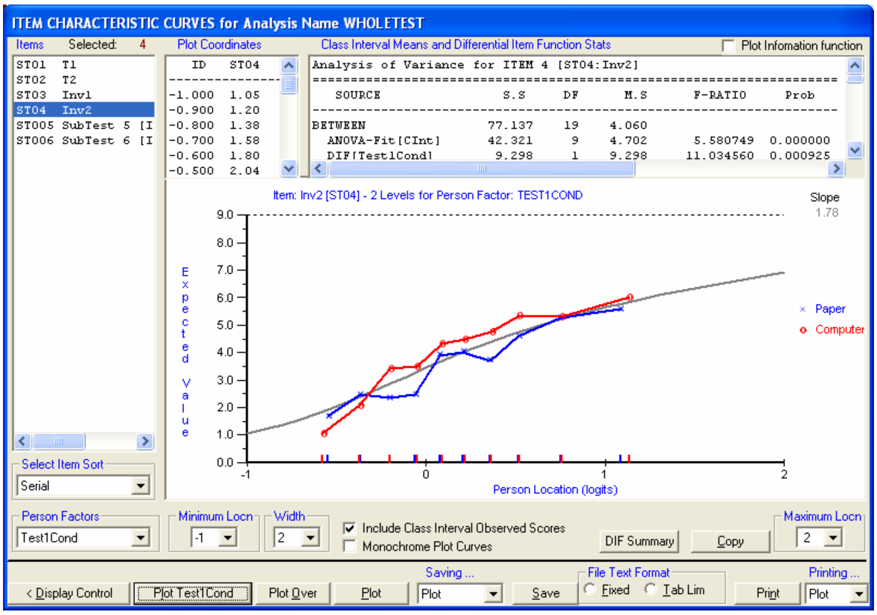
```
  ID   I0132
---------------
-3.000  0.01
-2.900  0.01
-2.800  0.01
-2.700  0.01
-2.600  0.02
-2.500  0.02
```

```
Analysis of Variance for ITEM 132 [I0132:I1_HH_OR2_S_GR2]
======================================================================
   SOURCE              S.S      DF      M.S       F-RATIO     Prob
----------------------------------------------------------------------
BETWEEN              48.241     19     2.539
  ANOVA-Fit[CInt]    15.997      9     1.777     1.946728   0.042678
  DIF[Test1Cond]     27.264      1    27.264    29.860450   0.000000
```

7.18 Performance of Investigation B

The rounded DIF figure of 0.001 for Investigation B (shown in the Rasch chart below as Inv 2) indicates a highly significant differential performance between students taking this investigation on paper and computer in favour of the computer version.

Figure 72: Rasch item characteristic curve for Investigation B in paper and computer modes



```
ST01   T1
ST02   T2
ST03   Inv1
ST04   Inv2
ST005  SubTest 5 [I
ST006  SubTest 6 [I
```

```
   ID     ST04
---------------
-1.000   1.05
-0.900   1.20
-0.800   1.38
-0.700   1.58
-0.600   1.80
-0.500   2.04
```

```
Analysis of Variance for ITEM 4 [ST04:Inv2]
=============================================================================
   SOURCE                 S.S      DF      M.S       F-RATIO      Prob
-----------------------------------------------------------------------------
BETWEEN                 77.137     19     4.060
  ANOVA-Fit[CInt]       42.321      9     4.702      5.580749   0.000000
  DIF[Test1Cond]         9.298      1     9.298     11.034560   0.000925
```

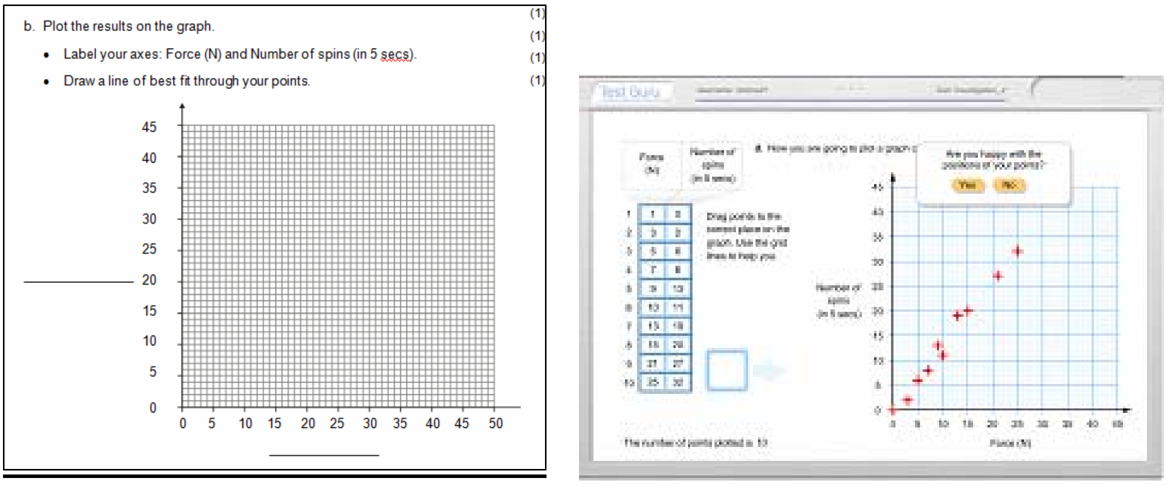
The following charts are the individual items characteristic curves within Investigation B showing significant differential performance between paper and computer modes at a level of at least p< 0.005.

7.19 DIF items within Investigation B

Table 59: Performance data on Investigation B, Q1b

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
1b	EE	DR	0.54	0.58	0.585	0.389	0.000P

Figure 73: Screenshots of Q1b in paper and computer mode



b. Plot the results on the graph. (1)

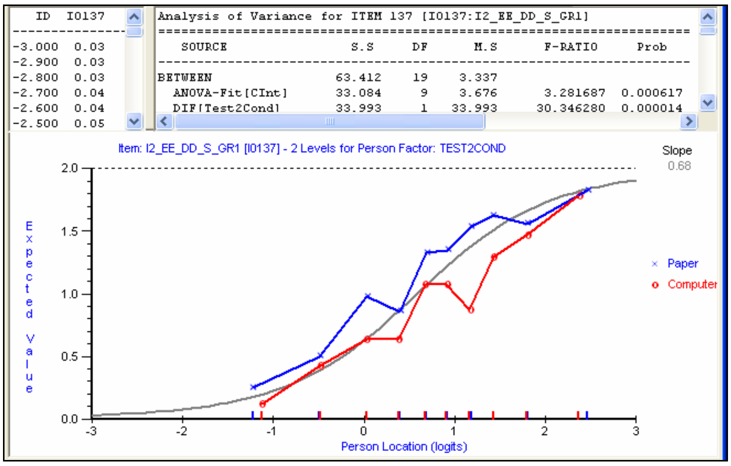
- Label your axes: Force (N) and Number of spins (in 5 secs). (1)
- Draw a line of best fit through your points. (1)

(1)

45
40
35
30
25
20
15
10
5
0
0 5 10 15 20 25 30 35 40 45 50
Test Guru
Force (N)
Number of spins (in 5 secs)
How are you going to plot a graph
Drag points to the correct place on the graph. Use the grid lines to help you.
Are you happy with the positions of your points?
Yes
No
The number of points plotted is 10
Force (N)

This item was targeted at a medium level of difficulty and involved students plotting data onto a graph. On paper they were given data and graph paper to draw and plot their graphs, on computer students generated their own results and plotted their graph using drag and drop functions. The overall facility values in both modes were similar, although the paper version had much higher discrimination values. The Rasch item characteristic curves shown below in Figure 74 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 74: Rasch item characteristic curve for Q1b in paper and computer modes



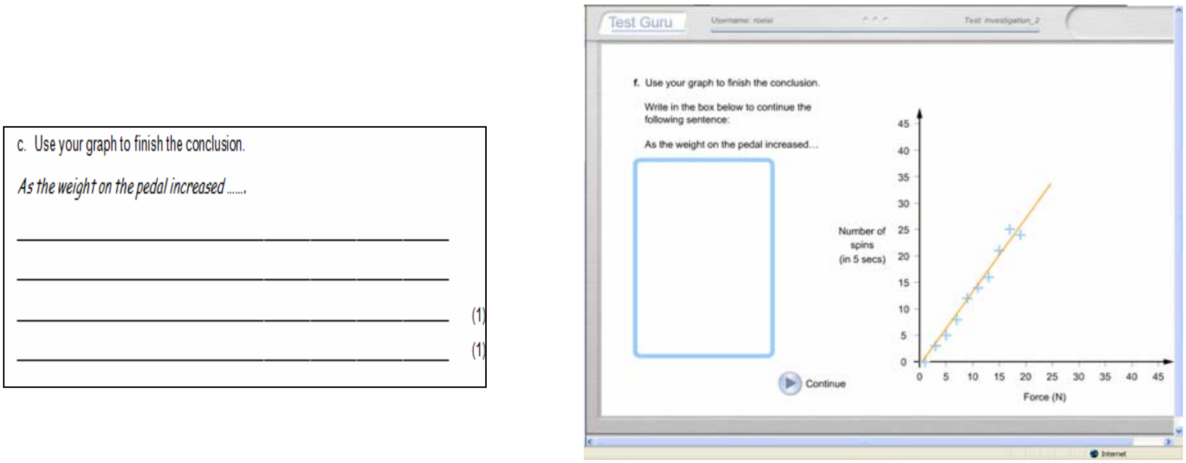
```
  ID   I0137
--------------
-3.000  0.03
-2.900  0.03
-2.800  0.03
-2.700  0.04
-2.600  0.04
-2.500  0.05
```

```
Analysis of Variance for ITEM 137 [I0137:I2_EE_DD_S_GR1]
==========================================================================
  SOURCE                S.S      DF     M.S       F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                63.412    19    3.337
  ANOVA-Fit[CInt]      33.084     9    3.676     3.281687  0.000617
  DIF[Test2Cond]       33.993     1   33.993    30.346280  0.000014
```

Table 60: Performance data on Investigation B, Q1c

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
1c	GG	OR	0.39	0.37	0.525	0.451	0.000P

Figure 75: Screenshots of Q1c in paper and computer mode



c. Use your graph to finish the conclusion.

As the weight on the pedal increased ......

(1)

(1)

Test Guru

Username: [illegible]

Test: investigation_2

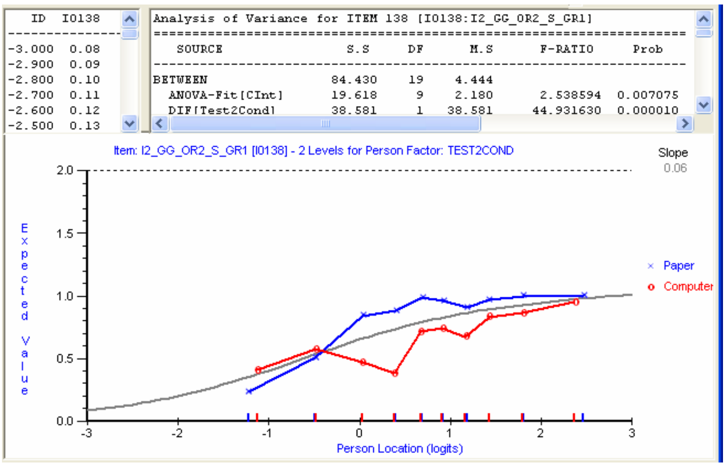
f. Use your graph to finish the conclusion.

Write in the box below to continue the following sentence:

As the weight on the pedal increased...

This item was targeted at a medium level of difficulty and involved students providing an open response, describing the relationship of the two variables using their plotted graph. In the paper version students had their plotted graph on the opposite page to this item. The computer version presented their graph alongside the answer space. There was plenty of space provided on paper and on computer for students to write an open response. The overall facility value on paper and computer were similar and the discrimination values were good in both modes. The Rasch item characteristic curves shown below in Figure 76 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 76: Rasch item characteristic curve for Q1c in paper and computer modes



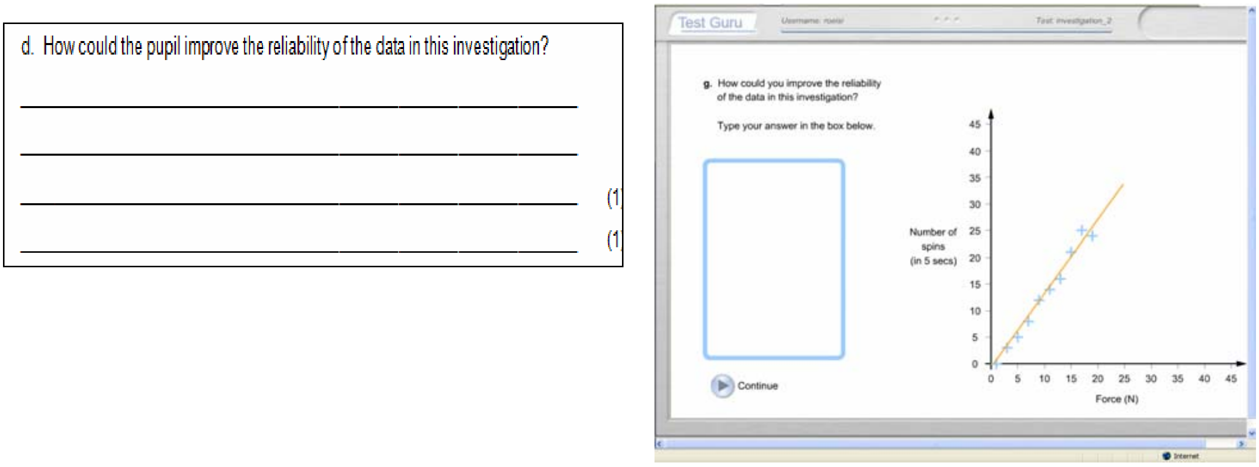
```
  ID   I0138
---------------
-3.000  0.08
-2.900  0.09
-2.800  0.10
-2.700  0.11
-2.600  0.12
-2.500  0.13
```

```
Analysis of Variance for ITEM 138 [I0138:I2_GG_OR2_S_GR1]
=====================================================================
  SOURCE                S.S     DF      M.S      F-RATIO     Prob
---------------------------------------------------------------------
BETWEEN               84.430    19    4.444
  ANOVA-Fit[CInt]     19.618     9    2.180     2.538594  0.007075
  DIF[Test2Cond]      38.581     1   38.581    44.931630  0.000010
```

Table 61: Performance data on Investigation B, Q1d

Paper Q	Computer Q	Computer Q Type	Paper Facility	Computer Facility	Paper Discrimination	Computer Discrimination	Sig. Diff, P or C
1d	HH	OR	0.21	0.24	0.279	0.325	0.000P

Figure 77: Screenshots of Q1d in paper and computer mode



d. How could the pupil improve the reliability of the data in this investigation?

____

____

____ (1

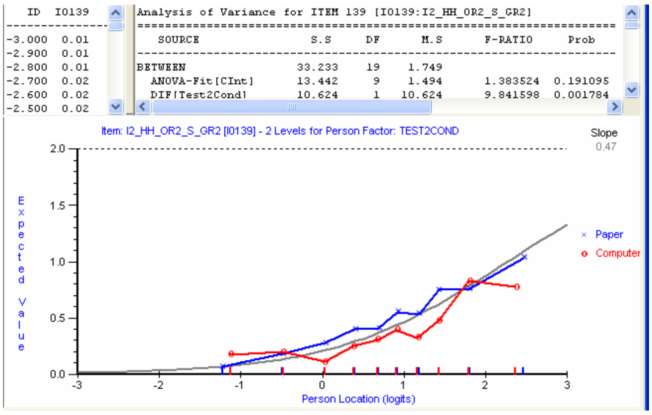
____ (1

g. How could you improve the reliability of the data in this investigation?

Type your answer in the box below.

This item was targeted at a high level of difficulty. requiring students to give an open response answer, commenting on how the reliability of the investigation could be improved. The overall paper facility for paper (0.21) was similar to computer (0.24), with both modes achieving reasonable discrimination values The Rasch item characteristic curves shown below in Figure 78 showed a fairly consistent difference in performance across the student ability range in favour of paper.

Figure 78: Rasch item characteristic curve for Q1d in paper and computer modes



```
  ID  I0139
---------------
-3.000  0.01
-2.900  0.01
-2.800  0.01
-2.700  0.02
-2.600  0.02
-2.500  0.02
```

```
Analysis of Variance for ITEM 139 [I0139:I2_HH_OR2_S_GR2]
==================================================================
   SOURCE                  S.S      DF       M.S        F-RATIO    Prob
------------------------------------------------------------------
BETWEEN                   33.233    19     1.749
  ANOVA-Fit[CInt]         13.442     9     1.494        1.383524  0.191095
  DIF[Test2Cond]          10.624     1    10.624        9.841598  0.001784
```

7.20 Overall performance- Gender

The following table shows the mean raw scores and percentages achieved by males and females, for both Tests A and B in paper and computer modes.

Table 51 shows very little difference between scores based on gender. There was no significant difference in performance, in either mode, between genders for either Test A or Test B.

Table 62: Overall performance - Gender

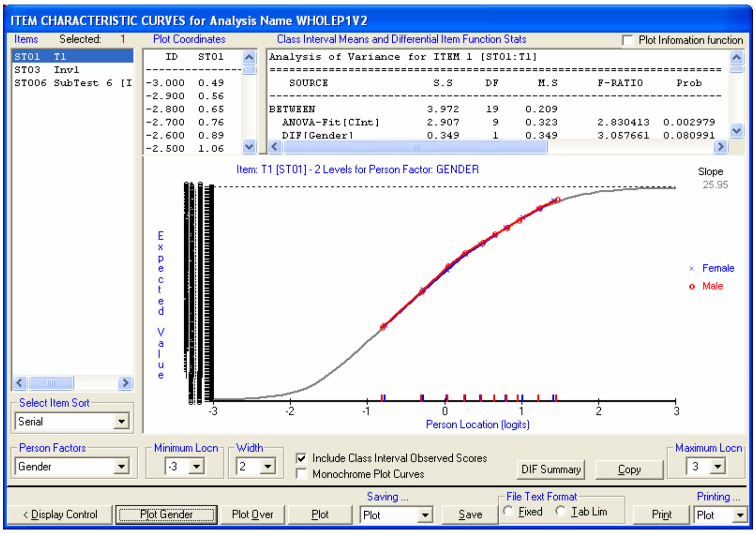
		Paper		Computer	
	Gender	Raw score	% Score	Raw score	% Score
Test A	F	57.74	71.28	46.84	57.83
	M	57.52	71.01	47.55	58.70
Test B	F	44.35	57.60	52.76	68.52
	M	46.35	60.19	51.03	66.27

For each of the two tests, and in each mode, Rasch analysis was carried out to compare the performance of males and females in whole test and individual item characteristic curves.

Paper Test A

The rounded DIF figure of 0.081 shown in Figure 79 below in the Rasch characteristic curves for Test A indicates that there was no significant difference in the performance between males and females across the whole paper Test A

Figure 79: Rasch item characteristic curve for Test A on paper in Gender mode



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLEP1V2

Items Selected: 1

```
ST01  T1
ST03  Invl
ST006 SubTest 6 [I
```

Plot Coordinates

```
  ID    ST01
---------------
-3.000  0.49
-2.900  0.56
-2.800  0.65
-2.700  0.76
-2.600  0.89
-2.500  1.06
```

Class Interval Means and Differential Item Function Stats

```
Analysis of Variance for ITEM 1 [ST01:T1]
==========================================================================
  SOURCE                S.S      DF       M.S       F-RATIO      Prob
--------------------------------------------------------------------------
BETWEEN                3.972     19     0.209
  ANOVA-Fit[CInt]      2.907      9     0.323      2.830413   0.002979
  DIF[Gender]          0.349      1     0.349      3.057661   0.080991
```

Item: T1 [ST01] - 2 Levels for Person Factor: GENDER

Slope 25.95

Select Item Sort: Serial

Person Factors: Gender

Minimum Locn: -3

Width: 2

☑ Include Class Interval Observed Scores

□ Monochrome Plot Curves

Maximum Locn: 3

File Text Format: ○ Fixed ○ Tab Lim

Saving ...: Plot

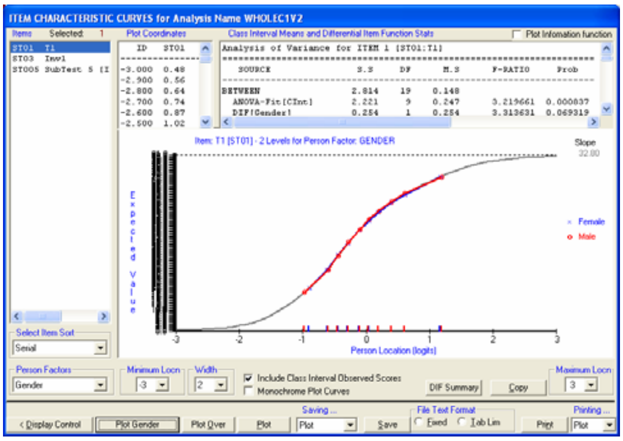
Printing ...: Plot

Although there were not any significant gender differences across paper test A as a whole, there were significant gender differences for a few items within Test A. Data, screen shots and Rasch item characteristic curves for these items are shown in Appendix P.

Computer Test A.

The following Rasch whole test curve shown below in Figure 80 below showed that there was not any significant gender difference (DIF) on the Computer A test as a whole (rounded to 0.070).

Figure 80: Rasch item characteristic curve for Test A on Computer in Gender mode



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLEC1V2

```
ST01   T1
ST03   Inv1
ST005  SubTest 5 [I
```

```
   ID    ST01
---------------
-3.000   0.48
-2.900   0.56
-2.800   0.64
-2.700   0.74
-2.600   0.87
-2.500   1.02
```

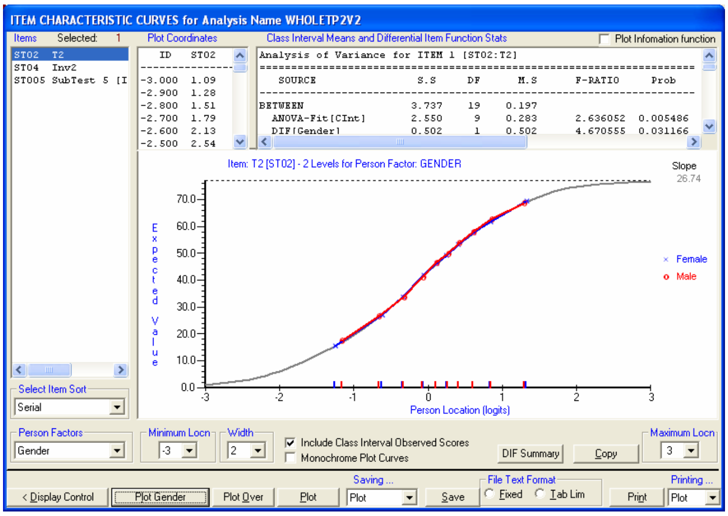
```
Analysis of Variance for ITEM 1 [ST01:T1]
==========================================================================
  SOURCE                 S.S      DF      M.S        F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                 2.814    19     0.148
  ANOVA-Fit[CInt]       2.221     9     0.247       3.219661   0.000837
  DIF[Gender]           0.254     1     0.254       3.313631   0.069319
```

Although there were not any significant gender differences across Computer Test A as a whole, there were significant gender differences for a few items within Test A. Data, screen shots and Rasch item characteristic curves for these items are shown in Appendix P.

Paper Test B

The following Rasch whole test curve shown in Figure 81 below showed that there was not any significant gender difference (DIF) on the Paper Test B as a whole (rounded to 0.031).

Figure 81: Rasch item characteristic curve for Test B on paper in Gender mode



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETP2V2

```
ST02   T2
ST04   Inv2
ST005  SubTest 5 [I
```

```
   ID    ST02
---------------
-3.000   1.09
-2.900   1.28
-2.800   1.51
-2.700   1.79
-2.600   2.13
-2.500   2.54
```

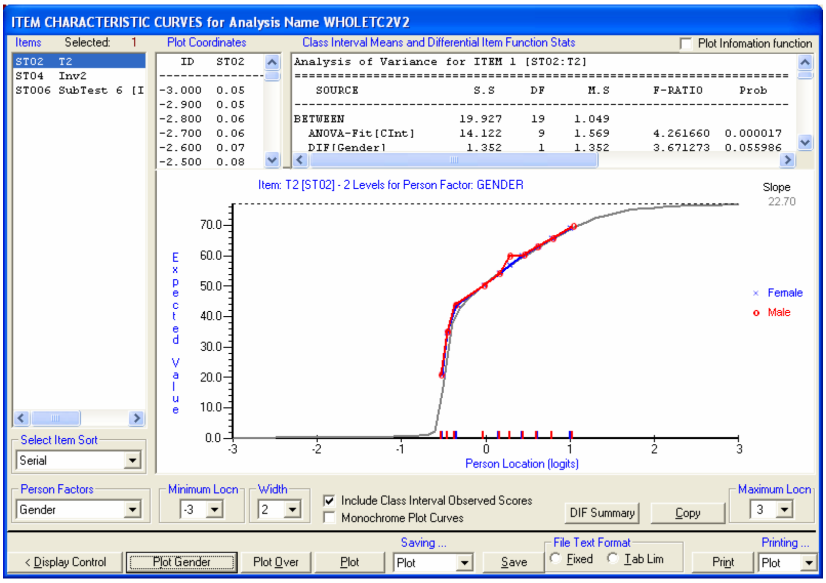
```
Analysis of Variance for ITEM 1 [ST02:T2]
==========================================================================
  SOURCE                 S.S      DF      M.S        F-RATIO     Prob
--------------------------------------------------------------------------
BETWEEN                 3.737    19     0.197
  ANOVA-Fit[CInt]       2.550     9     0.283       2.636052   0.005486
  DIF[Gender]           0.502     1     0.502       4.670555   0.031166
```

Although there were not any significant gender differences across Paper Test B as a whole, there were significant DIF gender differences for a few items within Test B. Data, screen shots and Rasch item characteristic curves for these items are shown in Appendix P.

Computer Test B

The following Rasch whole test curve shown in Figure 82 below showed that there was not any significant gender difference (DIF) on the Computer Test B as a whole (rounded to 0.060).

Figure 82: Rasch item characteristic curve for Test B on Computer in Gender mode



ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETC2V2

Items Selected: 1

```
ST02  T2
ST04  Inv2
ST006 SubTest 6 [I
```

Plot Coordinates

```
  ID     ST02
--------------
-3.000  0.05
-2.900  0.05
-2.800  0.06
-2.700  0.06
-2.600  0.07
-2.500  0.08
```

Class Interval Means and Differential Item Function Stats

```
Analysis of Variance for ITEM 1 [ST02:T2]
=============================================================================
  SOURCE                 S.S      DF      M.S        F-RATIO     Prob
-----------------------------------------------------------------------------
BETWEEN                19.927     19     1.049
  ANOVA-Fit[CInt]      14.122      9     1.569       4.261660   0.000017
  DIF[Gender]           1.352      1     1.352       3.671273   0.055986
```

Select Item Sort: Serial

Person Factors: Gender

Minimum Locn: -3

Width: 2

☑ Include Class Interval Observed Scores

☐ Monochrome Plot Curves

Maximum Locn: 3

Saving ...: Plot

File Text Format: ○ Fixed ○ Tab Lim

Printing ...: Plot

Although there were not any significant gender differences across Computer Test B as a whole, there were significant DIF gender differences for a few items within Test B. Data, screen shots and Rasch item characteristic curves for these items are shown in Appendix P.

There were not any significant gender differences in either the whole investigations or any items within them, and therefore I have not shown any data or Rasch curves.

7.21 Summary

This chapter has shown the outcomes of classical test and Rasch analysis on the paper and computer-based tests and investigations. Their performance as assessment instruments have been compared and any significant modal differences identified.
In the next chapter, the complementary qualitative evidence will be presented.

Chapter 8

Qualitative Results

8.1 Introduction

There were a number of sources of qualitative evidence within my research. These included students taking questionnaires when they had completed a paper-based test and investigation, another questionnaire when they had completed a computer-based test and investigation, and a final comparative questionnaire after they had completed both versions. In addition to questionnaires, samples of students and teachers were interviewed.

Each questionnaire (shown in appendices F, G, H, I, J,) asked students to apply ratings to questions (usually 1 indicating ease or preference and 4, the most difficulty or least preferred option). Each rating was accompanied by an open response box for students to qualify their rating with reasons if they wished to. The questionnaire data can be found in Appendix Q.

The following tables show the rating data from the three forms of questionnaires. There are a few points to note regarding this data.

- The number of students completing the first two questionnaires (the paper- and computer based tests and investigations respectively) were lower than the final comparative questionnaire. This was the result of logistical issues for the participating schools. Due to timetabling restrictions, many schools were not able to provide time for students to complete the tests and investigations and the accompanying questionnaires in a lesson. The priority was that they completed the comparative questionnaire, and therefore the first two questionnaires only had a 25% response rate, whereas the final comparative questionnaire had a response rate of 80%. Despite the relatively low response rate for the first two questionnaires, response numbers of between 220 to 320 still enabled the gathered data to be representative and reliable.
- The data shown below does not include all the questions contained in the questionnaire. The data focuses on the comparative questions. Data and information from other questions will be referred to where appropriate.
- There is a difference between 'no response' data and 'not completed' data. 'No response' indicates that the student chose not to answer a particular question,

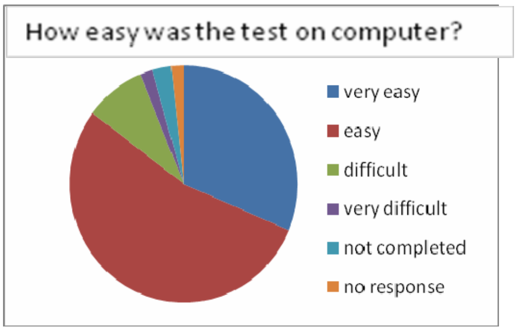
whereas 'not completed' indicated that the student did not have time to complete the questionnaire

8.2 Basic Findings of the Individual Test and Investigation Questionnaires

The questionnaire responses from the paper-based tests and investigations could be taken as control data for any comparative views on the computer assessments. Paper-assessments are the default mode of testing in this country, and the schools participating in this trial were used to taking formative and summative tests on paper, but not in on-screen forms. Questionnaire responses were also an indication of views and opinions of the content and perceived quality of the assessments. As the curriculum coverage and development of questions were the same for both versions of the tests and investigations, any differences would be accountable to attitudes to the mode of assessment.

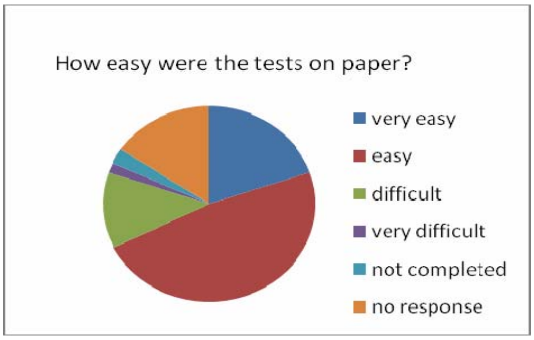
In the separate questionnaires taken straight after the paper and computer-based tests and investigations, in four out of the five central questions the computer mode was preferred to the paper-based mode in terms of the perceived ease and fitness for purpose of the assessments. Figure 83 below 85% of students found the tests very easy or easy on computer:

Figure 83: Pie Chart to show student views on ease of computer test



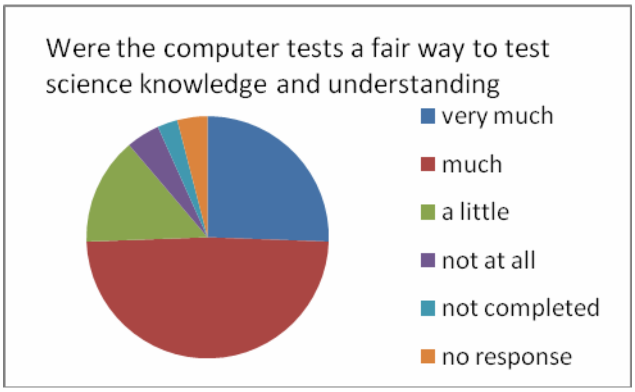
This compared with 68% finding the same tests very easy or easy on paper shown in Figure 84.

Figure 84: Pie Chart to show student views on ease of paper test



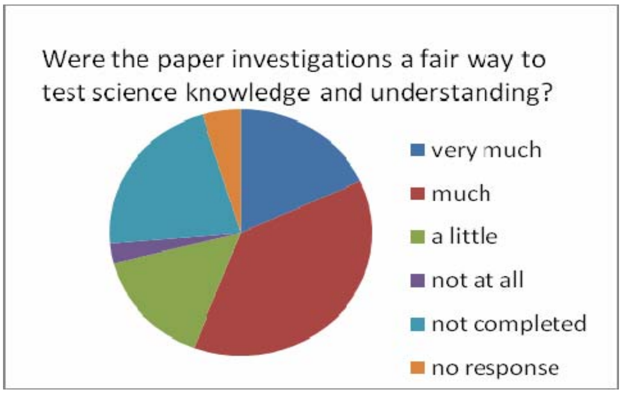
80% of students thought that the computer tests were a fair way of testing science knowledge and understanding, as shown in Figure 85.

Figure 85: Pie Chart to show student views on fairness of computer test



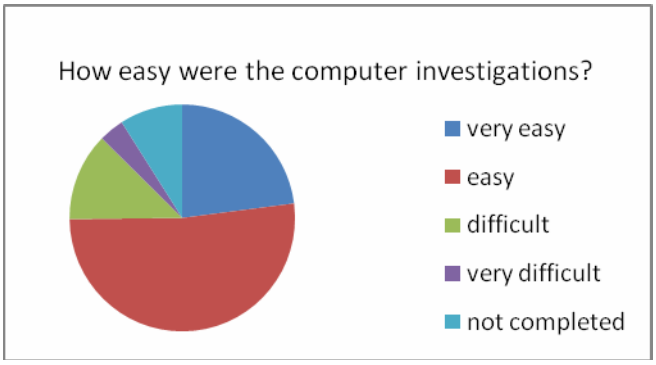
This compared to 56% who thought the paper versions were fair, shown in Figure 86.

Figure 86: Pie Chart to show student views on fairness of paper test



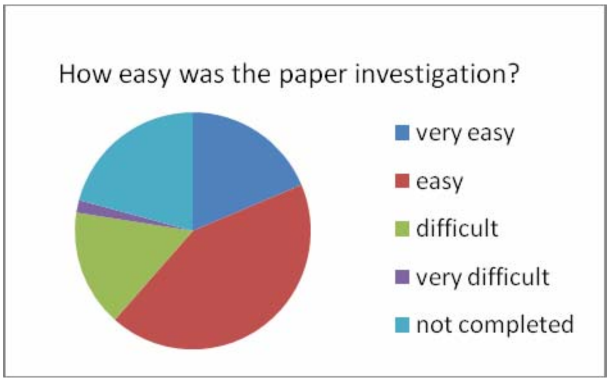
In terms of the investigations, 75% of students thought that the computer investigations were very easy or easy, as shown in figure 87.

Figure 87: Pie Chart to show student views on ease of computer investigations



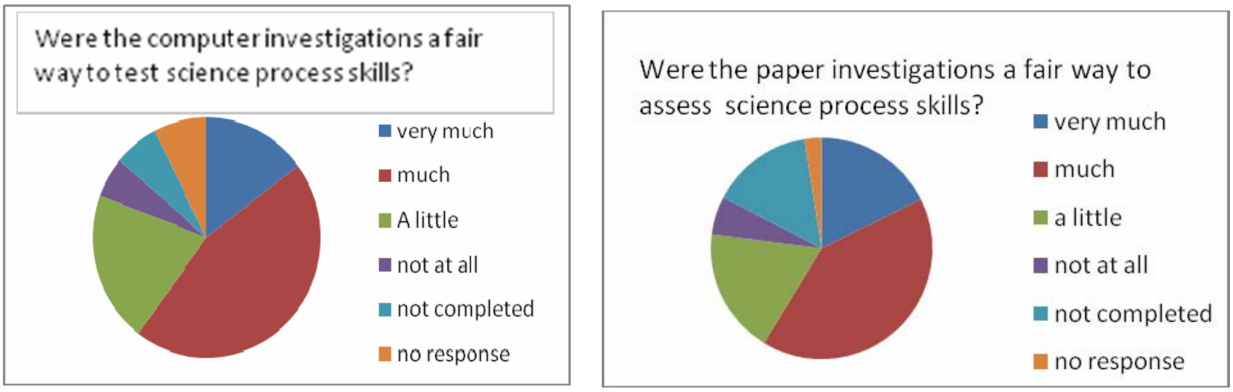
This compared to 62% who thought the paper-based versions were, as shown in Figure 88.

Figure 88: Pie Chart to show student views on ease of paper test



Interestingly, at the time of taking the investigations, students rated the computer and paper-based investigations equally when asked whether they were fair methods of testing science process skills, obtaining 60 % and 59% respectively, as shown in Figure 89.

Figure 89: Pie Charts to show student views on fairness of computer and paper-based investigations



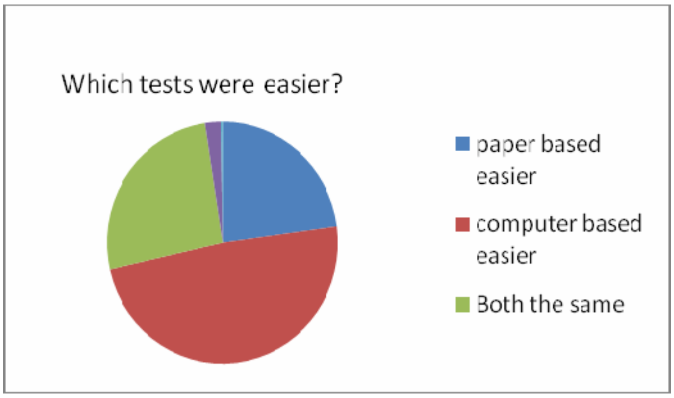
8.3 Basic Findings of the Comparative Test and Investigation Questionnaires

The comparative questionnaires, which had a sample higher completion rate of approximately 800 (80%), reflected the findings of the individual questionnaires.

8.3.1 The Tests

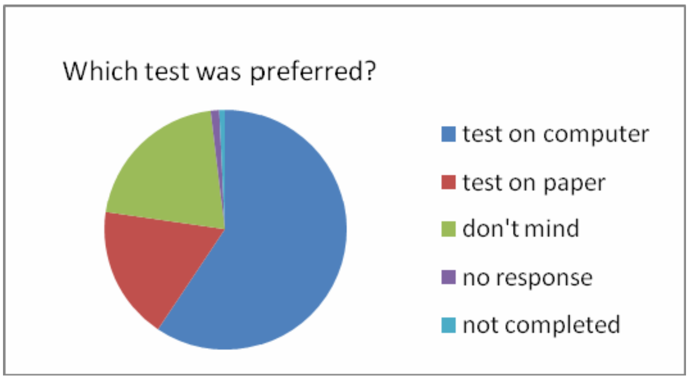
In terms of the tests, Figure 90 shows that nearly half (49%) of students thought the computer tests were easier than the paper versions. Approximately a quarter (23%) thought the paper versions were easier, with the remaining quarter of the sample (26%) rating their ease to be the same.

Figure 90: Pie Chart to show student views on the comparative ease of the paper and computer-based tests



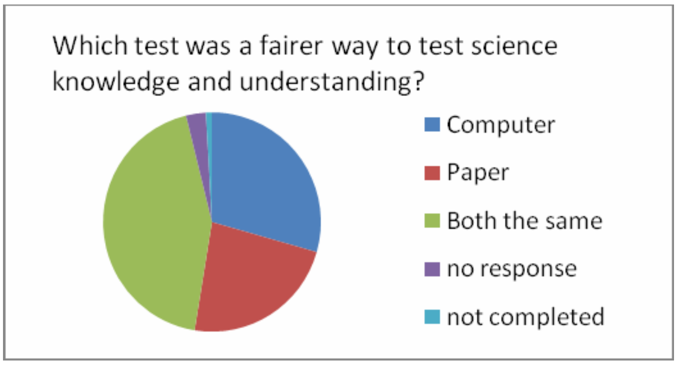
The preference for the mode of test, shown below in Figure 91, was clearly in favour of the computer versions, 59% and 18% favouring computer and paper respectively. 21% stated no preference.

Figure 91: Pie Chart to show student views on the comparative preference between the paper and computer-based tests



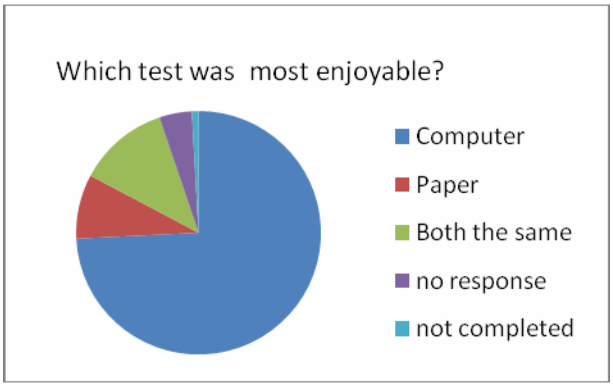
Attitudes to whether the different modes were fair ways of testing knowledge and understanding were interesting. Figure 92 below shows approximately a quarter preferring each mode (29% computer, 23% paper) and the remaining half(44%) rating them the same.

Figure 92: Pie Chart to show student views on the comparative fairness of the paper and computer-based tests



In terms of enjoyment, the computer tests were clear winners, figure 93 below showing a 74% preference. Only 9% preferred the paper tests, with the remainder rating both modes equally.

Figure 93: Pie Chart to show student views on the comparative enjoyment of the paper and computer-based tests



8.3.2 The Investigations

The findings of the comparative investigation questionnaires were similar to those of the tests, but not quite so positive in favour of the computer versions. Figure 94 below shows that 41% of students thought the investigations were easier on computer compared to 29% rating the paper versions easier. 21% rated them equally.

Figure 94: Pie Chart to show student views on the comparative ease of the paper and computer-based investigations

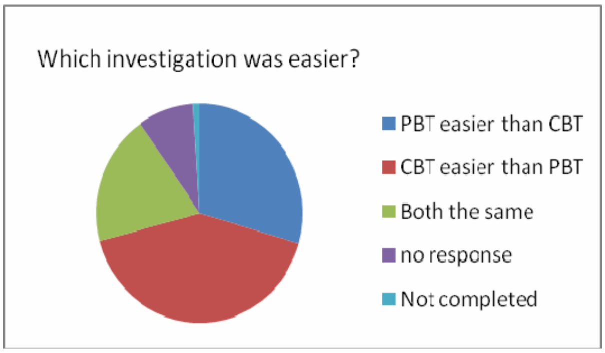
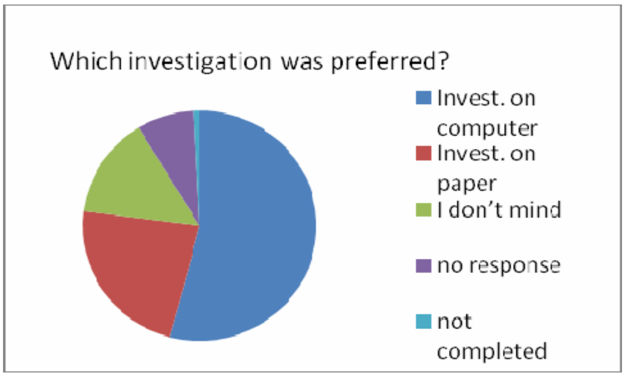


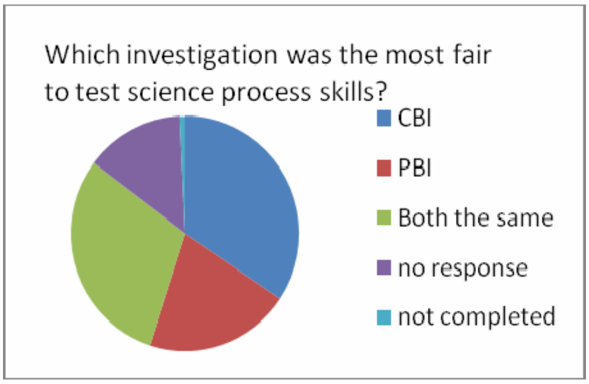
Figure 95 below shows that over half the students (54%) would prefer the computer versions as opposed to 23% who preferred the paper versions. 14% of students didn’t mind which mode was used.

Figure 95: Pie Chart to show student views on the comparative preference between the paper and computer-based investigations



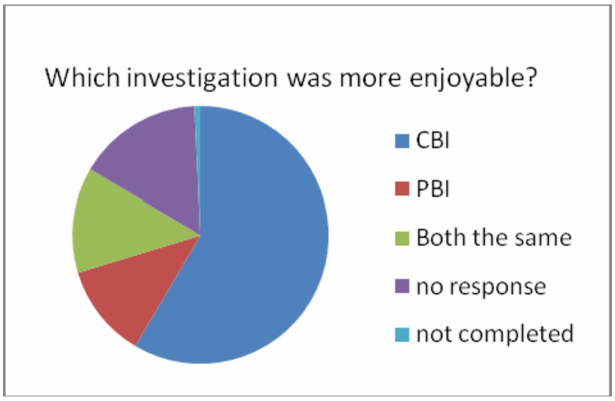
The ratings for the fairness of the assessment of scientific process skills were similar, Figure 96 below shows the computer mode scoring 34% preference, rather than the 24% preferring the paper versions. 30% of students considered them both fair.

Figure 96: Pie Chart to show student views on the comparative fairness of the paper and computer-based investigations



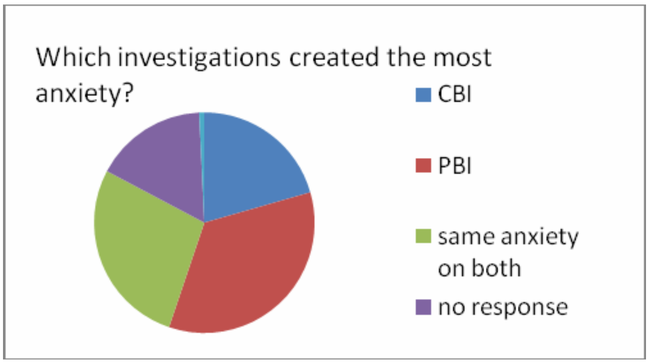
Similarly to the tests, in terms of enjoyment, Figure 97 below shows the computer versions were much more highly rated, scoring 59%, compared to 12% who enjoyed the paper versions more. 13% of students enjoyed both, and 16% could not rate the idea of enjoyment to an investigational assessment to either mode.

Figure 97: Pie Chart to show student views on the comparative enjoyment between the paper and computer-based investigations



The final comparative question asked students to rate their anxiety levels when taking these assessments in paper and computer modes. Figure 98 below shows paper-based assessments were rated as inducing more anxiety than computer versions (35% as opposed to 21%). 28% of students felt anxious taking assessments in either mode and 16% did not compare or rate anxiety levels.

Figure 98: Pie Chart to show student views on the comparative anxiety created by the paper and computer-based investigations



8.4 Expansion of the Basic Qualitative Findings

8.4.1 The Tests

Although the quantitative data indicated that student performance was slightly better on paper than computer, most of the qualitative data expressed a clear preference for the computer test versions.

Common themes from student questionnaires and interviews for preferring computer versions were that the computer-based test were '*better*', '*more fun*', '*more relaxed',* '*more enjoyable'* had '*pictures*' were '*active*', '*had colour*', and were '*quicker to finish*' with clear pictures and diagrams. These themes will be expanded in this section.

Nearly 60 per cent of students in the sample indicated a preference for doing the tests on computer and over a quarter didn't mind which mode they used to complete the tests. Reasons students gave for these preferences included the fact that the computer-based tests were '*more interesting and motivating',* that they did not have to '*write as much*', that they could easily amend or correct their answers and they were '*fun and active*'. Similar reasons were given in interviews, with students commenting that using the computer compensated for poor handwriting, and that it was easier to erase mistakes and errors.

The computer-based tests were also perceived as being less daunting and exerted less exam stress. One student commented that *'it didn't seem like a real test'* and *'it seemed a lot more relaxed.'*

Just under a fifth (19%) of the students said that they preferred to do the test on paper. There were a range of reasons given, however logistical concerns far outweighed those related to the assessments themselves.

The most significant issue for students when expressing a preference for paper–based assessments was the download time between questions on the computer versions; '*the computer test took time to load, whereas in the paper test it was much more straightforward and I could just get on with it', 'the questions took ages to load', 'you don't have to wait for the next page to load on paper as it is already there.'*

A number of students preferring paper assessments were concerned that tests using computers and the internet were not reliable; *'things can go wrong on computers and you can lose your work', 'Because I was worried if I would do something wrong and would have to start over again', 'in case the computer crashed', 'on paper I know nothing can go wrong.'*

Familiarity, or the lack of it was mentioned by a few students, not in terms of being completely unfamiliar with computers, but rather the comparison of being presented with assessments in different modes; '*I've being doing tests on paper all my life', 'because it was a new experience and I didn't know what to expect', 'exams have always been done on paper, we should maintain this tradition.'*

There were a small number of students who expressed anxiety about ICT generally; '*laptops don't like me, they don't ever work for me' 'I find computers stressful.'* However, for more students, a surprising concern was one of cheating and malpractice '*we can't cheat on paper'*, '*it's easier to cheat on the computer*', *you have the possibility to research on the internet and answer the question' 'there's less chance of some one looking at your answer on paper.'*

Accessibility issues did not feature very much in either student or teacher responses. A few students raised concerns such as '*the screen tires my eyes and makes concentration difficult' and 'the computer test made me feel tired and gave me a headache.'* Teachers accessibility issues were more concerned with possible poor hand-eye co-ordination problems some students might have using a mouse to answer questions.

In terms of assessment related issues, generally, there were few comments from students or teachers commenting on why paper-based tests were better or more preferred. A few students referred to perceived frailties in computer marking systems; '*because they (computers) have set right answers but with paper you can pick up marks or half marks in some questions', 'if you use a computer you are just another number.'* A few students referred to working practices, *'on the paper you are able to make notes to help your understanding', 'because on paper you can write things down to remember', ' it was much harder to do calculations to work out answers on the computer', ' traditional method of pen and paper is much easier.'*

Remaining concerns of a small number of students included the comparison of question layout between paper and computer, '*you could easily see the question in front of you (on paper), and flick between pages', 'you can see two pages at a time on paper but only one on the computer',* and the perceived lack of writing space on screen, ' *you can write as much as you want on paper', 'I didn't like the word limit on the computer-based test.'* There was not a word limit for the assessments on-screen.

As evident in the qualitative data, the computer based assessments were preferred by more students in every comparative aspect. In terms of the functionality of the computer assessments, the majority of students both in questionnaires and interviews remarked that they found the interactivity of the 'drag and drop', drop down lists and 'tick boxes' very straightforward and easy to use. Main reasons given by students on their preference for using computer mechanisms rather than paper-based alternatives were that they saved time and effort, '*It was simpler just to click and move things than to waste time writing it all out', 'It was just easier to look at and process', 'because all you have to do is click on things and you don't have to write as much', 'I am a lot quicker with computer than writing it down.'* For research purposes, two different mechanisms were used in the drag and drop questions. In some questions, once options were dragged from an available list, they were removed as options, whereas in other questions, options were re-useable. Students did not expect that the computer tests would provide this re-useable option, and a few students were confused by this facility. Interestingly, this option is common practice in paper-based questions, and was an option on the paper versions, but was not commented as being an issue by students.

The video clips were popular with students. Common views were that the clips allowed them to see clearly what was happening and acted as an aid to remembering actual experiments in contrast to diagrammatic forms used in paper versions. Comments such as they helped '*to understand what was going on'* and '*you can actually see how things happen'* were common.

Students also perceived the coloured pictures on the computer versions to be useful. The paper versions used standard black and white photos. The colour was felt by the majority to provide clarity when answering observational questions. Most students did not mention colour in relation to any assessment advantage, but rather that it enhanced the appearance and engagement of the assessments, '*the pictures were more colourful', 'the* *colour diagrams were clearer', 'I concentrate more with pictures', 'because some of the pictures actually moved and you understand it more.'*

Some questions involved the use of information boxes. These were available through an icon on the screen which then opened up an informational table or diagram, which could then be moved around the screen or minimised as required. In the paper versions, all required information was given in the form of tables and charts, and questions were generally presented as double paged spreads. This was not possible on computer as screen space was limited. Generally, students were comfortable using these on-screen information boxes and found them helpful and straightforward to use. They also liked that they could be moved around the screen. A few students commented that they '*did get in the way a bit of the questions*', but this was a minority view.

Teachers were more concerned with the use of the on-screen information boxes. They acknowledged that they were useful, but had concerns that there was a need to continually move them around the screen and switch between windows. One teacher commenting '*I think [students] might give up quite quickly and just almost guess the answer if they had to open it up and close it down a couple of times.'* Another felt this type of question was easier to answer on paper '*paper copy is better because you've got all the information in front of you; you don't even have to change pages … questions are printed: you have two pages facing pages with one question, so when the paper is opened, you get all the information in front of you, whereas with the pop-ups, it wasn'*t.'

A main question put to students and teachers concerned the use of open responses in different modes. As mentioned before, paper-based open responses is the default mechanism used by students in exams in this country, therefore students taking part in this trial did not express strong opinions on the open response sections of the paper assessments. Their views were based on the differences that the computer based tests presented.

A strong theme emerging from the student questionnaires and interviews was that computer-based open responses reduced the writing, neatness and legibility burden in a high stakes assessment, and that this would be advantageous to students and increased accessibility; '*writing takes me longer'*, '*on paper you try to write neater so the examiner can read it but on computer it doesn't matter, it is always neat.'* Most students perceived this facility as much more favourable than written responses, '*on paper, your hand starts* *to hurt', 'I don't really like writing on paper because my hand writing is not very neat', 'I worry that they won't be able to read my writing or see scribbling out.'* Allied to this was the opportunity to review and amend their responses; '*if you get it wrong you can rub it out', 'I could change my answer and go back and edit what I put.'*

A few students did comment that answering the open response questions on-screen could slow down students who had limited or slow typing skills

In terms of available space on-screen to respond, most students felt that there was enough space to type in their answers. The mechanism employed on-screen was that a response box was provided with an expected length of response in mind, however if a student typed a longer response, this would be accommodated. Some students mentioned that there was not enough space in the answer boxes to type in their answers and they were worried that they would '*run out of room.*' One student commented that when typing answers she kept them short whereas if she was writing on paper she would have '*rambled on*'. Conversely another student said that he wrote more on the computer-based tests remarking that it felt like doing less work than if he was answering it on paper.

Some students assumed that the size of the answer box indicated the amount that should be written, and they adjusted their responses accordingly. This reflects current practice on paper-based assessments. Several students felt that typing answers allowed them to express their understanding better. They also liked that they could easily change answers without having to rub or cross out. Students commented that they would have liked an opportunity to '*have a digital notepad or something … which would help'* as it would have provided them with the opportunity to make notes before constructing their response. These students felt that this opportunity was available on paper but not on-screen.

In interviews, teachers expressed few views on the logistical issues of writing on-screen compared with paper. However, they were strongly in favour of open-responses being part of on-screen assessment, rather than only comprising of fixed closed objective questions. This view was also evident in student questionnaires and interviews. There was a commonly held view that writing open responses to some questions allowed students to show their knowledge and understanding more than they could do with only fixed responses.

In student questionnaires and interviews, the predominant view was that that the computer-based test was straightforward to use and that you did **not** have to be '*good at using computers to do well in the test*' and if you knew the basics that would be adequate. It was remarked that the test was very well laid out and that the instructions were very clear. One student said that he was '*useless on computers*' but still '*found it easy.*' Interestingly a view taken by some of the students was that there are very few people now who do not have the basic computer skills required to take these assessments.

Teachers were also positive about the tests commenting on their clear and attractive presentation, layout and interactivity. In general, they liked the functionality and interactivity of the tests, the ease of navigation and the fact that the tests did not require any specialist computer skill to complete.

8.4.2 Authenticity and fitness for purpose

Many students thought that both the computer-based tests and paper tests were a fair way to assess their science knowledge and understanding. They felt that the both tests were made up of similar sort of questions that the '*questions tested your knowledge not the format.*'

Teachers liked the use of colour and felt it made the computer-based tests more visually appealing than paper-based tests and, therefore, possibly engaged students for longer than the paper-based tests. One teacher saying '*I loved the computer one, I really did, because when I saw the first one where they had the skeleton one and they had to say which organs does the ribs protect? That's really good because they've got a live model, everything's labelled, it's clear, it's big, and it's lovely to see. You've got the visualising …* '.Teachers particularly liked the inclusion of videos in the tests as opposed to just pictures or diagrams. A teacher commenting that he '*thought it was really good to see the experiment like that ….* [it] … *was great because you could actually see it happening, rather than just a picture of something which children don't relate to.*'

Teachers liked the mix of questions in the test particularly that they were not all multiple-choice. They liked the incorporation of free responses questions into the test. One teacher stated that the style of questions got students thinking.

Teachers felt that the test was assessing their science and knowledge as they felt that the students only needed basic computer skills to do the test. One teacher commented that she *'thought it was good; easy to use and it wasn't testing your computer skills, which is what I was worried about initially.'*

8.4.3 The Investigations

The questionnaires and the interviews with students and teachers asked some similar questions about the investigations as those for the tests, but as the investigations involved completely different response mechanisms and assessed process rather than content skills, the qualitative evidence included more discussion about authenticity and fitness for purpose than the tests. This emphasis is reflected in the following section.

In the questionnaires, three quarters of the students said that they found the computer-based investigations easy or very easy. Reasons they gave for finding it easy included that it was '*easier to understand', 'easier to use', 'interactive*' and they found using the '*simulator*' helped them. Some students did indicate that they had difficulty collecting data from the simulator.

Over half of the students said that they preferred the computer-based investigations and the majority of students in interviews expressed a comparative preference for doing the investigations on computer in questions. Reasons students gave for preferring the computer-based investigation during interviews reflected those made in questionnaires. The reasons being that they were '*fun*', '*easier to do', 'less time consuming'* and '*gave you more things to do rather than just reading about it, It's like you've got more help … you could see really clear what you had to do*.'

In interviews, students commonly referred to the increased motivation when taking the computer version, 'I was *more switched-on all the time, cos when you're writing over and over again, it gets boring and tiresome, but when you do this then it kind of keeps you switched-on all the time.*'

Of those students who preferred the paper-based investigation, the majority of reasons were logistical in nature, reflecting the same issues raised in the computer-based tests; that students had to wait for next screens, worries about the test crashing and the loss of data '*there is no chance of the paper version not working'* and some lack of confidence in the marking facility of the computer. A few students expressed the view that the computer programme was a less rigorous assessment than the paper-based version. The difficulty of manipulating the simulator, collecting results and particularly plotting a graph and then drawing a line of best fit were mentioned by students preferring paper-based versions '*it's easier to do things like line of best fit on pape*r'. Some students also expressed concern about writing open responses on screen '*paper lets you explain your answer more fully.'*

Students had mixed opinions on how fair they thought the computer-based investigations were at assessing their investigative skills compared to the paper-based versions. Just over a third of the students thought that the computer-based test was a fair way to test their process skills with just over a fifth thinking the paper-investigation was fair. 30 per cent thought both investigations were fair.

Students who thought the computer-based investigations were fair liked the opportunity to collect their own data, '*you can virtually carry out a mock investigation instead of reading and analysing someone else's results*' was a common response. Students enjoyed the interactivity of the computer-based investigation and the use of images and videos and felt their inclusion made them '*more interesting'* and '*helped their understanding.*'

Students who thought that the paper-based investigations were a fairer way to test their investigative skills gave reasons ranging from the fact that they felt that it '*tested them more*', were '*more challenging*' and '*because you can write more stuff.'*

Key comments made by students who said they thought that both modes of investigations were both fair in testing their science process skills included that '*there was no difference in the style and standard of questions*' *and* that they both '*made you answer questions about science investigations'*, were equally '*challenging'* and were just '*different formats.'*

In interviews the majority of students felt that the computer-based investigations were a fair way to assess their science process skills. One student remarked that the investigation '*did go through a wide range of everything you need to kno*w'. Another preferred doing the computer-based investigation over a practical-based investigation, criticising the latter because he thought that as they are often done in a group '*some people just let one person collect the data, another person does something else*' and as a result he felt that as a student he is '*not really getting assessed in all of them and in fact is only really only getting assessed on one cos you're the only one who's done that one thing.*' Another student said that he thought it was fairer than a paper-based test as '*you were collecting your own data.'*

In terms of the functionality of the investigations, and comparative differences between paper and on-screen versions, questions to students and teachers were divided up into stages of the investigation, from data collection, to graph completion and then open response analysis of data.

8.4.4 Data Collection

This part of the investigation was significantly different in paper and computer modes. The paper version followed standard practice found in summative assessments; an investigation was described, and then data from the investigation was given to students. The computer-based investigations presented students with a simulated environment, where they had to select and collect data that they would then use.

The majority of the students in questionnaires and in interview found the computer simulators in the investigation easy to use. They found the instructions of how to use them simple and straightforward to follow *'they were laid out really nicely'* and *'labelled really well'* were common comments.

One student remarked on how the simulator helped her to visualise what was happening and therefore helped in making the investigation easier, describing it as *'much easier for me than to just imagine it.'*

Students also commented on how they liked the opportunity to collect their own range of data using the simulators and they felt it was more like doing a '*proper investigation'* than when being done on paper. One student remarked that *'doing it yourself … gets you more involved in it and makes you think more.'* Another student said that *'the ability to adjust the temperature so'* she could *'record'* her *'own results was something'* that she *'obviously couldn't do on a paper exam.'*

Teachers felt that the simulations were a very good way of allowing students to collect data themselves as it gave them the opportunity to think about what data they were collecting. They also felt that it could provide students with a good set of results to use to plot graphs and analyse results. Teachers' comments also included the fact that they thought it *'was good for them to collect data because they needed to work out when the best time was to collect the data'* and that the investigations were *'great because you had to think about what results to take, how to put them in the table.'*

8.4.5 Graph Completion

Some students in the interviews did say that they found it '*tricky* 'to plot the points accurately onscreen, and found the gridlines difficult to see giving reasons such as, the colours used were '*quite pale', 'crosses were a bit big'* and '*the lines are really small'*. Student who used the touchpad on laptops and not a computer mouse experienced difficulties manoeuvring the cross into the correct position on the onscreen graph.

Students also expressed that they would have liked the facility to go back to change their results. '*it didn't tell you how big the graphs were going to be, so I didn't get a wide enough range of results so when I came to my graph they were all cramped in one corner.'*

Teachers felt that their students found labelling of the graphs straightforward. One teacher commented that she '*particularly liked having to label the axes',* and *'having the drag and drop'* mechanism to plot the points.

Teachers considered the plotting of points using the drag and drop mechanism to be straightforward for students to complete. One teacher pointed out that she liked the way '*you dragged those over and plotted the points'.* However, another teacher did not like the plotting mechanism and found it '*incredibly frustrating'* and *'didn't like …at all.'*

8.4.6 Lines of Best fit

The majority of students interviewed described creating the 'curve of best fit onscreen for the 'photosynthesis' investigation *'confusing', 'difficult' 'unclear' and 'tricky'*. One student described it a 'a bit weird to use'. Many were not clear about the functionality of how to draw the curve and one student said he 'took ages' and in fact he '*didn't quite get to finish it'.*
Some students did use the demo button and found it useful. Others didn't, finding it of '*no help*' and '*confusing*'. One student commented that when he clicked the demo he 'was

wondering what was going on'. Another student remarked that if they had practice at creating the curve before the assessment it would have helped.

A couple of students commented that they would have rather done the graph on paper than on the computer. Another student mentioned that given the opportunity she would like to have gone back and corrected her graph. Another thought that '*it was a good way of backing up your skills by using an interactive method*, *but I think graphs should really* be left to hand'. Generally, students found the creation of the line of best fit in the 'Forces' investigation straightforward.

Teachers views reflected those of their students, particularly for the drawing of a curve of best fit in the *photosynthesis* investigation. They felt that manipulating the curve would be challenging for students and that they would need practise to familiarise themselves with its functionality. A couple of teachers raised accessibility concerns, for example students with dyspraxia would find it difficult and frustrating trying to manipulate the points to draw the graph and to draw a line of best fit. In such cases it was felt that the computer-based investigation would not be assessing their science process skills rather than computer and fine motor skills.

Teachers pointed out the problems that students traditionally have in science when drawing graphs on paper. They suggested that the computer investigation could be used to improve students' graph drawing skills. Teachers saw advantages to creating the graph pointing out its use as a teaching tool. They described how they could use the program to '*carry out an investigation quite instantly and then plot a graph'* showing students *'how to use a graph'.* Many teachers mentioned that '*not knowing how to plot graphs'* for *a lot of students up until Year 11'* is *'a really weak area.'* One teacher noted that it was still very important for students to be able to plot and draw graphs on paper.

8.4.7 Analysis of data

The students liked the opportunity to type in their own answers in the free response boxes. However, they did have varying views on the space available to type their answer. Some felt that it limited the answers they could give and one commented that '*I was going to write something else but I had to take out something I'd already written to write what I was going to write afterwards.'* and '*on some of them you had to delete your answer and write again in shorter, make it shorter.*' Most students however felt that there was enough room.

Some students felt more comfortable with writing their responses on the paper-based test. One student said that it was easier to '*explain in your own words more and it was your own handwriting so you could cross out a mistake if you had it, but then if you're on the computer, you kind of just feel stuck because you don't know what to write, and your eyes start aching.'*

Teachers felt that there should be more space on the computer versions for students to write their free response answers, '*they tend to waffle a bit and I think they fill up the room quite quickly.'*

8.4.8 Authenticity and Fitness for Purpose

Several of the students interviewed thought that the computer-based investigation should not be exclusively used as a summative method of assessment, however they could be used alongside a practical assessed investigation ensuring that students '*kept doing the practical experiments*'.

One student commented that she felt that continuing to do class-based practicals was important as doing the practical work '*teaches you what you need to improve for future*' and could help in obtaining a '*better understanding of the results that you get*'. Another student said he '*quite like doing the experiments*' and did not want to end up doing them '*just on the computer*'. It was also mentioned by a student that the computer-based investigation didn't '*show how used you are to actually taking part in experiments and setting them up and completing them.'*

The final comparative questionnaire indicated that nearly 60 per cent of the students enjoyed doing the computer based investigation contrasting with just over 10 per cent enjoying the paper-based test. Students preferring the computer version referred to their familiarity and confidence with using computers. They perceived it as being '*easier*' and liked the incorporation '*of colour and video.*' Students who enjoyed the paper-based investigation referred to their familiarity with doing this type of assessment on paper. Some students complained about getting a headache when using the computer or feared that the computer could break down and lose their work.

Both interviews and questionnaires indicated that students liked the interactivity of the investigation and that there was no time wasted '*setting up all the equipment*'. Students in interviews thought that it was assessing their scientific skills '*I think the investigation was really good, cos it's like you're doing an experiment*', '*more hands-on*' and that '*you have to really know what's going on.*' One student described how he found it a lot easier to do because of the fact that you don't have to set the experiment up. This then meant that he knew he could collect data, whereas if he had problems setting up a practical, he would not be able to show his data handling and analysis skills.

Students felt that it could be used as a learning tool. One student suggested that '*it would like teach you how to do the investigation and when you actually had to do proper ones it would work better because you'd done the same before'.*

All teachers interviewed viewed the investigations as a useful, complementary and '*good additional tool'* to practical work but not as a replacement. It was seen as a good alternative method for students to create sets of '*decent results'* for analysis, when for example, they sometimes have problems collecting data good enough to analyse, either because they have problems with '*equipment not working*', *'constrained with time or with amounts of equipment that you have.'*

One teacher pointed out that she saw it as an aid to teaching and learning and said that she thought '*it would definitely help in the classroom and complement practical work because I think it would give the students some idea of what should happen in a practical as opposed to what often doesn't happen in a practical and would also help with practicals that you can't necessarily do in the classroom.'* Another teacher suggested that although the '*Forces*' investigation would not be easy to set up as a practical in the lab she could see herself using the computer program as a '*demonstration'.*

The use of the interactive whiteboard to go through the computer investigation with a whole class to facilitate the use of the graph plotting and drawing functions was highlighted by some teachers as a possible teaching approach. One teacher liked the fact that the investigation was '*already there prepared'* and believed she would find this very useful in her teaching and considered there to be a '*lot of flexibility*' for using the '*investigation in teaching investigation skills.*'

The investigations were also viewed as useful tools for '*doing up tables, plotting graphs etc'.* Teachers in particular liked the opportunity it gave students to collect and manipulate their own data.

Teachers thought that the investigation was an authentic tool to assess students' process skills. A teacher considered the computer-based investigation to be a very good way of assessing because '*it's a sort of real experiment*' and she felt that students '*could relate to it more than just a paper test.'*

Teachers also liked the interactivity of the investigations and believed the computer-based investigations were an '*authentic'* approach for assessing students' data handling and analysing skills and that '*the investigations on the computer were better than the investigations on paper because they had to collect their own data'.* The opportunity for students to collect their own data was regarded to be a good function of the investigations, one teacher pointing out that when students are faced with a table of data '*it doesn't mean anything to them, but because they actually collected the data, I think it's more meaningful for them on the computer-based test.'* She also highlighted the importance of students being able to understand that the use of simulations and modelling and predicting, as this is '*another skill that scientists use in the real world*'. This was echoed by another teacher who thought it would be useful '*for the children to see that in some cases simulations are used outside school to extract data.'*

Teachers saw a place for the investigation in supporting teaching and engaging students in the learning of investigation skills, particularly for collecting, handling and analysing data. Teachers remarked that the investigation was good at getting students to think about the experiment in terms of *'what they're doing and why',* allowing them to make decisions as they '*needed to work out when the best time was to collect the data and they could always change their mind and then pick some better data, especially with drawing the graphs.'* One teacher thought that the students would be more likely to do better in the computer-based investigations as '*they're more likely to remove data and put data back …. and redraw their graphs.'* The inclusion of anomalous data in the software was a feature which was popular with teachers.

One teacher thought that the computer investigation '*was really good because it was … active'* and if there were no technical problems and the programme worked correctly it would be '*fairly easy to implement*'. She observed that '*it was quite a quick and easy way of testing their skills …. you could immediately see that some of the pupils latched on* *straightaway, knew what to do, knew how to collect the data, whereas others were really struggling with it.*' Consequently she thought from '*that point of view it was a really useful tool.'* The same teacher also thought that the interactive nature of the computer-based investigation as a learning tool was better than the paper-based investigation as she thought '*it would appeal to a wider range of abilities*' and be more engaging for the students.

Some teachers acknowledged that individuals learn in different ways and the computer-based investigations would be a chance to enhance learning opportunities and vary activities offered to students. They therefore felt it was useful in the context of personalised learning. A couple of teachers felt that the computer-based investigations could be valuable exercises as alternatives to practicals as they could provide a more controlled focus. Sometimes, when doing practicals, students '*get very excited ... they're focussing on being up and about rather than thinking about the science and what they're actually doing.'* Simulated investigations could also be useful to overcome health and safety issues for particular practicals. One teacher pointed out that she could see that the investigations could be used in the assessment of coursework.

However some teachers did identify limits to the computer-based investigations. These teachers indicated that there would still be a need for students to plan their own investigations, manipulate real apparatus and collect and analyse first hand data. They felt that these process skills should still have place in the assessment of science process skills.

8.5 Summary

This chapter has shown the outcomes of the questionnaires and interviews from students and teachers. Their views on the appropriateness of the science tests and investigations in different modes have been described in a basic and expanded manner. Having now presented the quantitative and qualitative data and evidence, the next chapter analyses both strands of evidence and explores the emergent issues.

Chapter 9

Analysis

9.1 Introduction

The two previous chapters provided the quantitative and qualitative evidence from trialling comparative on-screen and paper-based science tests and investigations with approximately 1000 students. This chapter will synthesise the differing evidence strands and explore any emergent key issues.

In terms of student performance, there were significant differences between students taking tests and investigations in paper-based and on-screen modes. Although there were some significant differences within items, there were no significant differences between genders across the tests and investigations and therefore this variable will not be explored any further, as there was no empirical or qualitative evidence to suggest it might be an issue when considering the assessment of students in different modes.

9.2 Group sample differences

Before going into detail in the analysis of the test and investigation performance across different assessment modes, it is necessary to clarify how differences in the sample and the tests affected the evidence outcomes.

Ideally, in this type of study, the student group samples 1 and 2, and the test and investigations A and B should be as equivalent as possible in order to carry out a comparative study. The evidence obtained indicates that there were active differences, and they therefore need to be factored into any analysis.

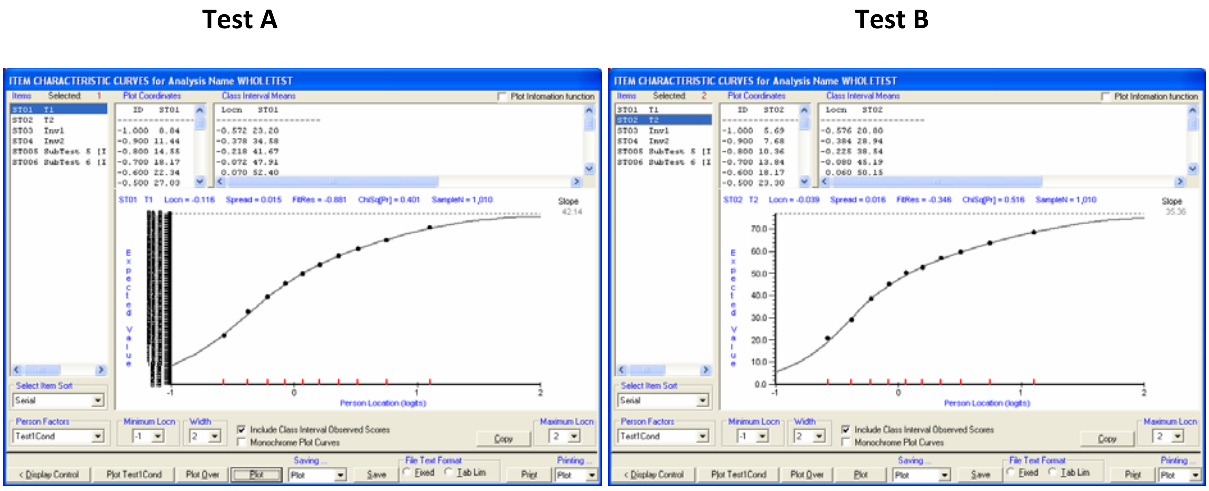
Table 5 on page 114 showed the imbalance of student ability across Group 1 and Group 2. Although teacher assessment levels of students operating at Levels 5 and above were nearly identical for both groups, the number of students operating at levels 6 and 7 was considerably higher in Group 1.

This was in part due to a school effect. There were difficulties in obtaining a large sample of students to take part in this trial, for students to complete the paper and on-screen versions, and for the schools to submit any prior attainment data, including teacher assessment levels. Therefore attributions to groups were taken on the basis of teacher advice on the range of ability of their students, without necessarily any prior attainment data to hand. One particular school, which provided a large number of students should have provided a nationally representative sample, and accordingly were all put into Group 2. When prior attainment data was obtained (after the students had taken the tests), the profile of students at this school indicated a slightly less able cohort than anticipated, This one school effect alone accounted for the ability imbalance between groups 1 and 2.

9.3 Test and investigation Performances

The two tests and investigations were designed and developed to have the same levels of demand. This was done on the basis of levelling questions using level descriptors on the National Curriculum scale, and then putting together equivalent tests and investigations. This process is described in Chapter 6. This exercise was undertaken by experts in the field, experienced in the development and construction of national curriculum science tests. The fact that these equivalent tests and investigations had differences in difficulty bears witness to the power of pre-testing. When national curriculum tests are constructed, all items are pre-tested first before they are put into test versions and then pre-tested again as complete tests. This means that levels of demand and difficulty are not made on expert judgement alone; they are supported by empirical data. It is interesting to note that the two equivalent tests and investigations produced for this comparative study, although the levels of demand were probably correct, had unexpected levels of difficulty.

Figure 99: Rasch Item Characteristic Curves for Tests A and B



Test A

ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

Items Selected: 1

```
ST01  T1
ST02  T2
ST03  Inv1
ST04  Inv2
ST005 SubTest 5 [I
ST006 SubTest 6 [I
```

Plot Coordinates

```
  ID    ST01
---------------
-1.000  8.84
-0.900 11.44
-0.800 14.55
-0.700 18.17
-0.600 22.34
-0.500 27.03
```

Class Interval Means

```
Locn   ST01
--------------------
-0.572 23.20
-0.378 34.58
-0.218 41.67
-0.072 47.91
 0.070 52.40
```

ST01 T1 Locn = -0.116 Spread = 0.015 FitRes = -0.881 ChiSq[Pr] = 0.401 SampleN = 1,010

Slope 42.14

Select Item Sort: Serial

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

☐ Monochrome Plot Curves

Maximum Locn: 2

Saving ...: Plot

File Text Format: ○ Fixed ○ Tab Lim

Printing ...: Plot

Test B

ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

Items Selected: 2

```
ST01  T1
ST02  T2
ST03  Inv1
ST04  Inv2
ST005 SubTest 5 [I
ST006 SubTest 6 [I
```

Plot Coordinates

```
  ID    ST02
---------------
-1.000  5.69
-0.900  7.68
-0.800 10.36
-0.700 13.84
-0.600 18.17
-0.500 23.30
```

Class Interval Means

```
Locn   ST02
--------------------
-0.576 20.00
-0.384 28.94
-0.225 38.54
-0.080 45.19
 0.060 50.15
```

ST02 T2 Locn = -0.039 Spread = 0.016 FitRes = -0.346 ChiSq[Pr] = 0.516 SampleN = 1,010

Slope 35.36

Select Item Sort: Serial

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

☐ Monochrome Plot Curves

Maximum Locn: 2

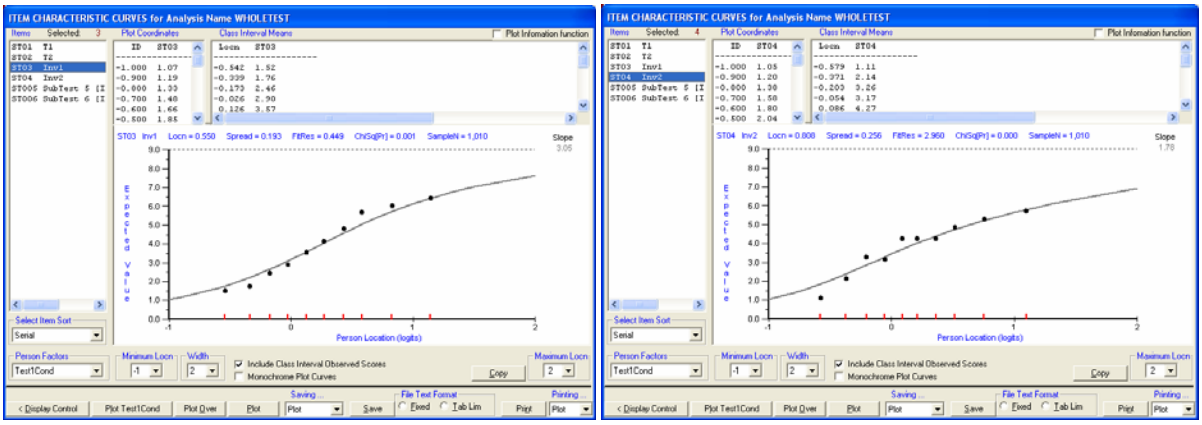
Saving ...: Plot

File Text Format: ○ Fixed ○ Tab Lim

Printing ...: Plot

The two Rasch characteristic curves shown above in Figure 99 show that although the general pattern of performance was similar for Tests A and B, the performance levels in Test B were lower than in Test A, the performance (axis Y) across the ability range (axis x) in Test B being consistently lower than that of Test A

Figure 100: Rasch Item Characteristic Curves for Investigations A and B



Investigation A

ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

Items Selected: 3

```
ST01   T1
ST02   T2
ST03   Inv1
ST04   Inv2
ST005  SubTest 5 [I
ST006  SubTest 6 [I
```

Plot Coordinates

```
   ID     ST03
---------------
-1.000   1.07
-0.900   1.19
-0.800   1.33
-0.700   1.48
-0.600   1.66
-0.500   1.85
```

Class Interval Means

```
 Locn    ST03
--------------------
-0.542   1.52
-0.339   1.76
-0.173   2.46
-0.026   2.90
 0.126   3.57
```

ST03 Inv1 Locn = 0.550 Spread = 0.193 FitRes = 0.449 ChiSq[Pr] = 0.001 SampleN = 1,010

Slope 3.05

Select Item Sort: Serial

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

□ Monochrome Plot Curves

Maximum Locn: 2

Saving ...: Plot

File Text Format: ○ Fixed ○ Tab Lim

Printing ...: Plot

Investigation B

ITEM CHARACTERISTIC CURVES for Analysis Name WHOLETEST

Items Selected: 4

```
ST01   T1
ST02   T2
ST03   Inv1
ST04   Inv2
ST005  SubTest 5 [I
ST006  SubTest 6 [I
```

Plot Coordinates

```
   ID     ST04
---------------
-1.000   1.05
-0.900   1.20
-0.800   1.38
-0.700   1.58
-0.600   1.80
-0.500   2.04
```

Class Interval Means

```
 Locn    ST04
--------------------
-0.579   1.11
-0.371   2.14
-0.203   3.26
-0.054   3.17
 0.086   4.27
```

ST04 Inv2 Locn = 0.808 Spread = 0.256 FitRes = 2.960 ChiSq[Pr] = 0.000 SampleN = 1,010

Slope 1.78

Select Item Sort: Serial

Person Factors: Test1Cond

Minimum Locn: -1

Width: 2

☑ Include Class Interval Observed Scores

□ Monochrome Plot Curves

Maximum Locn: 2

Saving ...: Plot

File Text Format: ○ Fixed ○ Tab Lim

Printing ...: Plot

The two Rasch characteristic curves shown above in Figure 100 for the investigations demonstrated a similar pattern of performance to that of the tests, Investigation A performing consistently higher across the ability range than Investigation B.

Difference of performance across the tests and investigations were in part due to differences in the ability profile of the two groups. However, differences in performance on seemingly equivalent assessments demonstrate one of the aspects of reliability discussed in Chapter 2. Assumptions are often made that tests or examinations are equivalent across series or years on the basis of an understanding of question or item demand. However, any two questions, even if they are based on the same concept or construct will often have differing levels of difficulty (actual performance and facility values) to the students that take them.

The context, style, layout and language of questions can all contribute to varying performance. The only way to eliminate this variable is to either use exactly the same questions every time or to pre-test all items. Public examinations in England do not use either of these strategies. They consist of equivalent papers using levels of demand as the predictive indicator, and then threshold cut scores are established on the basis of actual performance (ie the level of difficulty of the exam across the cohort). This therefore means that score thresholds can change across exam series, although the percentage success rate might remain constant.

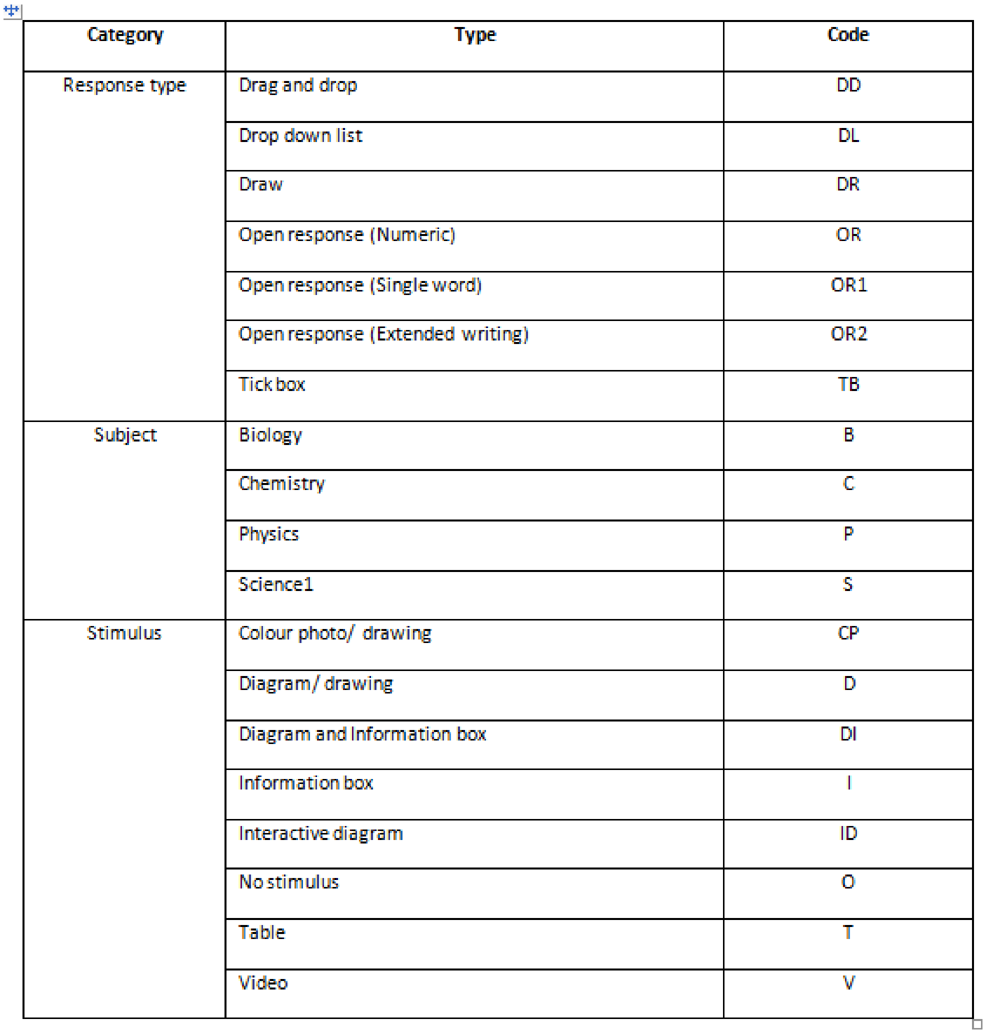
The imbalance between the ability of the group samples and the difficulty of the tests and investigations were difficult variables to control and led to seemingly contradictory results; Group 1 scoring significantly better on the paper assessments than the on-screen versions whereas Group 2 scored slightly better on the computer versions than the paper-based versions. As described in Chapter 7, factoring out differences in group ability and test difficulty resulted in an overall empirical outcome indicating a 2 mark overall difference in test performance in favour of paper in this comparative study. Due to the small number of items within the investigations and the fact that the paper and on-screen versions consisted of different mark totals, empirical comparative analysis of the investigations was not carried out. Analysis of the investigations will take place after the analysis of test performance.

In contrast to the quantitative outcomes, the qualitative evidence in Chapter 8 indicated a clear preference for the computer versions in every comparative category. This chapter will focus on the categories of items that demonstrated significant differences in performance, the nature of the differences between the items in different modes and possible reasons for the performance differences. This analysis will be considered alongside the qualitative views of students and teachers.

Chapter 7 set out all the relevant empirical evidence from the trials, and established a difference in performance favouring paper-based assessments. Also in Chapter 7, items within the tests that demonstrated highly significant differences in performance compared to general trends in each test version were identified. These data sets are called Differential Item Functions (DIFs) and indicate at any set significance value where comparative items perform with a marked difference to other items across the test.

All items within the tests were coded according to three particular categories; science subject areas (biology, chemistry, physics, how science works), the stimulus presented in the item and the response type required by the student. This coding list is shown in Table 63 below.

Table 63: Item Codes for Response, Subject and Stimulus Types



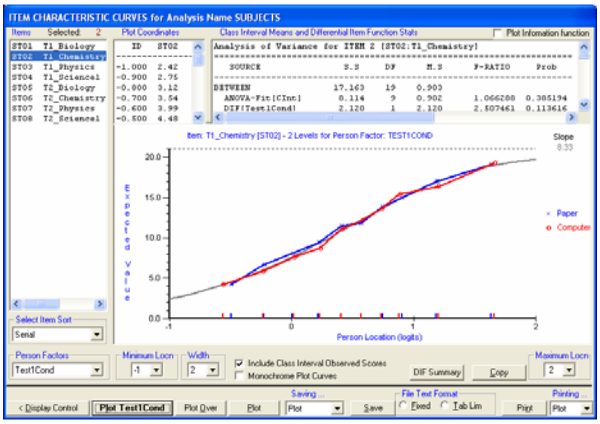
Category	Type	Code
Response type	Drag and drop	DD
	Drop down list	DL
	Draw	DR
	Open response (Numeric)	OR
	Open response (Single word)	OR1
	Open response (Extended writing)	OR2
	Tick box	TB
Subject	Biology	B
	Chemistry	C
	Physics	P
	Science1	S
Stimulus	Colour photo/ drawing	CP
	Diagram/ drawing	D
	Diagram and Information box	DI
	Information box	I
	Interactive diagram	ID
	No stimulus	O
	Table	T
	Video	V

9.4 Subject performance

When DIF items from Test A and B were analysed by type in order to see where DIFs occurred, it became clear that specific science subject area items were not active distractors. DIFs occurred across the areas of science. Rasch curves were produced on all the separate science areas across the tests to see if they showed any significant differences in performance between the paper and on-screen tests. They did not.

Figure 101 below shows the Rasch item characteristic curves between modes on all the chemistry items in Test A. Any differences in performance are shown not to be significant. This trend was replicated for all subject areas.

Figure 101: Rasch item characteristic curves for the chemistry items within Test A in paper and computer modes



```
ST01   T1_Biology
ST02   T1_Chemistry
ST03   T1_Physics
ST04   T1_Science1
ST05   T2_Biology
ST06   T2_Chemistry
ST07   T2_Physics
ST08   T2_Science1
```

```
   ID      ST02
-----------------
-1.000    2.42
-0.900    2.75
-0.800    3.12
-0.700    3.54
-0.600    3.99
-0.500    4.48
```

```
Analysis of Variance for ITEM 2 [ST02:T1_Chemistry]
===================================================================
  SOURCE              S.S      DF      M.S       F-RATIO     Prob
-------------------------------------------------------------------
BETWEEN             17.163     19     0.903
  ANOVA-Fit[CInt]    8.114      9     0.902     1.066288   0.385194
  DIF[Test1Cond]     2.120      1     2.120     2.507461   0.113616
```

While the science subject items were distributed evenly across both test versions, the stimulus and response types were used to best effect to assess the subject matter. The tests would not have been effective assessment instruments or fit for purpose if their design had been dictated by a set quota of stimulus or response types, however there was an active desire to use a variety of question and answer mechanisms for the purpose of the comparative study. Therefore, while there was a range of stimulus and response mechanisms used, there were differences in their usage across the tests. This issue will be discussed further in this section.

Analysis of the items that showed DIFs did indicate categories of stimuli or response types that resulted in significant differences in performance in differing modes. These categories will now be exemplified and discussed.

9.5 Stimulus Types

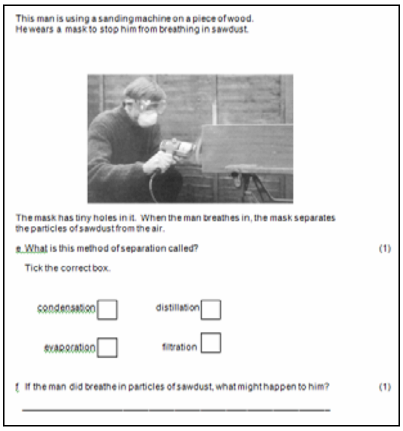
9.5.1 Use of Video

There were two items across the tests that used video as a stimulus. One did not result in any DIF performance, the other one did.

The first item is shown below in Figure 102. This item contained a still black and white photograph in the paper version, and a video sequence of the sander in action in the computer version. This item did not demonstrate any DIF in performance.

Figure 102: Screenshots of Test B, Q1 in paper and computer modes

Paper version **Computer version**



This man is using a sanding machine on a piece of wood.
He wears a mask to stop him from breathing in sawdust.

The mask has tiny holes in it. When the man breathes in, the mask separates the particles of sawdust from the air.

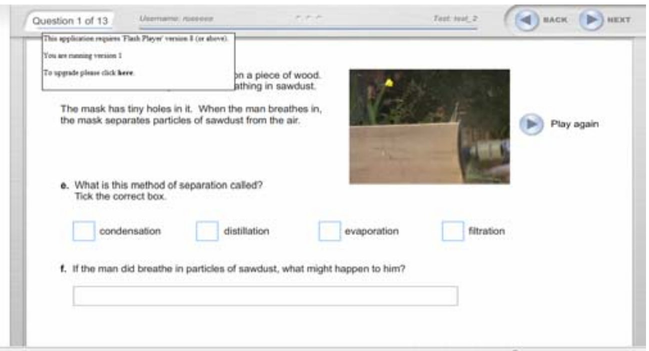
e What is this method of separation called? (1)

Tick the correct box.

condensation □ distillation □

evaporation □ filtration □

f If the man did breathe in particles of sawdust, what might happen to him? (1)



Question 1 of 13 Username: [illegible] Test: test_2

This application requires 'Flash Player' version 8 (or above).

You are running version 1

To upgrade please click **here**.

on a piece of wood.
athing in sawdust.

The mask has tiny holes in it. When the man breathes in, the mask separates particles of sawdust from the air.

e. What is this method of separation called?
Tick the correct box.

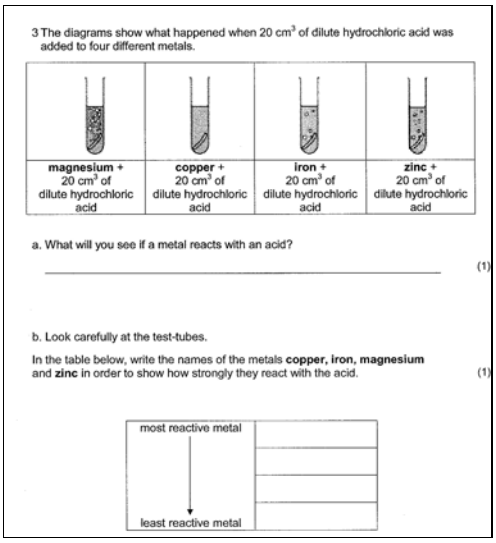
□ condensation □ distillation □ evaporation □ filtration

f. If the man did breathe in particles of sawdust, what might happen to him?

The second item is shown below in Figure 103. This item also used a black and white diagram in the paper version and a video sequence of the experiment in the computer version. This item did demonstrate a highly significant DIF in performance, favouring the paper version.

Figure 103: Screenshots of Test A, Q3b in paper and computer modes

Paper version **Computer version**



3 The diagrams show what happened when 20 cm^3 of dilute hydrochloric acid was added to four different metals.

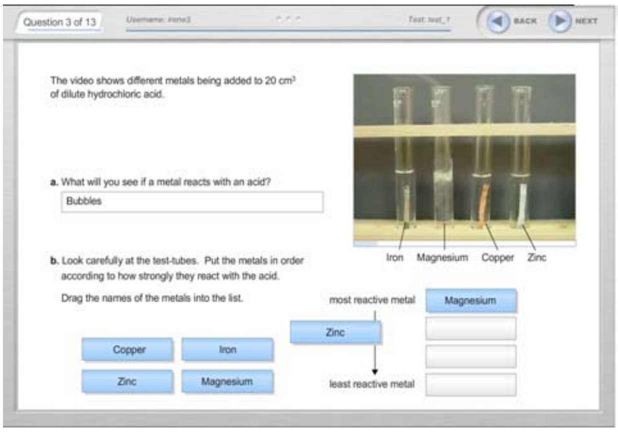
magnesium + 20 cm^3 of dilute hydrochloric acid	**copper** + 20 cm^3 of dilute hydrochloric acid	**iron** + 20 cm^3 of dilute hydrochloric acid	**zinc** + 20 cm^3 of dilute hydrochloric acid

a. What will you see if a metal reacts with an acid? (1)

b. Look carefully at the test-tubes.

In the table below, write the names of the metals **copper, iron, magnesium** and **zinc** in order to show how strongly they react with the acid. (1)

most reactive metal	
least reactive metal	



Question 3 of 13 Username: irene3 Test: test_1

The video shows different metals being added to 20 cm^3 of dilute hydrochloric acid.

a. What will you see if a metal reacts with an acid?

Bubbles

b. Look carefully at the test-tubes. Put the metals in order according to how strongly they react with the acid.

Drag the names of the metals into the list.

most reactive metal: Magnesium

least reactive metal

The two items shown above might appear to be using video for similar purposes, ie. as a stimulus for a question, however they are being used in very different ways and raise issues about the face and construct validity of paper and on-screen assessments.

In terms of face validity, students and teachers were overwhelmingly in favour of the use of video sequences in science questions. They felt their inclusion made the questions more authentic and engaging. They did not use the term construct validity, but it was clear from questionnaires and interviews that students and teachers thought that the use of video tested science with more validity than using 2D diagrams and photographs, as the representations were more authentic.

It was therefore interesting to compare the positive attitudes of students and teachers with the actual item performance and the manner in which video was used within the assessments.

In the first example shown above, the video sequence showing a sander in action was not an essential component of the item. It provided authenticity to the context and question, but did not in itself contain information required to answer the question. The second item shown above contains a video clip of an experiment in action which is an essential component of the question; the information required for the question obtained by careful observation of the sequence. This observational skill clearly assesses the construct of scientific enquiry with more validity than looking at a 2D diagram. It would be an expectation that students would have experienced science in this manner. It is a feature of an observation of this type that it does have a demand that the observation of the 2D diagram showing drawn bubbles does not have.

It is unsurprising, given the presentation and demand of the question in different modes that this item performed markedly better on paper than computer. Paper-based questions on this area of science have been stylised over time to present a version of an observational skill. However, the visual information is static and simple to decode in comparison to a dynamic but easier to miss image. The video sequence was repeatable to aid accessibility; however performance was still significantly better on paper. Although the demand of this question was low (Level 4), there were significant differences in performance across the ability range of students. Unfamiliarity of the medium in terms of being asked in an assessment to demonstrate observational skills may be in part responsible for lower performance on computer, however another possible reason is that because this sort of skill is not normally assessed in tests, it is not taught.

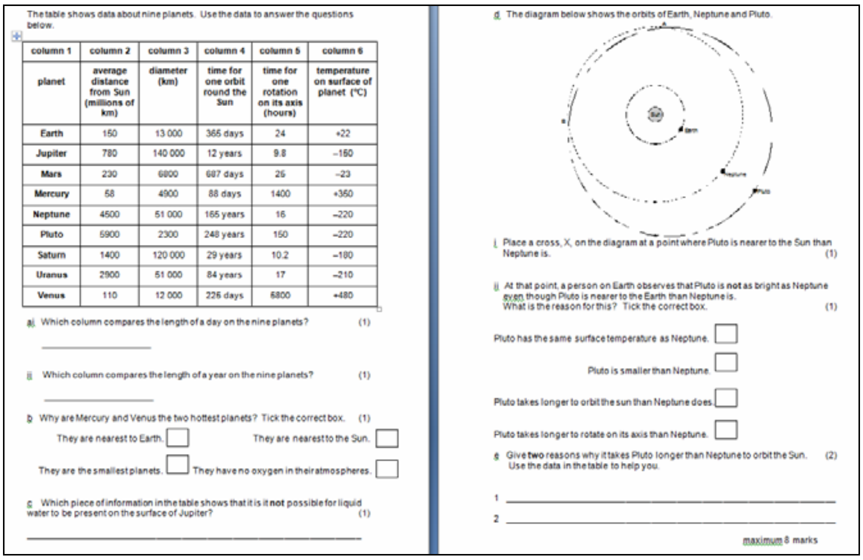
9.5.2 Information Boxes

One of the significant differences between presenting test questions on paper and on-screen is the amount of available space and the subsequent consequences on the style and layout of questions.

Questions that require interrogation of given data expose one of the challenges of presenting questions in differing modes. On paper, the default presentational style is to present questions alongside the required table of information. Occasionally this type of information can be presented in a tear-out sheet or a separate resource booklet. The key factor is the need to have the information to hand and available when answering the question. Presenting large detailed sets of data or information on-screen therefore presents a problem to a test developer where only one screen is visable at a time to the student. In this comparative study, data or diagrammatic information was made available to students using an information box mechanism. This enabled students to open up a table onto the screen, move it around and minimise it without having to flip to another screen away from the question. An example of this style of question is shown in Figure 104 below.

Figure 104: Screenshots of Test 1, Q13 in paper and computer modes

Paper Version



The table shows data about nine planets. Use the data to answer the questions below.

column 1	column 2	column 3	column 4	column 5	column 6
planet	average distance from Sun (millions of km)	diameter (km)	time for one orbit round the Sun	time for one rotation on its axis (hours)	temperature on surface of planet (°C)
Earth	150	13 000	365 days	24	+22
Jupiter	780	140 000	12 years	9.8	–160
Mars	230	6800	687 days	26	–23
Mercury	58	4900	88 days	1400	+350
Neptune	4500	51 000	165 years	16	–220
Pluto	5900	2300	248 years	150	–220
Saturn	1400	120 000	29 years	10.2	–180
Uranus	2900	51 000	84 years	17	–210
Venus	110	12 000	226 days	5800	+480

a) i Which column compares the length of a day on the nine planets? (1)

____

ii Which column compares the length of a year on the nine planets? (1)

____

b Why are Mercury and Venus the two hottest planets? Tick the correct box. (1)

They are nearest to Earth. □ They are nearest to the Sun. □

They are the smallest planets. □ They have no oxygen in their atmospheres. □

c Which piece of information in the table shows that it is it **not** possible for liquid water to be present on the surface of Jupiter? (1)

____

d The diagram below shows the orbits of Earth, Neptune and Pluto.

i Place a cross, X, on the diagram at a point where Pluto is nearer to the Sun than Neptune is. (1)

ii At that point, a person on Earth observes that Pluto is **not** as bright as Neptune even though Pluto is nearer to the Earth than Neptune is.
What is the reason for this? Tick the correct box. (1)

Pluto has the same surface temperature as Neptune. □

Pluto is smaller than Neptune. □

Pluto takes longer to orbit the sun than Neptune does. □

Pluto takes longer to rotate on its axis than Neptune. □

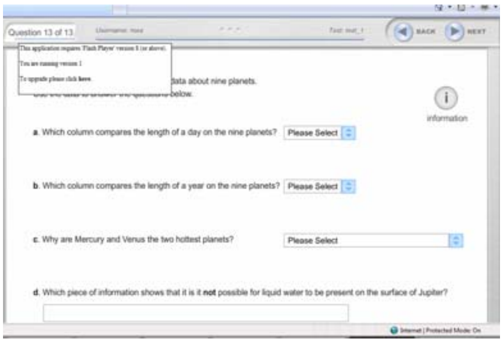
e Give **two** reasons why it takes Pluto longer than Neptune to orbit the Sun. (2)
Use the data in the table to help you.

1 ____

2 ____

maximum 8 marks

Computer version



This application requires 'Flash Player' version 8 (or above).

You are running version 1

To upgrade please click **here**.

data about nine planets.

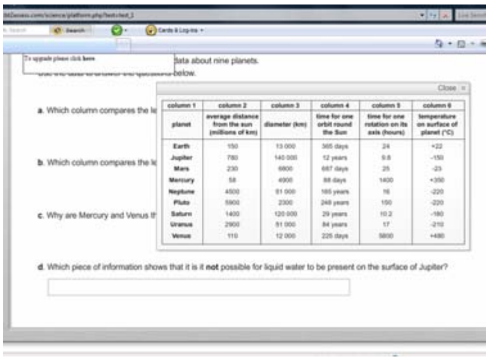
below.

a. Which column compares the length of a day on the nine planets? Please Select

b. Which column compares the length of a year on the nine planets? Please Select

c. Why are Mercury and Venus the two hottest planets? Please Select

d. Which piece of information shows that it is it **not** possible for liquid water to be present on the surface of Jupiter?



To upgrade please click **here**

data about nine planets.

below.

a. Which column compares the le

b. Which column compares the le

c. Why are Mercury and Venus th

column 1	column 2	column 3	column 4	column 5	column 6
planet	average distance from the sun (millions of km)	diameter (km)	time for one orbit round the Sun	time for one rotation on its axis (hours)	temperature on surface of planet (°C)
Earth	150	13 000	365 days	24	+22
Jupiter	780	140 000	12 years	9.8	-150
Mars	230	6800	687 days	25	-23
Mercury	58	4900	88 days	1400	+350
Neptune	4500	51 000	165 years	16	-220
Pluto	5900	2300	248 years	150	-220
Saturn	1400	120 000	29 years	10.2	-180
Uranus	2900	51 000	84 years	17	-210
Venus	110	12 000	225 days	5800	+480

d. Which piece of information shows that it is it **not** possible for liquid water to be present on the surface of Jupiter?

In all the questions where students needed to access information from an information box to answer a specific question there was a significant difference in performance favouring the paper version in at least one of the items within each of these questions.

From the qualitative evidence, students were unconcerned about using information boxes in the on-screen tests and found them helpful and straightforward to use. A few students did comment that they could obstruct questions, but generally they were not an issue for students. Teachers on the other hand, were more concerned that their students would find the information boxes difficult to access and use and that they would be disadvantaged by not having necessary data and information available in the manner in which paper based assessment information is provided.

The teachers concerns seem to be borne out by the performance evidence on these questions. Although these questions were generally targeted at higher ability levels, they produced significant modal differences in performance across all ability levels.

9.5.3 The use of colour photographs and diagrams

The use of colour in paper-based tests is not unheard of, however it is rarely used due to the prohibitive production costs. In an on-screen environment, the use of colour carries no extra costs, and can be used as a commonplace straightforward facility.

The use of colour in the on-screen tests was very popular with students and teachers. There was not a single negative comment obtained through questionnaires or interviews. Interestingly, their comments only surrounded the look and feel of the assessments rather than any potential assessment benefits. Commonly, students and teachers suggested that the use of colour in questions enlivened them and made the assessment experience more engaging.

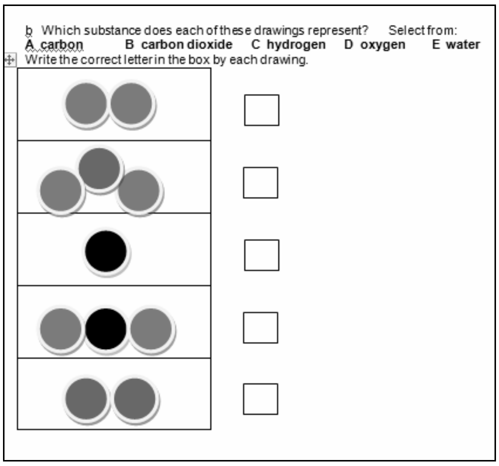
When analysing the assessment consequences of using colour, it is important to clarify the intended purpose and function of the colour in the question or item.

Colour photographs or diagrams were used in a number of items and many of them did not demonstrate any significant benefits or disadvantage to the students. However, the following items did, for possibly differing reasons, which will be discussed.

9.5.3.1 Advantageous use of colour.

Figure 105: Screenshots of Test 1, Q12b in paper and computer modes

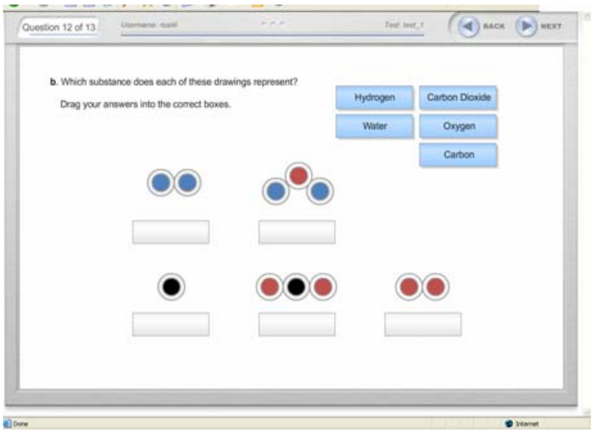
Paper Version



b Which substance does each of these drawings represent? Select from:
A carbon B carbon dioxide C hydrogen D oxygen E water
Write the correct letter in the box by each drawing.

□
□
□
□
□

Computer Version



Question 12 of 13

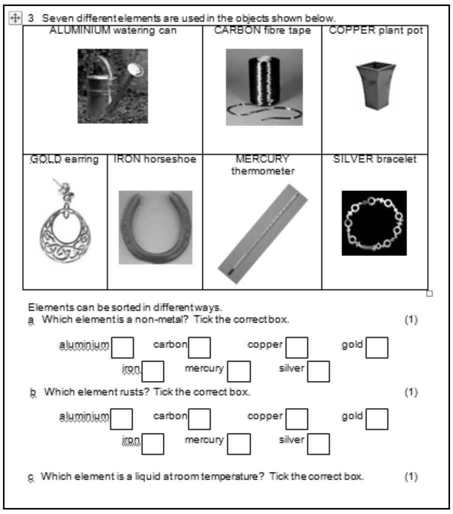
b. Which substance does each of these drawings represent?

Drag your answers into the correct boxes.

Figure 105 above shows an item that different in mode largely by the use of colour. The clarity in identifying different atoms through colour differentiation rather than the use of a grey scale would seem to be advantageous, and this was clearly supported by the empirical evidence. An interesting aspect was that the lower ability students, who would have found this item challenging, were not particularly advantaged by the colour. The group for whom it made the most difference were the students for whom this item was targeted (the more able students). It could be argued that the use of colour negated a construct irrelevant variable and enabled students who understood the scientific concept to demonstrate their understanding.

Figure 106: Screenshots of Test B, Q3b in paper and computer modes

Paper version



3 Seven different elements are used in the objects shown below.

Elements can be sorted in different ways.

a Which element is a non-metal? Tick the correct box. (1)

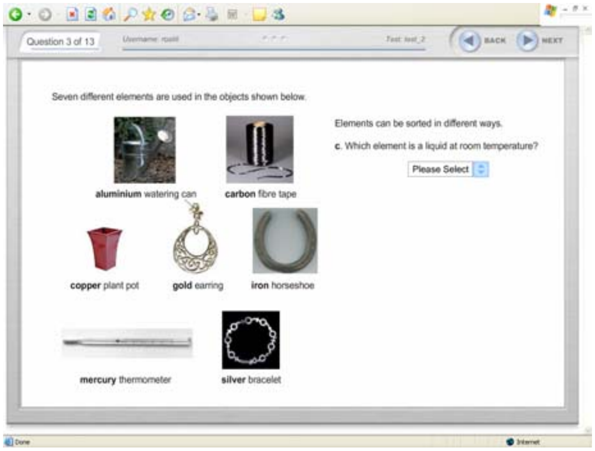
aluminium □ carbon □ copper □ gold □
iron □ mercury □ silver □

b Which element rusts? Tick the correct box. (1)

aluminium □ carbon □ copper □ gold □
iron □ mercury □ silver □

c Which element is a liquid at room temperature? Tick the correct box. (1)

Computer version



Question 3 of 13

Seven different elements are used in the objects shown below.

Elements can be sorted in different ways.

c. Which element is a liquid at room temperature?

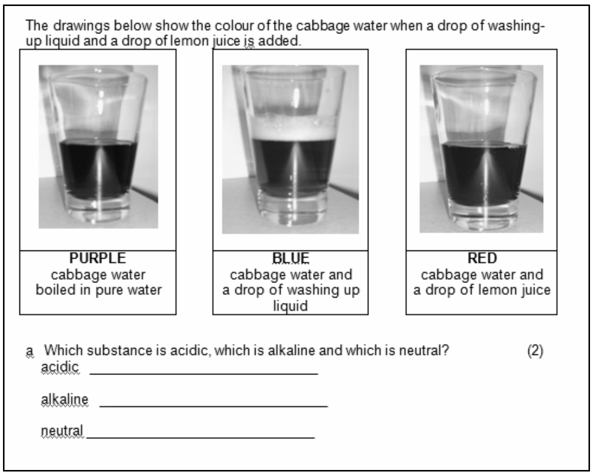
Please Select

Figure 106 above shows an item that showed a significant differential item function in favour of the computer version was part c. The clarity that the use of colour offers in this question significantly advantaged less able students.

9.5.3.2 Disadvantageous use of colour

Figure 107: Screenshots of Test 2, Q4 in paper and computer modes

Paper Version



The drawings below show the colour of the cabbage water when a drop of washing-up liquid and a drop of lemon juice is added.

PURPLE cabbage water boiled in pure water	BLUE cabbage water and a drop of washing up liquid	RED cabbage water and a drop of lemon juice

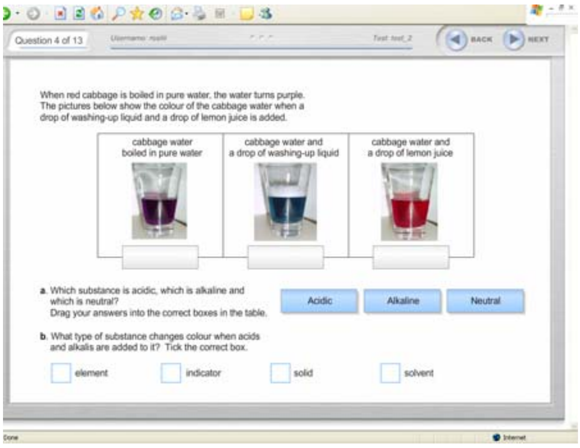
a Which substance is acidic, which is alkaline and which is neutral? (2)

acidic ______________________________

alkaline ______________________________

neutral ______________________________

Computer Version



Question 4 of 13

When red cabbage is boiled in pure water, the water turns purple. The pictures below show the colour of the cabbage water when a drop of washing-up liquid and a drop of lemon juice is added.

cabbage water boiled in pure water	cabbage water and a drop of washing-up liquid	cabbage water and a drop of lemon juice

a. Which substance is acidic, which is alkaline and which is neutral? Drag your answers into the correct boxes in the table.

Acidic Alkaline Neutral

b. What type of substance changes colour when acids and alkalis are added to it? Tick the correct box.

□ element □ indicator □ solid □ solvent

Question 4 in Test B shown above in Figure 107 raises interesting issues about construct validity of science assessment. The item above on the left shows how the question was presented on paper. The colour version on the right was the on-screen version. From first observation, it would appear that the computer version enhances the validity of the question. Students had to observe the colours of the beakers and then identify them as acidic, alkaline or neutral. This observational skill is a commonplace task for science students and therefore reflects the teaching and learning expectation and the subject construct to be assessed.

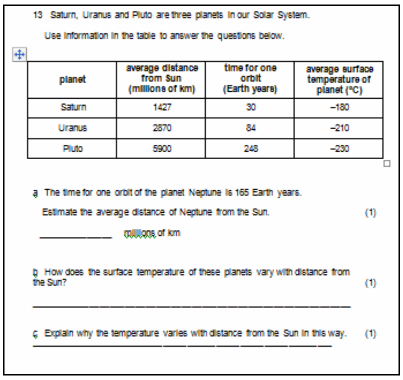
This item however showed a very significant DIF in favour of the paper version. The reason is simply explainable; the paper version cannot show colour, therefore the colour names are printed under the beakers. Visual information from the beakers is therefore redundant in paper mode. All the students needed to do in this question was to read the colour and then identify the colour name to acid, alkaline or neutral. The area of science to be assessed in this question is clearly limited in the paper version, and does not reflect the observational skills necessary to carry out this deductive process skill. The data from this item did not suggest that the use of colour in itself was disadvantageous, however students have become accustomed to gaining the required information they need by any simpler alternative means.

9.5.4 Tables of information

Apart from providing information boxes, there were items across both tests that contained tables of data on the paper and on-screen versions. If a data table could comfortably fit onto the screen alongside the question without the need to include an information box, this was the preferred choice. A surprising number of questions that contained data tables demonstrated a modal DIF performance. An example is shown in Figure 108 below:

Figure 108: Screenshots of Test B, Q13 in paper and computer modes

Paper Version



13 Saturn, Uranus and Pluto are three planets in our Solar System.

Use information in the table to answer the questions below.

planet	average distance from Sun (millions of km)	time for one orbit (Earth years)	average surface temperature of planet (°C)
Saturn	1427	30	–180
Uranus	2870	84	–210
Pluto	5900	248	–230

a The time for one orbit of the planet Neptune is 165 Earth years.

Estimate the average distance of Neptune from the Sun. (1)

_____________ millions of km

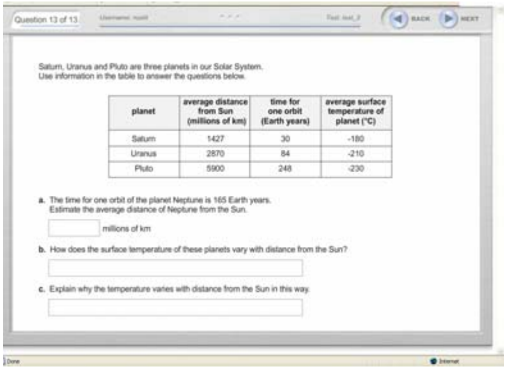
b How does the surface temperature of these planets vary with distance from the Sun? (1)

____

c Explain why the temperature varies with distance from the Sun in this way. (1)

____

Computer Version



Saturn, Uranus and Pluto are three planets in our Solar System.
Use information in the table to answer the questions below.

planet	average distance from Sun (millions of km)	time for one orbit (Earth years)	average surface temperature of planet (°C)
Saturn	1427	30	-180
Uranus	2870	84	-210
Pluto	5900	248	-230

a. The time for one orbit of the planet Neptune is 165 Earth years.
Estimate the average distance of Neptune from the Sun.

millions of km

b. How does the surface temperature of these planets vary with distance from the Sun?

c. Explain why the temperature varies with distance from the Sun in this way.

The two versions shown above were almost identical in their style and layout, as were many of the other data tables presented in paper and on-screen modes. All the items in the question shown above demonstrated highly significant DIFs favouring paper mode. An operational difference in all these question types was the way students had to interact with the data. They could not make any annotations on the computer version or use a ruler to locate and read off information. Although students and teachers did not comment that these item types presented any visual difficulties, the number of DIFs on these item types does suggest that there was some form of modal disadvantage, even if it was simply a lack of familiarity with the on-screen item types.

9.6 Response Types

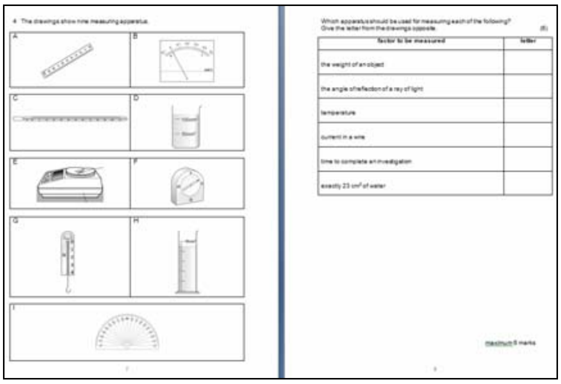
The test questions included a variety of response types. Analysis of items performing with highly significant DIFs highlighted a few issues surrounding some of these response mechanisms.

9.6.1 Tick boxes/drop down lists

There were a few items across both tests that were almost identical in terms of the layout of the questions and even the diagrams used. The only difference in these items was the on screen response mechanisms. An example is shown below in Figure 109.

Figure 109: Screenshots of Test A, Q4 in paper and computer modes

Paper Version



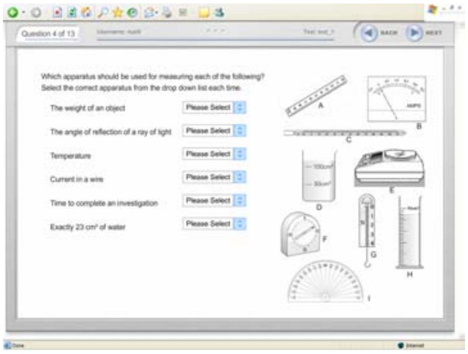
4 The drawings show nine measuring apparatus.

Which apparatus should be used for measuring each of the following?
Give the letter from the drawings opposite. (6)

factor to be measured	letter
the weight of an object	
the angle of reflection of a ray of light	
temperature	
current in a wire	
time to complete an investigation	
exactly 23 cm³ of water	

maximum 6 marks

Computer Version



Which apparatus should be used for measuring each of the following?
Select the correct apparatus from the drop down list each time.

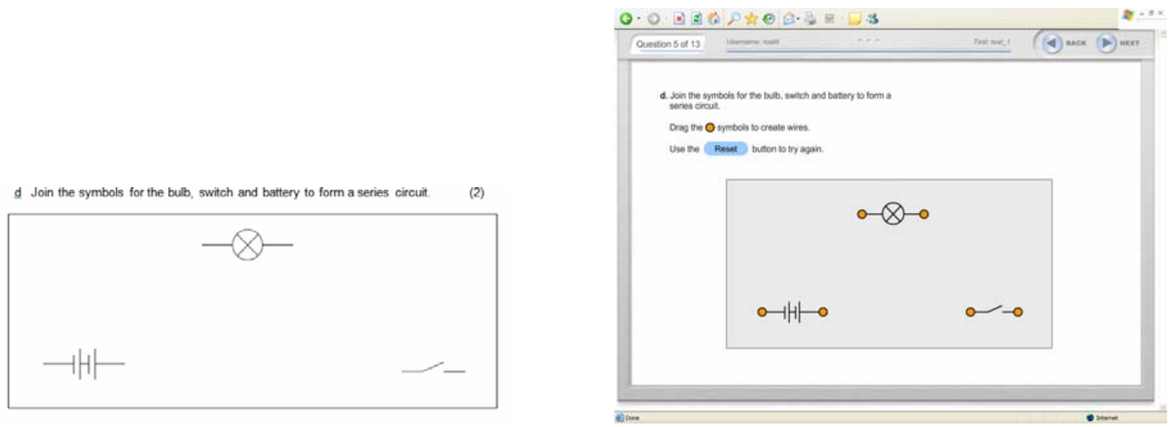
The weight of an object	Please Select
The angle of reflection of a ray of light	Please Select
Temperature	Please Select
Current in a wire	Please Select
Time to complete an investigation	Please Select
Exactly 23 cm³ of water	Please Select

It might be suggested that the identical artwork was larger and perhaps clearer in the paper-based versions, but it could also be suggested that students are not quite as comfortable in the use of these mechanisms as they think they are. Students rated these mechanisms as very easy to understand and to operate, however for some reason a number of these item types demonstrated significant DIFs in favour of paper-based versions. The modal differences were more marked at lower ability levels.

9.6.2 Drawing

There were only two items across the tests that required a drawing response mechanism. However, both of these items demonstrated different issues, which will be addressed in turn.

Figure 110: Screenshots of Test A, Q5 in paper and computer modes



Paper Version

d Join the symbols for the bulb, switch and battery to form a series circuit. (2)

Computer Version

d. Join the symbols for the bulb, switch and battery to form a series circuit.

Drag the ● symbols to create wires.

Use the Reset button to try again.

The first item is shown above in Figure 110. The paper version of this item would be fairly familiar for most students; they had to draw a series circuit from the given symbols. The computer version used a drag and drop mechanism where students picked up a connection at the ends of one of the given symbols and join it to another symbol. Although students would have been unfamiliar with the computer version mechanism, the difference in performance and subsequent DIF was higher than any other item on either test in favour of the computer version. Reasons for this DIF raises questions about the level of demand of this response type and also one of construct validity.

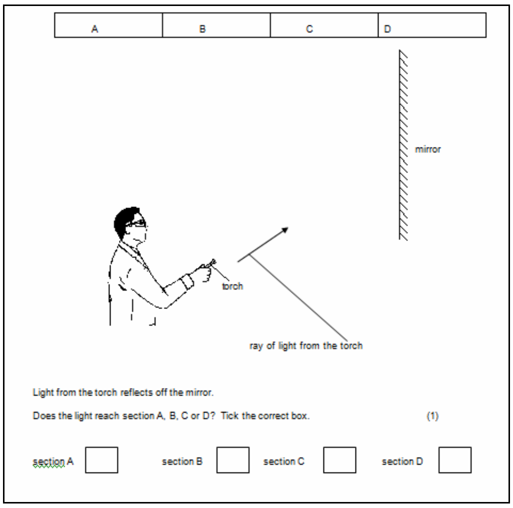
One of the main reasons students get paper-based versions of these questions incorrect is because of a lack of accuracy when connecting circuit lines to symbols. They often leave small gaps which if more than 0.1cm, is deemed to indicate a break in a circuit and therefore not creditworthy. These rubrics were applied to the marking of the paper versions. In the computer based versions, it was not possible for students to leave gaps. As soon as a circuit line approached a symbol it clicked together. Therefore the possible pathways that students could draw and connect to were more restricted on the computer version than on paper.

It is therefore unsurprising that performance was much better on the computer version than on paper. However whether the paper version actually is a more construct valid question is an issue. It might be questioned as to what the paper version is actually assessing; an understanding of constructing circuits or an exercise in visual acuity and drawing skills?

The second item is shown below in Figure 111. This item also used a drawing tool, however in this case, it was not used as a direct response mechanism, but as an aid to establishing the answer to the item.

Figure 111: Screenshots of Test B, Q6 in paper and computer modes

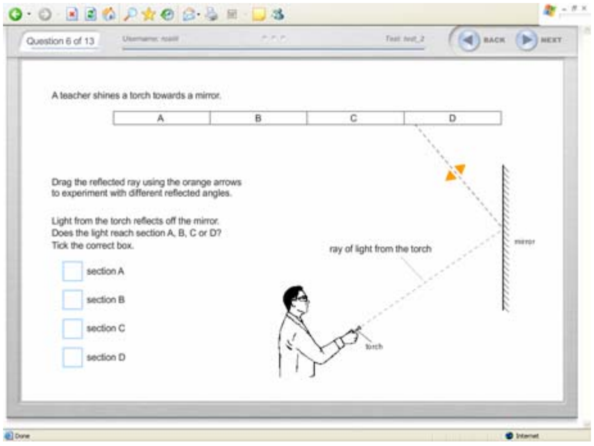
Paper Version **Computer Version**



Light from the torch reflects off the mirror.

Does the light reach section A, B, C or D? Tick the correct box. (1)

section A □ section B □ section C □ section D □



A teacher shines a torch towards a mirror.

Drag the reflected ray using the orange arrows to experiment with different reflected angles.

Light from the torch reflects off the mirror.
Does the light reach section A, B, C or D?
Tick the correct box.

□ section A
□ section B
□ section C
□ section D

The item above on the left shows the paper-based version. This is a well used way of asking questions about reflective angles. Students can use a protractor or ruler to extend the ray from the torch onto the mirror and then work out where on the screen at the top the reflected ray will be seen. In the computer version, students were not able to use a protractor or ruler, therefore a mechanism was provided for them to move the reflecting ray to an angle that they thought it would make and then select this position from the four options given.

The performance on paper for this item was as expected and the item discriminated moderately well. On-screen, this item performed very poorly and had a negative discrimination, ie. there was a negative correlation between overall performance across the rest of the test and the performance on this item; students doing well on the test as a whole were getting this item wrong and vice versa.

A large number of students gave the answer as option D in this item, which was the default setting of the ray on the screen. Many students ignored the instruction on the screen to drag the ray with the orange arrows to the reflected position, and simply gave the answer at the default position.

Performance on examples 1 and 2 were therefore completely different in their performance. Although students had no preparation for the on-screen tests, they picked up the mechanism of the circuit diagram and operated it with high success. In the reflected ray question, it may have appeared to be a simple closed response question, and students took it at face value, without reading the simple instruction on the screen. It is therefore unwise to assume that students are always as comfortable with on-screen mechanisms as they claim. An outcome of this question might be to ensure that students could not make this simple mistake; to construct the question with the reflected ray at a neutral angle.

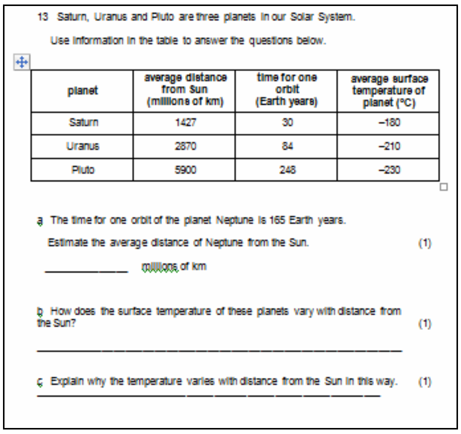
9.6.3 Open Responses

A number of open response items across both tests demonstrated significant modal differences. In every case high DIF significance was in favour of the paper versions. It might be considered that open responses are the least altered response type from paper to computer, i.e. the use of language. However, the empirical and qualitative evidence indicate that this response type had a number of performance issues associated with modality. This difference was exemplified more in the investigations which will be discussed in this chapter after analysis of the tests. However, the following question illustrates issues associated with certain open response types.

There were three category codings for open responses; numeric, short and extended response. The question below uses two of these item types. All the items in this question demonstrated highly significant DIFs. Numeric questions can often cause difficulties in on-screen versions due to the restricted marking mechanisms employed. All creditworthy responses must be pre-determined and programmed and therefore strict rubrics are applied to creditworthy answers in order for automatic marking to operate. This can lead to a lack of marker judgement in on-screen numeric answers compared to paper versions. It was interesting to note that in the one item that used an on-screen calculator, there was no significant difference in performance across modes.

Figure 112: Screenshots of Test B, Q13 in paper and computer modes

Paper Version



13 Saturn, Uranus and Pluto are three planets in our Solar System.

Use information in the table to answer the questions below.

planet	average distance from Sun (millions of km)	time for one orbit (Earth years)	average surface temperature of planet (°C)
Saturn	1427	30	–180
Uranus	2870	84	–210
Pluto	5900	248	–230

a The time for one orbit of the planet Neptune is 165 Earth years.

Estimate the average distance of Neptune from the Sun. (1)

____________ millions of km

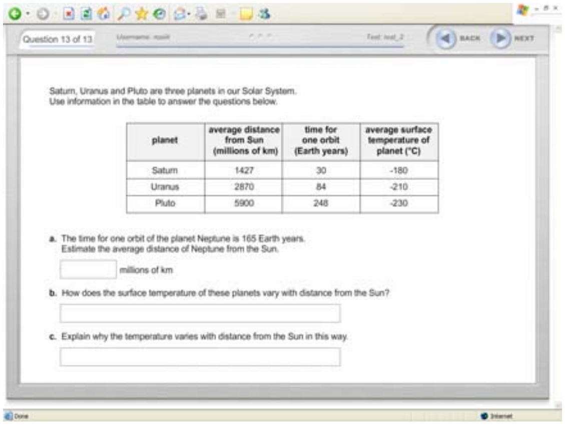
b How does the surface temperature of these planets vary with distance from the Sun? (1)

____

c Explain why the temperature varies with distance from the Sun in this way. (1)

____

Computer Version



Saturn, Uranus and Pluto are three planets in our Solar System.
Use information in the table to answer the questions below.

planet	average distance from Sun (millions of km)	time for one orbit (Earth years)	average surface temperature of planet (°C)
Saturn	1427	30	-180
Uranus	2870	84	-210
Pluto	5900	248	-230

a. The time for one orbit of the planet Neptune is 165 Earth years.
Estimate the average distance of Neptune from the Sun.

millions of km

b. How does the surface temperature of these planets vary with distance from the Sun?

c. Explain why the temperature varies with distance from the Sun in this way.

Figure 112 above show parts b and c of Q13 which were not automatically marked, they were captured and marked by expert markers. Both of these items showed significant DIFs across modes favouring paper.

Qualitative evidence illustrated interesting views and perceptions concerning open response items. A popular view from student questionnaires and interviews was that they preferred on-screen tests because there was less writing involved in these forms of assessment and when it was required, any poor handwriting skills would be negated by being able to type their responses and that they could easily amend or correct their responses. However, there were concerns by some students that typing their answers into the space given on-screen was restrictive compared to paper; they felt they had less space to respond.

The short open response sections in the on-screen versions were designed to look similar to a paper-based presentation. The available space usually indicates the expected length of a response by the space given in the answer box. This is a similar issue in paper-based assessments. The number of lines given as an answer space indicates the expected length of response; however, if a student extends a response more than the allotted line, they will not be penalised. This was also possible in the on-screen assessments. If a student wrote a more extended response, it would still be captured and marked, however the text scrolled across so that the whole response could not be read at a glance.

In most cases, the difference in performance between the modes in these short response items were not caused by students writing more in the on-screen versions, but markedly less than the space allowed. There was clearly a difference in perception of students preferring to answer open responses in an on-screen environment and believing that this would enhance their performance and then the outcome evidence demonstrating that in most cases their performance deteriorated in this mode.

9.7 Analysis of the Investigations

Each student in each sample group completed a test and investigation in paper-based and computer mode. Therefore, for example a student would take Test A and Investigation A in one mode and then Test B and Investigation B in the alternative mode.

Both Investigations tested the same process skills and the format of each investigation was identical. There were only two differences in their design. One was the context used. Investigation A used a laboratory context of a biology photosynthesis investigation, while Investigation B used an applied context of a physics forces investigation. The other difference was that the photosynthesis investigation produced a curved graph from a complete available range of data whereas the forces investigation produced a straight line graph from the available range of data.

The nature of the Investigations had to be different in differing modes. The intention of the on-screen version was to engage the students in an investigation and require them to collect and use their own data. This would not be possible in a paper-based format and data was given to students in this mode. This difference in approach could not be measured in any quantitative manner and resulted in the paper-based and on-screen versions being made up of differing mark totals. However, this comparative difference was meant to be judged qualitatively in terms of the views of students and teachers concerning the authenticity and fitness for purpose of the assessment of scientific process skills.

It was not possible therefore to compare performance in the initial stages of the investigations and note any DIFs, however there was a high level of concensus that the on-screen facility of collecting and then using their own data sets was highly desirable in the teaching, learning and assessment of science process skills.

Quantitative modal comparisons were possible once data had been collected. It was interesting that there were DIFs in performance in the same items across both investigations. These will now be considered.

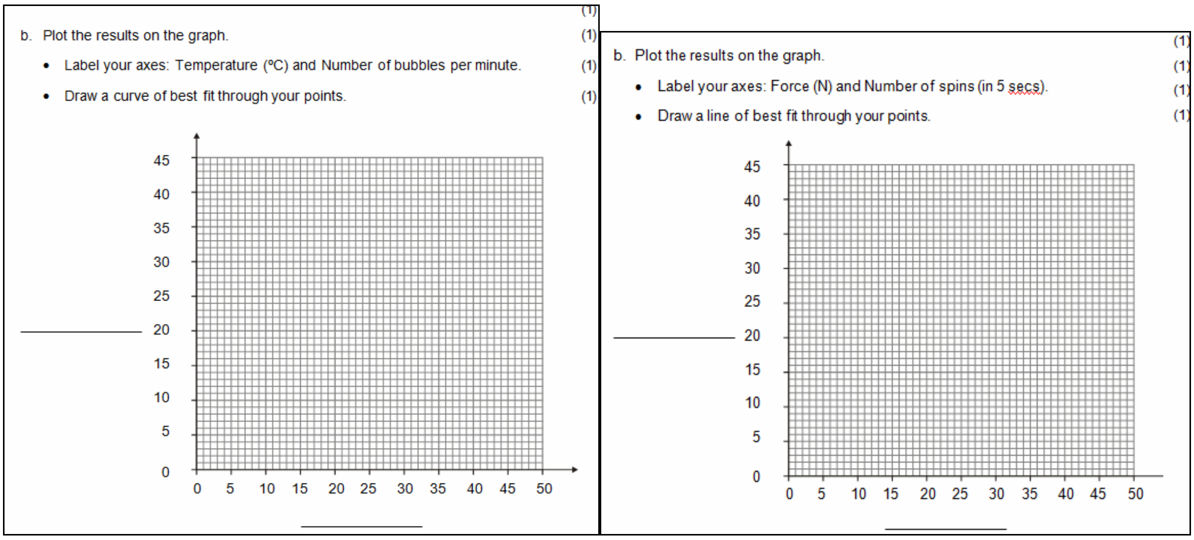
9.7.1 Graph Plotting

In both investigations, there was a DIF in the performance of plotting graphs in favour of the paper-based versions. This was a high DIF for the forces investigation, and smaller for the photosynthesis investigation. There was evidence from the qualitative data that some students had difficulty plotting points in the on-screen versions, particularly those using touch-pads rather than a mouse, however comparative qualitative data suggested that most students found the on-screen manipulative skills straightforward and preferred them to the paper-based versions. Most students and teachers concerns regarding the on-screen tests and investigations centred on the possible loss of data rather than any modal difficulty.

The tolerance levels for the graph plotting skills were set at the same level for both modes, however the marking mechanisms were different. In the on-screen versions, the graph plotting was automatically marked to the agreed tolerance limit. The paper versions were marked by humans. Whether the marking was more accurate on the on-screen versions, and therefore the higher levels of performance on paper were in part due to marking error is an interesting question.

Figure 113 below shows how the graph plotting sections were presented in both investigations across modes,

Figure 113: Screenshots of Investigation A and B, Q1 in paper and computer modes



(1)

b. Plot the results on the graph. (1)

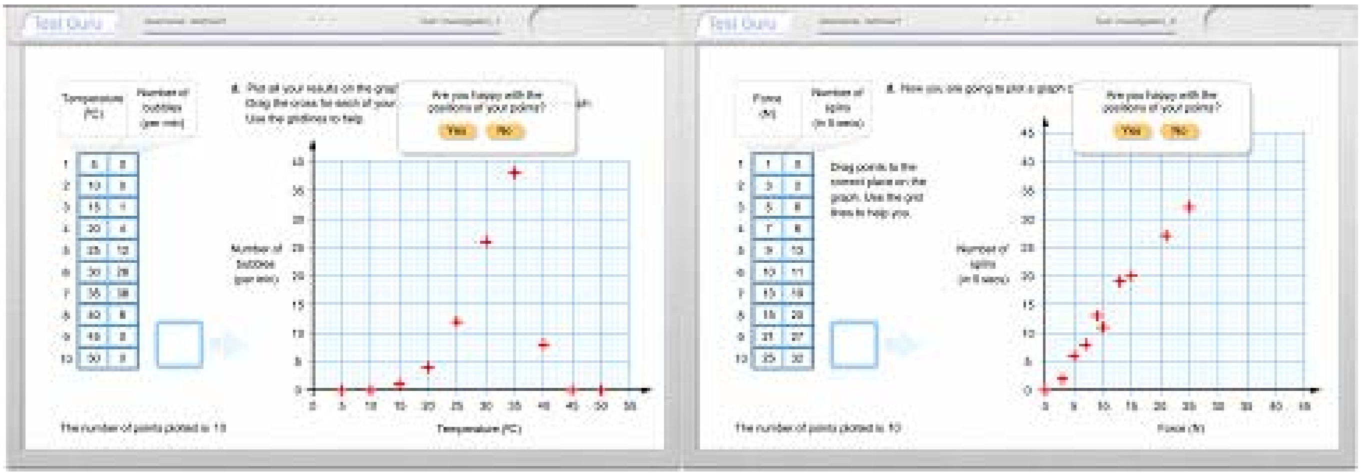
- Label your axes: Temperature (ºC) and Number of bubbles per minute. (1)
- Draw a curve of best fit through your points. (1)

b. Plot the results on the graph. (1)

- Label your axes: Force (N) and Number of spins (in 5 secs). (1)
- Draw a line of best fit through your points. (1)

(1)

Computer Versions



	Temperature (°C)	Number of bubbles (per min)
1	5	0
2	10	0
3	15	1
4	20	4
5	25	12
6	30	26
7	35	38
8	40	8
9	45	0
10	50	0

d. Plot all your results on the graph. Drag the cross for each of your ... Use the gridlines to help.

Are you happy with the positions of your points?

The number of points plotted is 10

	Force (N)	Number of spins (in 5 secs)
1	1	0
2	3	2
3	5	6
4	7	8
5	9	13
6	10	11
7	13	19
8	15	20
9	21	27
10	25	32

d. How you are going to plot a graph ...

Drag points to the correct place on the graph. Use the grid lines to help you.

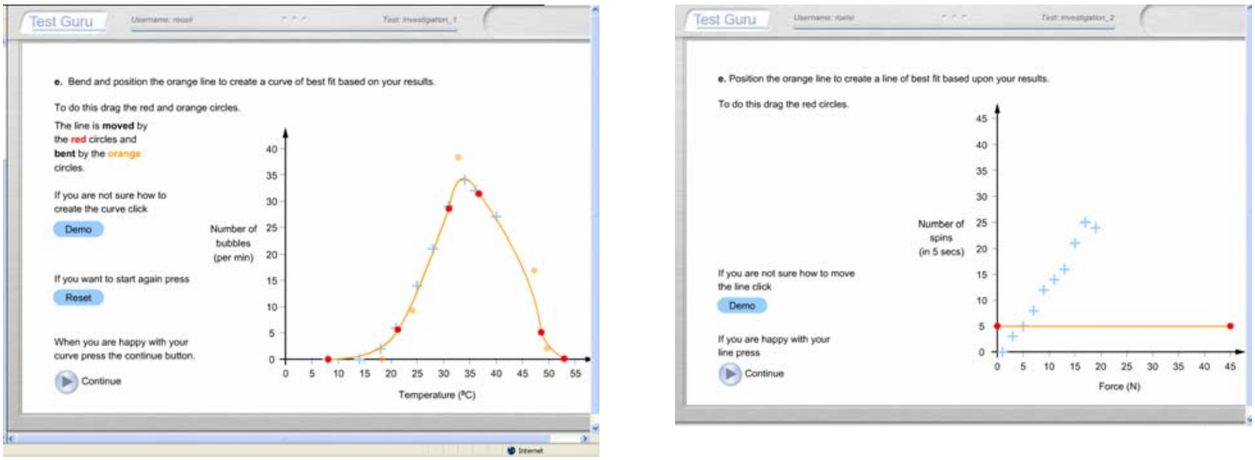
Are you happy with the positions of your points?

The number of points plotted is 10

After students had plotted their points on the graph, they then had to produce lines of best fit. In the photosynthesis investigation students had to produce a curve of best fit, whereas in the forces investigation they had to produce a line of best fit. The paper and on-screen formats are shown in Figure 114 below.

The on-screen mechanism was quite exploratory. A demonstration was available to students to show them how to manipulate lines and curves to produce lines of best fit. Interestingly, there was not any significant difference in performance in the paper and on-screen version for this item, although the performance was very low across modes.

Figure 114: Screenshots of the computer lines of best fit within Investigations A and B



e. Bend and position the orange line to create a curve of best fit based on your results.

To do this drag the red and orange circles.

The line is **moved** by the red circles and **bent** by the orange circles.

If you are not sure how to create the curve click

If you want to start again press

When you are happy with your curve press the continue button.

e. Position the orange line to create a line of best fit based upon your results.

To do this drag the red circles.

If you are not sure how to move the line click

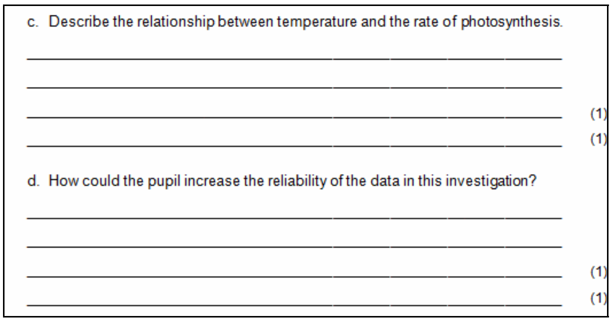
If you are happy with your line press

9.7.2 Open Responses

The open responses required in the investigations were the most extended types used in this comparability study. The style and layout of the two modes are shown in Figure 115 below:

Figure 115: Screenshots of Investigation A and B, Q1c in paper and computer modes

Paper version



c. Describe the relationship between temperature and the rate of photosynthesis.

____

____

____ (1)

____ (1)

d. How could the pupil increase the reliability of the data in this investigation?

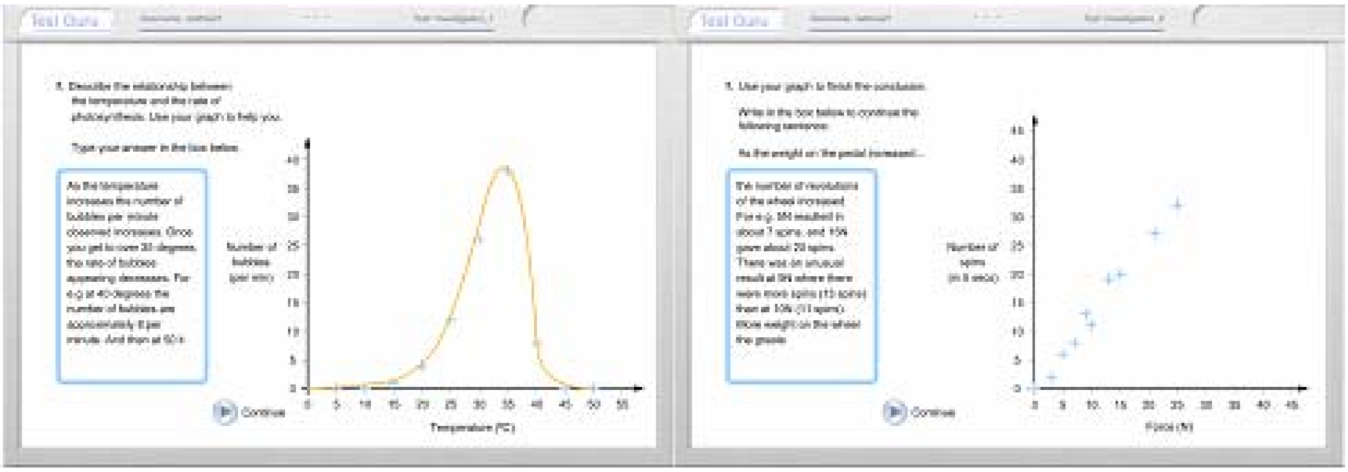
____

____

____ (1)

____ (1)

Computer Version



1. Describe the relationship between the temperature and the rate of photosynthesis. Use your graph to help you.

Type your answer in the box below.

As the temperature increases the number of bubbles per minute observed increases. Once you get to over 35 degrees, the rate of bubbles appearing decreases. For e.g at 40 degrees the number of bubbles are approximately 8 per minute. And then at 50 n

1. Use your graph to finish the conclusion.

Write in the box below to continue the following sentence:

As the weight on the pedal increased...

The number of revolutions of the wheel increased. For e.g. 5N resulted in about 7 spins, and 15N gave about 20 spins. There was an unusual result at 9N where there were more spins (13 spins) than at 10N (11 spins). More weight on the wheel the greater

The amount of space given to students in the on-screen versions was generous, compared to the space available in the paper-based versions. The presentation of these items allowed students to see their full response on-screen without the need to use scrolling. However, the general comparative trend indicated that students wrote less in the on-screen compared to the paper versions, and had achieved significantly lower levels of performance. Comparative performance for both group samples, across both investigations provided the same outcomes and therefore produced robust evidence in this study on significant differential performance when students were required to write extended open responses on computer.

9.8 Summary

This chapter has synthesised the quantitative and qualitative evidence from the tests and investigations and identified and discussed key areas that caused modal differences in performance. For these areas the empirical evidence of performance was contrasted with student and teacher views on the appropriateness of the assessments in different modes. Some of the performance differences were found in items that had differing stimuli between paper and on-screen versions; other differences were found in items that had differing response mechanisms. I have tried to categorize these differences as being the result of either construct relevant or irrelevant factors; that is, factors caused by differences in what is being assessed and differences caused by the mode. Sometimes however, these two factors overlap.

The different forms of stimuli used in the on-screen versions included the use of video sequences, colour diagrams and photographs and pop-up information boxes. Where these stimuli involved the assessment of different aspects of science constructs compared to paper-based versions, students performed less well in the on-screen versions. Paradoxically, these questions were much preferred by students and teachers as assessment instruments. When these stimuli did not affect the construct being assessed, generally they also resulted in lower student performances, which suggested that student confidence in using on-screen mechanisms was in part, optimistic.

The different response mechanisms used in the on-screen versions included tick boxes, drag and drops and drop down lists, drawing facilities and free text responses. Most of these mechanisms introduced construct irrelevant differences which resulted in slightly lower student performance. Similarly to the use of different stimuli, the performance evidence was at odds with the preference and confidence expressed by students.

The only construct relevant factor included in the response mechanisms was the circuit drawing tool. This resulted in students performing significantly better in the on-screen versions. This may have been the result of less opportunity for students to make drawing mistakes compared to the paper based versions, although it could also be suggested that the essential circuit understanding was assessed better on-screen as there was no emphasis on drawing acuity skills. The marking of the on-screen versions was also more reliable than the paper versions.

The next chapter will discuss how these differences impact on the issues of equivalence, reliability and validity.

Chapter 10

Discussion

10.1 Introduction

The second chapter of this thesis outlined the central concepts of the reliability and validity of assessments; their technical definitions, relationship to each other and their role, use and purposes in the measurement and consideration of the quality of assessment instruments. Any discussion on the appropriateness of on-screen assessments therefore must return to these key indicators in order to consider whether on-screen assessments can offer similar levels of reliability and validity in comparison to paper-based versions and preferably offer enhancements to the status quo.

Rather than discuss reliability and validity as distinctly separate aspects in considering the appropriateness of on-screen science assessments in this comparative study, they will first be discussed through the emergent issues of the comparative study. It should be fairly clear whether these issues have more links with reliability or validity, however there are overlapping themes. After discussing the main emergent issues, I will return to consider whether modal differences affect any of the threats to validity discussed in Chapter 2.

The first issue to consider is the measurement of internal reliability. Chapter 2 discussed the facets of reliability, how these areas can be quantifiably measured and the levels of confidence that can be placed on these figures. In terms of the measurement of internal reliability, how did the on-screen science tests and investigations measure up?

The Internal reliabilities of the assessments were calculated using Cronbach's Alpha, and then using the standard deviation, the Standard Error of Measurement (SEM) was calculated. This data is shown again Table 64 below:

Table 64: Standard Error of Measurement for Tests and Investigations A and B

Test Version	Cronbach's alpha	Standard Deviation	Standard Error of Measurement
Test A: Paper	0.919	13.20	3.76
Test A: Computer	0.927	16.48	4.45
Test B: Paper	0.930	15.99	4.23
Test B: Computer	0.930	13.34	3.53

Investigation Version	Cronbach's alpha	Standard Deviation(SD)	Standard Error of Measurement(SEM)
Investigation A: Paper	0.677	2.11	1.20
Investigation A: Computer	0.594	2.17	1.38
Investigation B: Paper	0.664	2.06	1.19
Investigation B: Computer	0.595	2.35	1.50

In Chapter 7, accounting for ability differences between the two student groups and difficulty differences between the test versions, a statistically significant 2 mark modal difference was established in favour of the paper-based versions. Interestingly, the SEM variance within the paper and computer-based tests was approximately twice the modal difference. This does not suggest that the paper and computer based tests were actually equivalent; they were proven not to be. However these figures do highlight how equivalence is not the only issue that can affects student scores and any potential differences between actual and true scores. SEM data is not usually a feature of public discussion concerning the reliability of assessments, and its value as a measurement tool is questionable, however it does remind us that assessment accuracy is almost impossible as any reliability measure will usually be less than 1.

Chapter 5 presented a table, from Texas University (Table 5, page 79) to show generalised evaluations of Cronbach's alpha figures and their acceptability for differing forms of assessment. The Cronbach's alpha figures for both test versions, across both modes were high enough to give confidence that, if any of these tests were to be used in an external assessment, they would perform as reliable assessments, and that they should consistently establish a consistent rank order of candidates.

The high figures obtained indicate that each of the tests were testing the same construct within each test (a pre-requisite of Cronbach's alpha). Considering that the on-screen and paper versions were based on the same questions, the high figures obtained across modes also indicates that the on-screen and paper-based tests were also assessing the same construct. There was a significant modal mark difference, which will be discussed in this chapter, however, one of my research questions focussed on whether modes of assessment would affect the intended constructs. I would suggest that they did not; the modes may have affected some construct validity aspects, but not the construct itself.

10.2 Equivalence

In Chapter 3, the issue of equivalence was discussed. This comparability study showed a significant difference of 2 marks in favour of paper-based assessments. This may not be a 'true' figure, once construct irrelevant variables are removed from the assessment environment. However, at this stage of research, testing a variety of differing stimulus and response mechanisms to students who had no preparation for these, in either teaching or learning strategies used by the school or in any on-screen test practice, it would have been surprising, even extraordinary if there had been no difference in performance. It might even have been disappointing if the assessment of science using a different approach to validity had resulted in exactly the same assessment outcomes.

However, this leads on to the next issue and a continuing dilemma concerning the introduction of computer based testing in England.

In high stakes test environments in England, the default examination mode is paper-based. There are a number of possible scenarios for the integration of computer-based exams:

A simple outcome would be that particular specifications could be developed with computer-based tests being the intended assessment mode. Any particular content or construct can be assessed with at least the same levels of reliability and validity of paper, and crucially there wouldn't be any equivalence issues. Standard setting would have to be adapted to accommodate scrutiny of on-screen responses, however, given that all the responses are stored electronically on a server, script scrutiny could easily access the whole cohort, rather than the paper-based position where usually only a small sample of scripts are available. This is a significant advantage.

A more realistic outcome would be that in the short to medium term, there would be school and student choice concerning the mode of assessment. The status quo now, in early 2010, is that if components of a qualification provide a choice of paper-based or computer based multiple choice questions, they are deemed to be equivalent in terms of outcomes and awarding decisions. This is under the conditions of a like-for-like assessment, where the appearance and layout of the questions are the same in each mode. While it is expected that exam boards monitor equivalence issues, there is a regulatory acceptance of equivalence (QCA, 2009).

In the case of on-screen tests consisting of the type of stimulus and response mechanisms contained in this comparability study, there is clearly not equivalence. So what to do?

The reliability of the tests in both modes was very high, and therefore they acted as effective assessment instruments. However some of the questions do seem to be assessing the construct of science in differing ways, and therefore the tests, although both having reasonable levels of construct validity, seem to have some differing forms of construct validity. At the same time the modes themselves present differing construct irrelevant factors. There would appear to be two options in the short to medium term; either continue to run comparative trials until stability of scoring emerges or offer differing modal versions, and incorporate judgemental and statistical weightings at the awarding stage.

By 'stability' of scoring I do not mean trialling until the scores are the same, but rather to a point where there is empirical and qualitative evidence to show that construct irrelevant factors have been identified and minimised, and therefore there is more confidence and reliability concerning the differences between the assessments across modes, across cohorts.

The second option seems to be dangerous; in terms of the defence of making 'adjustments' between modes on the basis of partial evidence. Even if a regulator was satisfied with such script scrutiny arrangements, it would be likely that if such methodologies were made available to schools and the public at large, there would be swift perception and judgement about which mode was 'easier' and therefore all the subsequent 'standards' and 'fairness' debates.

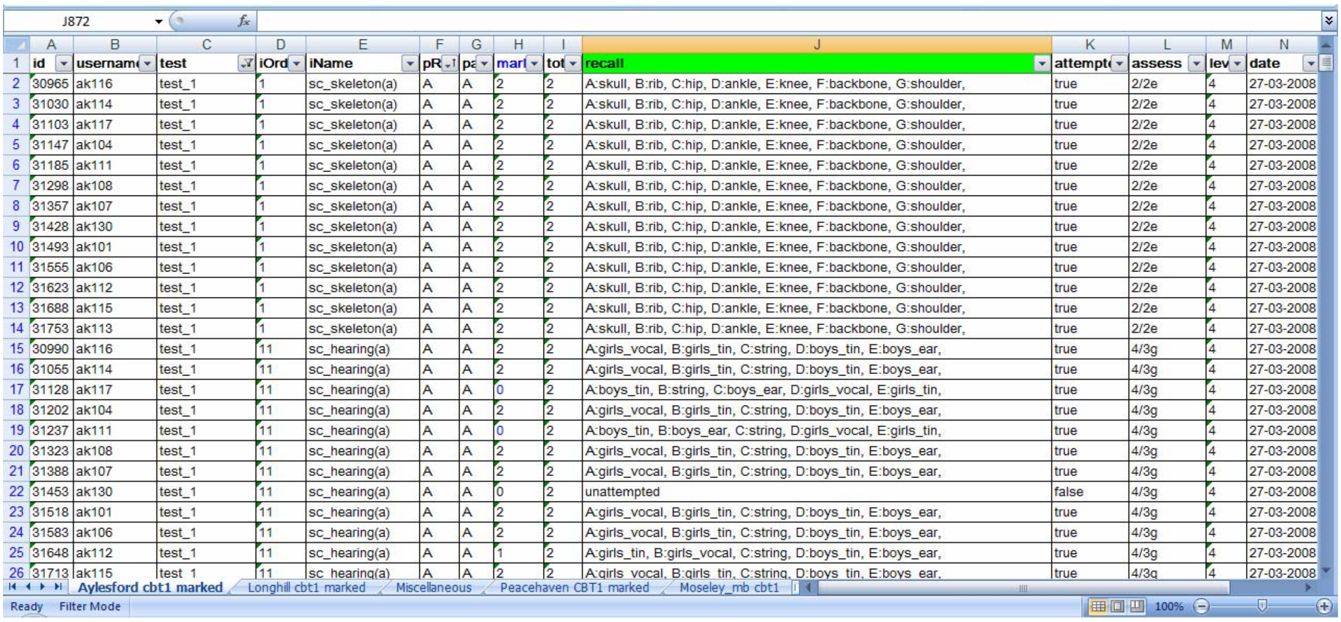
10.3 Marking Reliability

This comparability study did not carry out any analysis of marking reliability between the two modes, however there are a few issues that can be discussed at this stage.

Any system that can automatically score and record responses is going to have higher reliability than human marking. One of my underlying design principles for these science assessments was that objective questioning has a place in a summative assessment, but not alone. The on-screen versions used a variety of mechanisms to ask objective questions rather than simple multiple choice options. This was a deliberate choice, with the intention of engaging students more. The paper-versions also used a variety of student response mechanisms for the objective questions, for example joining lines and matching exercises. In terms of marking reliability, providing interesting answering mechanisms will necessarily decrease reliability, joined lines will be missed, correct responses not seen or rubric errors. These errors did not happen in the on-screen versions.

In terms of the marking of short and longer open responses, the situation was a bit different. In the on-screen versions, these responses were not automatically marked; they were captured by the platform, and then marked on-screen by expert markers. There were two marking mechanisms available for the on-screen versions. The marker could go into the test of a student and mark the response from the page presented to that student (which would also include any other responses on that screen). Alternatively, the marker could go into a spreadsheet of all answers for a particular item, and then mark all responses for a particular question. A sample of this spreadsheet is shown in Figure 116 below.

Figure 116: Screenshot of a marking spreadsheet



J872

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	usernam	test	iOrd	iName	pR	pa	marl	tot	recall	attempt	assess	lev	date
2	30965	ak116	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
3	31030	ak114	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
4	31103	ak117	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
5	31147	ak104	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
6	31185	ak111	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
7	31298	ak108	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
8	31357	ak107	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
9	31428	ak130	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
10	31493	ak101	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
11	31555	ak106	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
12	31623	ak112	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
13	31688	ak115	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
14	31753	ak113	test_1	1	sc_skeleton(a)	A	A	2	2	A:skull, B:rib, C:hip, D:ankle, E:knee, F:backbone, G:shoulder,	true	2/2e	4	27-03-2008
15	30990	ak116	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
16	31055	ak114	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
17	31128	ak117	test_1	11	sc_hearing(a)	A	A	0	2	A:boys_tin, B:string, C:boys_ear, D:girls_vocal, E:girls_tin,	true	4/3g	4	27-03-2008
18	31202	ak104	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
19	31237	ak111	test_1	11	sc_hearing(a)	A	A	0	2	A:boys_tin, B:boys_ear, C:string, D:girls_vocal, E:girls_tin,	true	4/3g	4	27-03-2008
20	31323	ak108	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
21	31388	ak107	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
22	31453	ak130	test_1	11	sc_hearing(a)	A	A	0	2	unattempted	false	4/3g	4	27-03-2008
23	31518	ak101	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
24	31583	ak106	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
25	31648	ak112	test_1	11	sc_hearing(a)	A	A	1	2	A:girls_tin, B:girls_vocal, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008
26	31713	ak115	test_1	11	sc_hearing(a)	A	A	2	2	A:girls_vocal, B:girls_tin, C:string, D:boys_tin, E:boys_ear,	true	4/3g	4	27-03-2008

The spreadsheet could also be filtered in any number of ways. For example, once all the responses were initially marked, all the credited responses could be filtered together and reviewed; likewise the incorrect responses could be put together and compared. This enabled the marker to double check the marking in a very quick and efficient manner. Also, because all the marks were available on the administration site, double marking by another marker was also very fast and efficient. Any marking differences between markers were recorded and reviewed.

Marking of paper-based versions was done using traditional practices. Scripts were marked by one marker, marks recorded on the pages and on the front of the script. These scripts were then sent to the second marker, who marked them again. Any differences were recorded, reviewed and resolved.

Once the papers were marked, they were sent to a data collection agency who keyed in the item and whole test scores. Reliability figures for accurate data transcription is high, however there is always some error associated with human transcription from script to spreadsheet.

There may have been issues concerning students using an on-screen mode to write open responses in terms of familiarity and modal differences in the style and content of writing, however the on-screen marking systems discussed above did aid marking reliability compared to the paper-based marking systems.

10.4 Validity Issues

Chapter 2 discussed various forms of validity before subsuming all of them with the component parts of reliability into a unified view of validity.

In terms of component validity types, this comparability study raised interesting issues when considering in particular face and construct validity.

There was no doubt that in terms of face validity the on-screen tests rated quite highly. Not only did students and teachers find the assessments stimulating and engaging, they also considered that the on-screen versions assessed science in a more appropriate manner than on paper. Not surprisingly, the very things that contributed to increases in face validity were also the aspects that were also considered as having greater construct validity, for example the use of video demonstrating science in action and the use of colour images. It was interesting, however, that many of the modal differences that contributed to an increase in face and construct validity were the ones that resulted in differences in performance, generally in favour of the paper-based versions.

On the other hand, there were a number of construct irrelevant variables that also seemed to affect student performance; for example the way students have to access information, draw circuits or even write open responses. Generally, these also seemed to discriminate against the on-screen versions.

Therefore there seems to be two variables at play that contributed to significant differences in performance between non-standard MCQ items in paper-based and on-screen modes; the variables being construct relevant and irrelevant factors. The construct relevant factors are the aspects that seem to be assessing some parts of the construct of science in differing ways to paper; those which it could be suggested are more authentically representative of the construct. In terms of comparability of exams, there are two key questions to consider when offering assessments in differing modes; is it acceptable to purposely assess different aspects of a construct, and if so, to what extent; and how are any differences quantifiable and therefore accounted for in awarding procedures?

The second variable is construct irrelevant factors. The challenges to these are different. There were clearly familiarity and logistical reasons in this comparability study why students performed differently in paper-based and on-screen modes. It could be argued (Heppell et al, 2004; Tattersall, 2009) that as more computer technology is used in the teaching, learning and assessment of science, many modal differences will reduce to a point where there will not be any construct irrelevant factors at play. The question in the short and medium term therefore is whether we should or could account for construct irrelevant factors which are probably variable across cohorts and not stable over time.

In Chapter 2, after the separate and unified approaches to reliability and validity were described, a broader view of threats to validity was explored using Messicks (1980) matrix and then Crooks et al (1996) eight staged linked model.

In the light of the results and their analysis in this comparability study, I will revisit the Crooks et al model and compare whether modes of assessment might influence any of the validity threats identified at the eight identified stages.

10.5 Revisiting Crooks et al framework for validity

10.5.1 Administration

This link is concerned with conditions affecting the students and the assessment itself; low motivation, anxiety, inappropriate assessment conditions in terms of the environment and the assessment instrument.

The evidence gained from the comparability study would suggest that for most of these issues, an on-screen mode offers advantages to paper.

Students considered the on-screen test versions more engaging, stimulating and less anxiety inducing than their paper counterparts. Clearly, there was a novelty value attached to the on-screen assessments; the students were not used to taking science tests in these forms. Whether they would retain their high motivational rating when they become the norm and standard practice remains to be seen, but it would be hoped that assessments that represent the way students engage and learn a subject would help to combat this particular threat to validity. Although both test modes were presented in highly controlled assessment conditions, they were not high stakes in terms of their outcomes and uses. Students often present different attitudes in low and high stakes conditions. However, considering the evidence obtained in this research, the on-screen versions did seem to reduce fear and anxiety compared to paper tests.

On-screen assessments do help to reduce the variables concerned with assessment conditions to an extent, in that the on-screen environment and instructions will be common to all. Time allocations for on-screen exams also can be tightly controlled. However, taking examinations in computer suites rather than the usual examination locations can raise issues of space and equipment allocation. There are also issues regarding the security and possible malpractice of on-screen tests in computer suites. This was a commonly held concern among students and teachers involved in this comparative study.

The last threat mentioned in this link is perhaps the most significant. Does the assessment (in this case, a modal question) result in a lack of accessibility or fairness in terms of the construct or the testing condition? This is not a simple question and does not provide straightforward answers.

It could be argued that the computer-based assessments allowed students access to aspects of science that the paper-based forms could not. This was a view taken by the majority of students and teachers in the trial. Therefore not only can computer versions generate more authentic scientific assessments, but they can also allow facilities such as font adjustments and oral presentations of questions on a large scale not possible in paper-based forms. However, there are equally significant accessibility issues associated with the on-screen versions. It is clear from these trials that many students were disadvantaged when taking assessments in an on-screen mode. Some of these disadvantages can be attributed to unfamiliarity issues, which will reduce in time, however it unlikely that they will all dissipate across school populations consisting of students from very different social, economic and cultural conditions. There is no doubt that if high stakes assessments move towards computerisation, there will need to be paper-based versions made available by choice or exception for some time. It is equally clear however, that assessments across modes are not the same if they move out of a simple MCQ (multiple choice question) format. The question remains whether actual or weighted statistical equivalence will be enough or acceptable to account for these differences, or are parallel assessments that actually test some different things are simply inequitable?

10.5.2 Scoring

This link is concerned with errors associated with marking; they include restrictions of creditworthy responses through the application of narrow mark schemes, overemphasis of some skills over others, inter and intra marking reliability and marking being either too atomistic or holistic.

This particular comparability study only really addressed one of these threats; that of inter and intra marking reliability. As discussed earlier, one of the advantages of the automated marking of objective questions is that responses are not missed, rubrics not applied or marks incorrectly transcribed. These might seem trivial matters, but they contribute to high stakes marking unreliability error more than most people think. The difference between basic MCQ (multiple choice questions) and objective questions can be found mainly in the response mechanisms. MCQ usually involve selection of an answer from four alternatives. Objective questions still consist of a correct response, however the answering mechanisms can be far more complex; eg. joining boxes together with lines or matching exercises. Although the marking rubrics are relatively straightforward, these types of questions lend themselves to inaccuracy when human marked, but complete reliability in an on-screen environment.

As described previously, most of the items in these science tests were automatically marked and therefore were constructed with a fairly narrow interpretation of skill sets.

The mark schemes for the paper and on-screen versions were identical for the closed and open responses, and therefore although some of the closed responses were perhaps a reflection of narrow skill sets, the open responses gave the opportunity for students to demonstrate a wider range of cognitive skills. As described previously, the on-screen marking mechanisms enabled these open-responses to have a high level of marking reliability, however still enabling expert human judgement to apply.

The investigations were designed to assess process rather than content skills. The on-screen versions combined automatically marked items alongside open-ended expert marked items. In particular, the automatic marking of plotted graph points and lines of best fit had significant reliability advantages compared to marking these aspects by hand.

10.5.3 Aggregation

This link explores aggregating scores which may distort any intended outcome or purpose to an assessment.

It may appear that this link is unaffected by presenting assessments in different modes, as the assessments did not differ in terms of the number of questions or the time allowed. However there are a few issues to consider, as the design of these assessments were built around on-screen delivery. As discussed previously, there are various drivers involved in a large scale move towards on-screen summative assessments; these include efficiency in terms of time and money, faster feedback to schools and students and the potential to assess facets of subject or cognitive constructs not possible through paper-based assessments. The on-screen tests used in this comparability study were made up in large part by objective questions which then were automatically marked. These types of questions are clearly faster and cheaper to mark than open-ended responses. They also considerably speed up the marking process, which is a desired outcome of on-screen assessment. However, there must be a question about whether these question types are generalisable across a whole subject construct, whether they would need to be supplemented by open-ended components and if so, what weighting might need to be applied to components?

This position might be more of an issue if the only form of test was basic multiple choice. In this comparability study, there were objective and open ended question types used. It could be argued that, if well designed, on-screen assessments of the nature of this study can avoid aggregation issues as they can assess a broader range of a construct and therefore be more representative of a student's knowledge and understanding. The inclusion of open-ended questions in an on-screen test will necessarily slow down the return of results if human expert marking systems are used; however, this might be the necessary compromise to overcome aggregation issues in terms of assessing the broadest range of cognitive skills in the most efficient manner.

10.5.4 Generalisation

This link explores the reliability of student performance in terms of whether the assessment is a dependable measure of a construct. One of the identified threats for this link was the issue of a ‘one-shot’ testing culture; that many tests and exams are largely dependent on the contents or contexts of a particular exam, at a fixed time and place which all may or may not suit the preparations and condition of the student at that particular time.

Despite movements towards raising the profile of teacher assessment or continuous assessment methodologies, high stakes assessments in schools have remained firmly in the domain of externally set and externally marked exams. This is unlikely to change in the foreseeable future, but what might change are the production, delivery and feedback mechanisms for these exams.

The production, delivery and marking methodologies of high-stakes, paper-based exams have hardly changed over the last fifty years. Individuals write papers, usually sampled from a specification, they are reviewed by a small number of people, finalised, printed and delivered across the country to schools and colleges. Exams are taken on a particular day at a particular time by students, collected and sent to hundreds or thousands of markers who are trained to mark that particular paper. Grade boundaries are then subsequently determined, disseminated and applied. These systems evolved in the light of the historical conditions with respect to high stakes assessments; fewer students took them, they were taken at one fixed time of the year (usually in the summer), high security had to be applied to avoid malpractice and feedback was only provided in the form of grades, not providing diagnostic or formative feedback. There were no reliability measurements applied within or across exam series and there was little interest in validity issues. More than anything, they were used as predictive measures of future performance.

The local conditions in terms of high stakes assessments are very different now, and there is an emergent technology to support it. There are far more students in the high stakes exam system, more exams need to be available more often, and they therefore need to be more consistent in their design and more reliable in their performance. Feedback of performance needs to be faster and diagnostic feedback highlighting strengths and weaknesses is now desired by teachers and students.

Although these issues may not seem to address the issue of generalisation, they do in terms of systemisation. If (or when) computers are used to support the changing requirements of the assessment system, they will be underpinned by psychometrics in a way that does not happen now. Items and tests can be tagged and constructed to assess constructs with more consistency within given reliability measures, and there will be less pressure on the system and students at fixed points of time, as assessments will be more readily available around the year in unitary and modular formats.

10.5.5 Extrapolation

This link explores how effectively an assessment samples a construct. There are two issues to consider in relation to how on-screen assessments affect this link. One is how coverage of the construct is established and then secondly how differing question types can support assessment. It has already been described how more systemisation in the design, construction and delivery of computer-based tests can result in more construct representation than paper-based systems. Computerisation alone will not solve issues such as bias or individual item validity, however there can be more assurance that more areas of a curriculum can be assessed through objective question types rather than simple MCQ, and assessment weighting and emphasis can be consistently applied within and across tests using effective tagging methodologies.

On the other hand, there is concern that only assessing a construct through objective questioning will not provide opportunities for assessing all essential elements of a domain. This might include higher order thinking and evaluative skills and also open response communication. At the present time in England, regulatory qualification and subject criteria do not allow the assessment of high stakes assessments to consist only of objective questions. Therefore, if on-screen testing is to be used in high stakes assessments, either additional paper-based components would need to be included or the on-screen assessments would need to be capable of assessing a construct in other ways than objective questions.

The results of this comparability study have shown that this can be done. Not only can on-screen assessments assess the range of question types that paper can, but the on-screen assessments allow students to interact with items and contexts in ways that are not possible on paper. The question therefore returns to equivalence issues; are the results from differing modes equivalent, and are they testing the same things?

Regulatory equivalence will necessarily restrict the movement towards on-screen assessments, and particularly through the three generations described by Bennett (1998) for as long as there has to be a choice of testing mode in high stakes assessments. This need not apply for formative and diagnostic assessments, however there will not be major breakthroughs in the design and delivery of high stakes summative assessments until this regulatory equivalence issue is resolved.

10.5.6 Evaluation

This link explores how information from assessments are interpreted and used. The use of on-screen assessments, through their design and construction may well enable stakeholders to be given a richer analysis of performance than currently available through paper-based systems. Total scores and grades can be broken down and feedback given on any number of criteria, from performance across content areas of a specification to performance of particular cognitive skills (for example, how well a student has demonstrated cognitive levels of demand). Effective tagging of items provides opportunities for on-screen assessments to be used for more than just a graded measure of overall competence.

The issue of exam data being taken at face value and the lack of interrogation of how well or reliably an exam assesses a construct will continue to be an issue as long as high stakes assessment continue in externally set, externally marked methodologies.

It might be hoped however that as on-screen assessments become mainstream and regulatory issues resolved, there will be more opportunity for assessments to have more construct validity in terms of the types of tasks students can carry out, the types of process skills they can demonstrate and the type of evidence gained from them. If this does happen, there should be less 'teaching to the test' approaches where only the aspects of a construct that can be easily measured are assessed and therefore subsequently taught.

10.5.7 Decisions

This link explores the outcomes of assessments in terms of 'standards' of achievement gained and also pedagogic actions taken on the basis of assessments. As discussed previously, on-screen assessments can take a range of forms, from basic MCQ to simulated environments. Setting construct standards on the basis of MCQ questions might not lead to improved confidence of standards in comparison to the current status quo. It would be unlikely that all the essential elements of a construct could be assessed using MCQ and objective questions. If MCQ and objective questions alone were used, any standard setting procedures would only be able to use statistical techniques, as there would not be any judgemental identifiers of performance. In addition, even if it could be argued that objective testing could assess a wide range of cognitive skills, this type of assessment could easily distort the way teaching and learning was applied. Standards would no doubt rise, in the manner described by Black, (1998); Wiliam, (2007); and Tymms, (2004), and the Lake Wobegon effect and Goodhart's Law could be readily applied.

Conversely however, there could be positive influences on standard setting through on-screen assessments. As described previously, one simple advantage of using computerised assessment systems is that they count correctly, do not miss out marks, and transfer marks without error. These systematic errors can affect standard setting decisions and the outcomes for students if they fall short of a specified standard through clerical error rather than through performance. In addition, if on-screen assessments include a range of response types, and can assess valued areas of a construct, then there might be enhanced confidence in the meaning of 'standards'. In addition, because enhanced forms of process skills can be assessed using on-screen environments there would be less risk of a Lake Wobegon effect or Goodhart's Law dominating the assessment agenda in terms of exams being centred on very narrow prescribed areas of learning.

There still however remains the recurring question concerning equivalence between on-screen and paper-based assessments, and the implications differing modes have on the setting of standards. If there are performance differences between modes, and perhaps even different aspects of a subject construct being assessed between modes, on what basis can a single standard setting procedure be applied?

10.5.8 Impact

This final link explores the consequences of assessments. The central question related to this comparability study is whether on-screen assessments can offer any positive advantages to the consequences of assessments compared to the current status quo. Do on-screen assessments offer greater dependability to support the purposes of assessment?

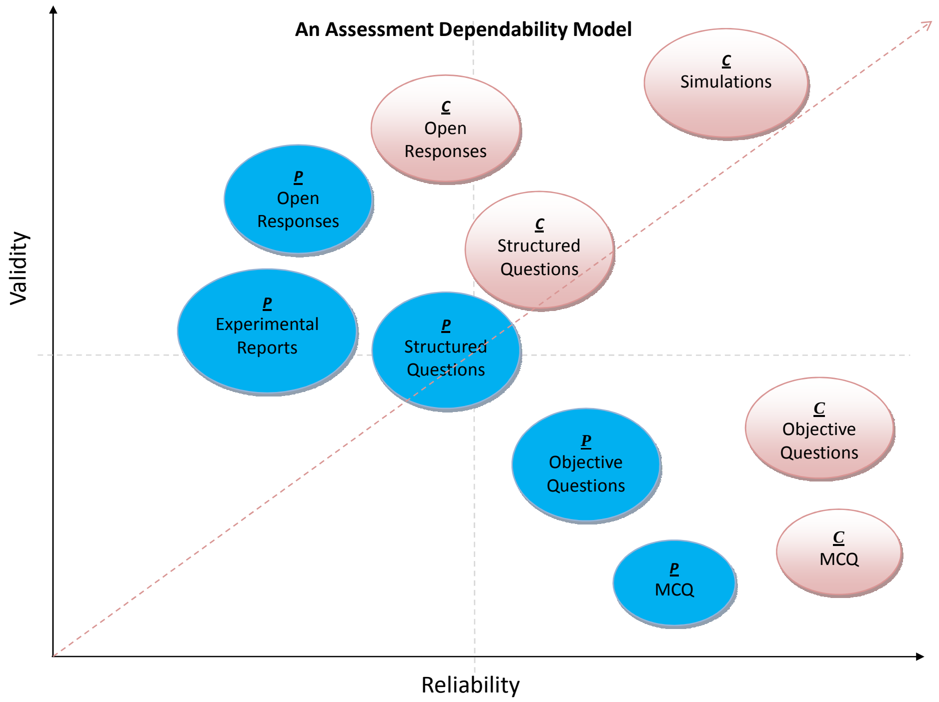
Newton (2007) described the three purposes of assessment as essentially the generation of data or evidence, decisions taken on the basis of that evidence and then the impact on teaching and learning.

This comparability study has demonstrated that on-screen assessment can offer enhancement to the face and construct validity of assessments. They can also offer enhanced reliability measures, particularly in relation to inter and intra marking. They can offer enhance motivation and engagement of students and rich feedback of the performance of a construct, including aspects not possible on paper. The positive benefits that on-screen assessment can offer might seem to suggest that these types of assessments are not only more dependable, but they also better support the purposes of assessment. However, there are significant hurdles to overcome before a potentially enhanced assessment system can be applied efficiently and fairly for all stakeholders. These key considerations will be discussed in the following concluding chapter.

10.6 Comparative dependability

In summary of this chapter, a model of the relationship between validity and reliability in on-screen and paper-based modes is offered. A visual model is shown below in Figure 117, and then described.

Figure 117: A Model to Compare Modal Dependability



Key

P Paper-based items

C On-screen items

The model in Figure 116 above is a representation of how the different science assessment types in different modes in this study could be compared and contrasted using reliability and validity as key variables.

The matrix represents low to high values of reliability and validity on the axes. The matrix also shows an ideal theoretical direction of travel through the origin, resultant at the top right quartile where an assessment could be considered to have both very high reliability and validity values. This would be an ideal assessment in terms of a unified approach to validity. In Chapter 2, the concept of a unified approach was discussed, using the term dependability. Dependability attempts to consider the relative reliability and validity of an assessment to suit its methodology and intended outcome. The plotted points on the model above could be considered to determine the dependability of an assessment, and then compare it with suggested alternatives.

The challenge for on-screen assessments should be that they achieve a higher dependability on this matrix when compared to their paper-based alternatives.

The model above compares the five parallel assessment types used in the paper and computer based versions of this research study. Their comparative positions could be open to discussion and differences of opinion, however I have placed them as shown and use a brief summary of the quantitative and qualitative evidence from this research study to justify their positions.

10.6.1 Basic multiple choice questions (MCQ)

Paper and on-screen versions of MCQ's are shown to have a similar level of dependability. In particular they are considered to have higher reliability than validity. This is not to suggest that they should be considered as unworthy assessment items, however, their purpose generally in high stakes assessments are to assess the lower cognitive levels. The on-screen versions are rated at slightly higher levels of reliability as they can be marked without error and slightly higher levels of validity as they can contain a greater variety of stimuli, for example video or animations.

10.6.2 Objective questions

In this category, the gap is wider along the reliability axis than the validity axis. As described in chapter 10, there can be high marking unreliability in paper based objective questions depending on the complexity of the answering mechanism or the mark scheme rubric. Non MCQ objective questions can be engaging and motivating for students, however they do contribute to marker error. The on-screen versions carry no risk to marking error. These question types were considered to have a higher validity rating than the MCQ's due to their higher levels of engagement, however the on-screen versions rated slightly higher on validity in the same way as MCQ's; they were able to utilise more authentic stimuli.

10.6.3 Structured questions

The gap between the paper and on-screen versions widens further for these questions. A combination of automatic and marker assisted technologies in the on-screen versions gives them higher levels of reliability. The available range of stimuli and response mechanisms gives them a higher validity level than the paper-based versions.

10.6.4 Open-response questions

The position of these items is probably the category that would be most open to differences of opinion. The on-screen versions were rated significantly more reliable than the paper-based versions due to the on-screen marking facilities that allowed markers to mark, review and amend open responses in a fast and efficient manner. The on-screen versions were given a slightly higher validity rating than the paper versions due to the variety of stimuli that were used to support them. However this item type was one of those highlighted as having construct irrelevant issues in terms of some students feeling inhibited by the response medium. Therefore in this study, higher validity was not applied for all students.

10.6.5 Investigational simulations/ reports

This category placed the on-screen and paper-based versions the furthest distance apart on the matrix model. The onscreen versions had high levels of reliability as they consisted of a combination of automatic and marker assisted marking technologies. The on-screen versions also were considered to have very high levels of validity as they enabled students to engage in scientific enquiry in an authentic manner and enabled a combination of content, process and cognitive skills to be assessed. The combination of high levels of reliability and validity for the on-screen investigations gave them the highest dependability rating of all the trialled item types. The paper alternatives that assess enquiry skills can have low reliability values due to the inherent error associated in marking graphical outputs. Their validity is also only considered to be moderate as they cannot engage students with an authentic investigational experience, they usually are restricted to using secondary data, and the range of skills that can be assessed is much more limited than the on-screen versions.

The assessment dependability model could be a useful tool when designing or reviewing assessments in any particular mode, however it would also be useful to compare and contrast any potential benefits or issues when changing assessment modes and item types are considered. While a high dependability is considered to be a desirable attribute, an assessment positioned at any position on this matrix could be justified if it has a clear purpose and rationale.

10.7 Summary

This chapter has discussed the three key themes of this comparative study, those of the equivalence and comparative reliability and validity of science assessments presented in paper and computer-based modes. All of these themes have then been incorporated into a unified view of comparative validity, using the model developed by Crooks et al (1996), which has also been used in Chapter 2 to discuss the threats to validity of existing paper-based assessment systems.

Finally I proposed a model to compare the dependability of assessments in different modes, which may offer a simple way of determining their appropriateness.

In the following final chapter I will review and evaluate my research findings.

Chapter 11

Conclusion

11.1 Introduction

My research has centred on changing assessment practices resulting from the shift from paper-based to on-screen assessment in schools. This research is a contribution to the development of an informed understanding on the empirical and interpretive issues to be considered as school high stakes assessments become computer based. This includes the potential impact on the attitudes and performance of school students when assessment modes change.

A comparability study was set up in the context of Year 9 and 10 science education in England to produce equivalent paper and computer based tests and investigations. These tests and investigations were trialled by 1000 students and the quantitative data from the marked assessments was then analysed alongside qualitative data and evidence collected from the students who took part in the trial and their teachers.

The three key areas of interest were to establish whether the same assessments in different modes had scoring equivalence and to compare and contrast the reliability and the validity of the assessments in their paper and computer versions.

Before these issues are discussed, it needs to be acknowledged that this research is an ecological study. It was set up in a particular subject area for a particular school age group using particular styles of assessments. While I intend that the findings are generalisable for future on-screen assessment consideration, the literature review on computer based testing in Chapter 3 emphasised that any individual findings should be treated with caution in terms of any suggested outcomes. However, this controlled research study has provided a body of evidence to contribute towards an understanding of the quantitative and qualitative challenges facing stakeholders as national high stakes assessments move towards on-screen delivery over the next ten years (Ofqual, 2009).

There is little doubt that this change will happen in England at least, for cost, efficiency and assessment related reasons. After Ken Boston's initial blueprint proposals back in 2005 and 2007, there has been slow progress in the shift towards changing assessment modes. However 2009 saw a concerted interest by regulators and awarding bodies to re-engage with e-assessment and pave the way towards change in terms of establishing an initial consensus statement on the direction of travel for on-screen assessments (see QCA consensus statement, 2009). Government and exam boards talk about 'when' not 'if' in consideration of large scale on-screen assessment change and not just for automatically marking some item types but marking all student work (Oates, 2009).

In terms of my key areas of research interest, I will present the conclusions of my research study.

11.2 Equivalence

The quantitative outcomes of this research study established that there was not equivalence between the scores of the paper and computer based tests. This was not an unexpected result as the tests consisted of a range of item types, some similar in each mode and others quite different in the stimuli or response mechanisms employed. The paper based tests were calculated to have a 2 mark advantage compared to the on-screen versions. This calculation was established once differences in the group profiles had been taken into account. While this 2 mark difference was statistically significant, it was half of the calculated standard error of measurement (SEM) for each test. This indicates that the scale of this issue was actually far less than the in-built error associated with the test. Equivalence could not be calculated for the investigations as they contained too few items and they had unequal mark allocations in each mode. Issues concerning equivalence are discussed in Chapter 3, the quantitative results are shown in Chapter 7 and the equivalence outcomes of this research study are discussed in Chapter 10.

Any differences in scores can be attributed to two variables: construct relevant and construct irrelevant factors. Construct relevant factors consist of the assessment of differing aspects of a construct and irrelevant factors are the performance differences caused by unfamiliarity or unfairness in the mode of delivery. It can sometimes however be difficult to distinguish between them. The theory of construct relevance and irrelevance is discussed in Chapter 3 and the discussion of these factors in relation to the outcomes of the tests and investigations used in this study are discussed in Chapters 9 and 10.

11.3 Reliability

The internal reliability was calculated for the tests in paper and on-screen modes, using Cronbach's alpha. The calculated figures for the tests in both modes indicated a high level of internal reliability; meaning that they each appeared to be assessing the construct of science, and establishing a reliable rank order of student performance. This does not mean that the tests in each mode were necessarily testing the same aspects of the construct, however the tests in either mode could be used in high stakes assessments in terms of their construction and performance. Cronbach’s alpha was also calculated for the investigations, achieving a low level of internal reliability. This however was indicative of too few items in the assessments for Cronbach’s alpha to be used as an appropriate or effective measure.

The reliability of marking across the two modes was not empirically studied, however the on-screen versions had two particular advantages compared to the paper versions. Approximately 75% of the on-screen tests and investigations were automatically marked, and therefore not liable to have any associated marking error. In addition, the marking facilities for the open-ended, on-screen questions reduced marker error by providing efficient filtering and marking review mechanisms.

The results of the quantitative research are shown in Chapter 7 and analysed alongside the qualitative evidence in Chapter 8.

11.4 Validity

Unlike reliability, validity cannot be empirically determined. Component aspects of validity are discussed in Chapter 2. Evidence concerning the validity of the science tests and investigations in this research study came from the qualitative data collected from the 1000 students who took in the trial and a sample of their teachers. This evidence is shown in Chapter 8, and analysed alongside the quantitative data in Chapter 9. In general, the on-screen tests and investigations were rated more highly than the paper-based versions in terms of being more engaging, authentic, fit for purpose and less stressful for students. Although some of this preference may be attributed to the novelty of the on-screen assessments, much of the collected evidence from students and teachers concerned the enhanced construct validity contained in the on-screen versions. These views applied equally to the tests and the investigations. The preferred mode in terms of the perceived validity of the assessments were the on-screen versions as stated. However, this preference was counterbalanced by students performing less well in the on-screen versions, and there was some evidence presented in Chapters 7 and 9 to suggest that students of lower ability were disadvantaged more than the more able students.

11.5 Dependability

Dependability, similarly to validity, has no performance measures associated with it. It is best described as the effective trade-off between reliability and validity in order to achieve the most fit for purpose assessment. Ideally, assessments would have the highest measurable reliability and the highest rated validity, however in practice, a compromise is usually arrived at to fulfil the regulatory requirements of externally set, externally marked high stakes assessments with the most valid method of their assessment. The resulting dependability will therefore be based on the qualification type, the subject area and the style of assessment used. Chapter 10 presented a model to compare the reliability and validities of differing assessment item types in different modes in this research study. This could be said to be a measure of dependability. Using this model, the on-screen item types achieved a higher dependability rating than the paper versions.

In conclusion to my research study, I will address a few unifying themes that concern the movement of assessments from paper to on-screen modes of construction, design and delivery.

11.6 Assessment Purposes

Chapter 2 discussed the various purposes of assessment, and how any intended or unintended consequential purposes placed upon the outcome can affect the validity and fitness for purpose of the assessment, regardless of its initial intention. Newton (2007) outlined three main purposes of assessment; generating results, enabling decisions as a result of assessments, and impacting on teaching and learning.

The assessments developed in this research study were designed for use in a high stakes environment; that is, assessments usually used for selection or entry requirements (eg. GCSE's). If the three purposes outlined above are considered, what difference would a movement towards on-screen assessment make?

11.6.1 Generating results

It might be assumed that this purpose would be unaffected by changes in assessment mode. However, as described in Chapters 9 and 10, higher marking reliability alone would result in the generation of more reliable results for students, and therefore there would be more confidence in the assessment outcomes. In addition, if content and process skill elements of a construct are better represented using on-screen item types, then there will also be greater confidence in the generated result.

11.6.2 Enabling decisions

This purpose follows on from the one previously discussed. If on-screen assessments can assess a more complete range of a subject construct, incorporating many aspects that paper assessments cannot, then any entry decision or predictive indictors arising from that assessment will be more valid in terms of the assessment coverage and outcomes.

11.6.3 Impact on teaching and learning

There is the potential for this purpose to be most affected by movements towards on-screen assessments. If assessment item enhancements can be incorporated into high stakes assessments, there may be less opportunity or incentive to 'teach to the test' and therefore teaching and learning could be less distorted by the external assessment system. In addition to this, the opportunity to offer diagnostic feedback is built into on-screen technologies. As all items are tagged, teachers and students can interrogate their performance in various ways after results are returned. This can enable students to review and improve on their performance if required and it allows teachers to review their teaching approaches and strategies for current or future individuals or classes.

11.7 Principles, practices and paradigms of assessment

If high stakes assessments do move towards on-screen modes, the principles of assessment construction and design will have to change as well as the modes of delivery and marking. This might entail changes in the regulatory subject criteria and the codes of practice in terms of permissible assessment item types and the way in which items and tests are authored, standardised and awarded.

The test and investigation items designed and used in this research study were an attempt to assess a broad range of content and process skills using a variety of stimuli and response types. The qualitative feedback from students and teachers indicated that the variety of assessment approaches were important features, not only in terms of the effective assessment of the construct, but also the ability to actively engage and interest the students.

Wiliam (2008) suggested that MCQ tests and items, if well designed, can be very effective assessment instruments as they can allow a range of cognitive levels to be assessed, and enable a wide range of a construct to be assessed in relatively short assessments. If high stakes assessments do move into on-screen automated modes, it is most important that MCQ and objective questions are constructed and designed to a high level of quality as a matter of course and not by exception. This will entail significant training and development for existing assessment writers in order to ensure that the assessment dependability is supported by high construct validity as well as high reliability measures, particularly in construct areas that have not normally been assessed through paper modes.

It would be hoped that assessment practices in schools would also change in the light of the potential that on-screen assessments can offer. There is the opportunity to remove high stakes assessments, and their associated teaching and learning strategies from a narrow interpretation of a construct to a more holistic representation of the intended content and skills required from a specification.

As assessments move towards different on-screen item types which can assess areas of constructs not possible on paper, can this result in a paradigm shift in terms of assessment? Many of the conditions discussed by Kuhn (1970) seem to apply when the changing assessment needs and technologies combine to enable the assessment of different constructs in new ways. Paper-based assessments have existed for hundreds of years; however no matter how hard and concerted the effort, they cannot assess constructs in a dynamic manner. The integration of a different medium to receive and transmit information and evidence has the potential to view assessment and measurement systems in radically different ways, not just in terms of constructs, but also the relationship between the assessor and the student.

It could be argued that as long as school-based, high stakes on-screen assessments in England are still firmly routed in first generation usage (Bennett, 1998), there will be no effective change to any assessment paradigm. However if and when 3rd generation on-screen usage becomes standard assessment practice, new assessment paradigms may emerge.

My initial research questions focused on the comparative reliability and validity of science assessments presented in different modes. My research outcomes to these questions are positive. High internal test reliability can be achieved in either mode, and probably higher overall reliability can be achieved in computer-based assessments through more effective marking interfaces. There is also the potential that computer-based assessments can be more authentic, face and construct valid, as long as they are designed to include valued aspects of a construct not possible in paper-based versions. My research was based in the subject area of science, however many of the outcomes are generalisable across subject areas.

In terms of the three personal professional perspectives I gave in Chapter 1; those of a teacher, a governmental regulator and now working at an exam board, do my research outcomes present any conflicts of interests? I do not think that they do. There is considerable research and development required to the operational systems, pedagogy and assessment instruments before computer-based assessments will be available and equitable for all. However, if managed with care, there is little to fear and much to gain in the shift towards on-screen assessment in schools.

References

Aiken. L.R. & Groth-Marnet, G. (2006). *Psychological Testing and Assessment.* Pearson Education.

Akpan, J.P. (2001). Issues Associated with Inserting Computer Simulations into Biology Instruction: A Review of the Literature. *Electronic Journal of Science Education,* Vol 5, (3), March 2001.

Al-Gahtani, S.S. & King, M. (1999). Attitudes, satisfaction and usage: factors contributing to each in the acceptance of information technology. *Behaviour and Information Technology*, 18(4), pp. 277-297.

Anastasi, A. (1990). *Psychological Testing* (6 edn). New York: Macmillan.

Andre,T. and Haselhuhn, C. (1995). Mission Newton! Using a computer game that simulates motion in Newtonian space before or after formal instruction in mechanics. *Paper presented at the American Educational Research Association Annual Meeting*, April 1995.

Andrich, D. (2004). Controversy and the Rasch model: a characteristic of incompatible paradigms? *Medical Care,* 42, pp.1-16.

APU (Assessment of Performance Unit) (1984). Science at ages 11, 13 and 15- *Three reviews of APU surveys* 1980-1984. London HMSO.

Bailey, K.D. (1967). *Methods of Social Research.* The Free Press. London: Collier Macmillan.

Baird, J. & Mac, Q. (1999). *How should examiner adjustments be calculated? -* A discussion paper. AEB Research Report, RC13.

Baird, J., Greatorex, J. & Bell, J.F. (2002). *What makes marking reliable? Experiments with UK examinations.* AQA Research Report, RC191.

Barthlomew, D.J. (2000). The measurement of Standards. In Goldstein, and Heath, A. *Educational Standards*, The British Academy, pp. 39-56.

Bartram, D. (1990). Reliabilty and Validity. In Beech, J.R. and Harding, L. (eds) *Testing People: A practical guide to psychometrics.* NFER-NELSON.

Becta (2006) Research report: *Becta Review 2006*, Evidence on the progress in ICT in education.

Bennett, R.A. (2001). How the Internet Will Help Large-Scale Assessment Reinvent Itself. *Education Policy Analysis Archives*, 9, (5).

Bennett, R.A. (2002). Inexorable and inevitable: The continuing story of technology and assessment. *The Journal of Technology, Learning and Assessment*, 1,(1) pp.1-24.

Bennett, R.A. (1998). *Reinventing assessment: speculations on the future of large-scale educational testing.* Princeton, NJ: Educational Testing Service. Policy Information Center.

BERA. (2004). *Revised ethical guidelines for educational research.*

Bergstrom, B. *(1992). Ability Measure Equivalence of computer Adaptive and Pencil and Paper Tests: A Research Synthesis.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Black, P. and Wiliam, D. (1998). Inside the black box: raising standards through classroom assessment. London, Kings College.

Black, P. and Wiliam, D. (2002). *Standards in Public Examinations.* London: King's College, School of Education.

Black, P. & Wiliam D. (2006). The Reliability of Assessments, in Gardner, J. (eds) *Assessment and Learning,* London: Sage, pp.119-132.

Black, P. (1998). *Testing: Friend or Foe? The Theory and Practice of Assessment and Testing.* London: Falmer Press.

Black, P. (2003). Testing, testing, testing: listening to the past and looking to the future, *School Science Review,* 85, (311), pp. 69-77.

Blake, N. (1997). Research, Development and Tacit Capability in the Education System, *Cambridge Journal of Education, 27 (2),* pp. 223-234.

Blatchford, I.S. & Blatchford, J.S. (1997). Reflexivity, Social Justice and Educational Research, *Cambridge Journal of Education, 27 (2),* pp. 235- 248.

Blaxter, L.; Hughes, C.; and Tight, M. (1996). *How To Research,* Buckingham: Open University Press.

Bloom, B.S. (1956). *Taxonomy of Educational Objectives: the classification of educational goals*. Handbook: Cognitive Domain. New York: David McKay Co.

Blunkett, D. (2000). Influence or irrelevance? Can social science improve government? Speech to the Economic and Social Research Council, 2nd February 2000, reported in *Research Intelligence,* 71, pp 12-21.

Boston, K. (2005). Assessment, Reporting and Technology. System-wide assessment and reporting in the 21st century. Tenth annual roundtable conference, Strategy, technology and assessment.

Boston, K (2007a). Tipping points in education and skills. *Speech to QCA Annual Review* , 2006.

Boston, K. (2007b). *Evidence given at the DCSF Select Committee on Testing and Assessment.*

Brennan, R.L. (2001). An essay on the history and future of reliability from the perspective of replications. *Journal of Educational Measurement,* 38 (4), pp. 295-317.

Bridges, D. (1999) Educational research: pursuit of truth or flight into fancy?, *British Educational Research Journal,* 25(5), pp. 597-616.

Broadfoot, P. & Black, P. (2004). Redefining assessment? The first ten years of Assessment in Education. *Assessment in Education, 11* (1).

Broadfoot, P. (1996). Educational Measurement: the myth of measurement. *Education, Assessment & Society.* Buckingham: Open University Press.

Brooks, R., and Tough, S. (2006). Assessment and Testing: Making space for teaching and learning. Ippr. London.

Brosman, M.J. (1998). The impact of computer anxiety and self-efficacy upon performance. Journal of Computer Assisted learning, 14(3), pp. 223-234.

Bross, T.R. (1986). The microcomputer-based science laboratory. Journal of computers in mathematics and Science Teaching, 5. (3), pp. 16-18.

Brown, F. (1980). Perspectives on validity. *NCME Measurement News, 23,* pp. 3-4

Bruner, J. (1960). The process of education. New York: Random House.

Bryce, T.G.K. and Robertson,I.J. (1988). The singer, not the song: a response to "Beyond processes". Studies in Science Education, 15, pp. 135-143.

Bunderson, C.V., Inouye, D.K. and Olson, J.B. (1989). The four generations of computerized educational measurement. In R.L. Linn (Ed), Educational measurement (3rd ed., pp 367-407). London: Collier Macmillan.

Calderhead, J. (1981). Stimulated recall: a method for research on teaching. *British Journal of Educational Psychology* 51, pp. 211-217.

Cannell, C.F. & Kahn, R.L. (1967). *The Dynamics of Interviewing.* New York: Wiley.

Choi, S.W. & Tinker, T. (2002). Evaluating comparability of paper and computer-based assessment in a K-12 setting. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, LA.

Chua, S.L., Chen, D.T and Wong, A.F.L. (1999). Computer anxiety and its correlates: a meta-analysis. Computers in Human Behaviour, 15(5),pp. 609-623.

Clausan-May, T. (2001). *An Approach to Test Development.* Slough: NFER.

Clesham, R. (2004). Standards in science key stage tests: how the QCA sets the attainment levels each year, *School Science Review,* 85**,** 312.

Cohen, L., Manion, L. and Morrison, K. (2007). *Research Methods in Education,* 6th Edition. London: Routledge.

Cresswell, M. J. (1996). Defining, Setting and Maintaining Standards in Curriculum-embedded Examinations: Judgemental and Statistical Approaches. *Assessment: Problems, Developing and Statistical Issues.* Edited by H. Goldstein and T. Lewis. John Wiley & Sons Ltd.

Creswell, J.W. (1994). *Research Design: Qualitative and Quantitative Approaches*. Thousand Oaks, California.: Sage.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Wilson.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests, *Psychometrika,* 16, pp. 297-334.

Cronbach, L.J. (1988). Five perpectives on validity , In WAINER, H., & BRAUN, H. I., (eds). *Test Validity,* pp. 3-17, Hillsdale, NJ, Erlbaum.

Cronbach, L.J., and MEEHL, P.E. (1955). Construct validity in psychological tests, *Psychological Bulletin,* **52,** 4, pp. 281-302.

Crooks,T.J., Kane,M.T., & Cohen, A.S. (1996). Threats to the Valid Use of Assessments. *Assessment in Education, 3,(3).*

Denzin, N.K. (1978).*The Research act: A theoretical introduction to sociological methods (*2nd ed.). New York: McGraw-Hill.

Dewey, J. (1910). How We Think. Lexington, MA: D.C. Heath.

DIIA (2003). Test Item Analysis & Decision Making. The University of Texas at Austin. www.utexas.edu/academic/diia accessed December 2008.

Driver, R. (1983). The Pupil as Scientist? Open University Press.

Driver, R. and Bell, B. (1986). Students thinking and learning of science: A constructivist view. School Science Review, 67, pp. 443-456.

Driver,R., Leach,J., Millar,R., & Scott,P. (1996) *Young Peoples Images of Science.* Buckingham: Open University Press.

Duggan,S., Gott,R., Luben, F. and Millar,R. (1994). Evidence in Science Education; articles in journals, edited books and conference proceedings. *The PACKS project*. York, University of York.

Ebel, R.L., & Frisbie, D.A. (1986). *Essentials of Educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Edward, A. & Talbot, R. (1999). *The Hard-pressed researcher, A research handbook for the caring professions,* 2nd Edition.

Eisner, E. (1992) Objectivity in educational research, *Curriculum Inquiry,* 22, pp. 9-15.

Ericsson, K.A. and Simon, H.A. (1984). *Protocol Analysis: Verbal reports as data.* Cambridge, MA: MIT. Press.

Evans, K. (1985). The Development and Structure of the English School System. London: Hodder and Stoughton.

Foucault, M. (1957). La recherché scientifique et la psychologie. In *Ditset Ecrits Vol1.* Paris : Gallimand.

Foucault, M. (1977-78). Security, territory, population, Lectures at the College of France. Paris: Gallimand.

Foucault, M. (1981-82). The Hermeneutics of the subject, Lectures at the College of France. Paris: Gallimand.

Gagne, R.M., Wager, W. and Rojas,A.(1981). Planning and authoring computer-assisted instruction lessons. Educational Technology, 21(9): pp. 17-26.

Galton, F. (1884). *Heredity genius.* New York: Appleton. Cited in Goldstein, H. (1994). Recontextualizing Mental Measurement. *Educational Measurement: Issues and Practice, 13, 1*

Gardner, H. (1992). Assessment in Context: The Alternative to Standardized Testing, in Gifford, B.R. and O'Connor, M.C. (eds) *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction,* Boston and Dordrecht: Kluwer, pp. 77-117.

Garrett, H.E. (1937). *Statistics in Psychology and Education.* New York, NY: Longmans, Green . Cited in Wiliam, D. (1994). Reconceptualising Validity, Dependability and Reliability for National Curriculum Assessment, in Hutchison, D and Schagen, I. (eds) *How Reliable is National Curriculum Assessment.* NFER. pp. 11-34.

Gipps, C. and Murphy, P. (1994). *a Fair Test? Assessment, Achievement and Equity,* Buckingham: Open University Press.

Gipps, C.V. (1994). *Beyond Testing: Towards a Theory of Educational Assessment,* London: Falmer Press

Glaser, R. (1991). *Expertise and assessment*. In M.C. Wittrock & E.L. Baker (Eds), Testing and Cognition (pp. 17-30). Englewood Cliffs, NJ: Prentice-Hall.

Glen, S. (2000). The dark side of purity or the virtues of double-mindedness?, in: H. Simons & R. Usher (Eds) *Situated Ethics in Educational Research (* London, Routledge, Falmer).

Goldstein, H. (1994). Recontextualizing Mental Measurement. *Educational Measurement: Issues and Practice, 13, 1.*

Goldstein, H. and Heath, A. (2000). Educational standards. Proceedings of the British Academy 102, Oxford University Press. pp.158 (eds).

Gott, R. and Duggan, S. (2002). Problems with the assessment of performance in practical science: which way now? Cambridge journal of education, 32(2), pp. 183-201.

Green,S. and Nickson,M. (1997). Research in effective assessment, Key stage 2 Science. Paper presented at BERA.

Green,A.(1999) Technical Education and State Formation in Nineteenth Century England and France, in Moon,B. and Murphy,P.(eds) *Curriculum in Context (*44-62). London: Open University.
Greene, J.C., Caracelli, V.J., & Graham, W.F. (1989). Towards a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis,* 11(3), pp255-274.

Griffiths, M. (1997). Why Teachers and Philosophers Need Each Other: philosophy and educational research. *Cambridge Journal of Education,* Vol, 27, 2, pp. 191-202.

Halliday, J. (2002). Researching Values in Education. *British Educational Research Journal,* Vol. 28, 1, pp. 49-62.

Hambleton and Cook, (1977). Latent trait Models and their use in the analysis of Educational Test Data, *Journal of Educational Measurement.*

Hammersley, M. & Gomm, R. (1997). Bias in social research. Sociological Research Online, 2, www.socresonline.org.uk/socresonline/2/1/2.html

Hammersley, M. (1995). *The politics of Social Research,* London, Sage.

Hammersley, M. (2005). Countering the “new orthodoxy” in educational research: a response to Phil Hodkinson. *The British Educational Research Journal, 31: (2),* pp. 139-155.

Handy, C. (1994). *The empty raincoat.* London, UK: Hutchinson. Cited in Wiliam, D. (2000a). The meaning and consequences of educational assessments. *Critical Quarterly 42 (1),* pp. 105-127.

Hargreaves, D.H. (1996). *Teaching as a research-based profession: possibilities and prospects,* Teacher Training Agency Annual Lecture Annual Lecture, (London, Teacher Training Agency).

Harlen, W. (1994). Developing public understanding of education- a role for researchers, *British Educational Research Journal,* 20,(1), pp. 3-16.

Harrison, E.F. (1999).*The Managerial Decision-making Process (New York, Houghton Mifflin).*

Henrysson, S. (1971). Gathering, analysing, and using data on test items. In R.L. Thorndike (Ed.), *Educational Measurement* (p141). Washington DC: American Council on Education.

Heppell,S.; Chapman, C.; Millwood.; Constable, M,; Furness, J. (2004). *Building Learning futures…* a research project at Ultralab, ARU.

Hodkinson, P. (2004). Research as a form of work: expertise, community and methodological objectivity, *British Educational Research Journal, 30 (2),* pp. *9-26.*

Hogarth, S.; Bennett, J.; Campbell,B.; Lubben, F. and Robinson, A. (2005). A systematic review of the use small-group discussions in science teaching with students aged 11-18, and the effect of different stimuli (materials, practical work, ICT, video/film) on students’ understanding of evidence. Research Evidence in Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of education.

Hoinville, G. and Jowell,R. (1989). *Survey Research Practice.* London: Gower.

Homan, R. (1990). Institutional controls and educational research, British *Educational Research Journal, 16*, pp. 237-248.

Hook, C. (1981). *Studying Classrooms.* Deakon University.

Horkay, N., Bennett, R., Allen, N., Kaplan, B. and Yan,F. (2006). Does it Matter if I take my writing test on computer? An empirical study on mode effects in NAEP. *The Journal of Technology, Learning and Assessment. Vol 5.(2).*

House of Commons, Science and Technology Committee (2002). *Science education from 14 to 19*. Third report of session 2001-2, 1. London: HMSO.

James, M. (1998). *Using Assessment for School Improvement.* Oxford: Heinemann (school Management Series).

Jenkins, E. (2007). School science: a questionable construct? Journal of Curriculum Studies, Vol, 29, (3), pp. 265-282.

Jick, T.D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly,* 24, pp. 602-611.

Johnson, M. & Green, S. (2004). On-line assessment: the impact of mode on student performance. Paper presented at BERA.

Kane, M.T. (1992). An argument-based approach to validity, *Psychological Bulletin,* 112, pp. 527-535.

Kehoe, J. and Jerard, M. (1995). Basic Item Analysis for Multiple-Choice Tests, *Practical Assessment, Research & Evaluation,* 4(10).

Kellnor, P. (1997). Hit-and-miss affair. *Times Education Supplement,* 23. Cited in, WILIAM, D. (2001b). What is wrong with our educational assessments and what can be done about it. *Education Review, 15, 1.*

Kerlinger, F.N. (1973). *Foundations of behavioural research (2nd Ed)* New York: Holt and Reinhart and Wilson.

King, N. (1998) Template analysis, in G. Symon and Cassell (eds) *Quantitative Methods and Analysis in Organizational Research.* London: Sage.

Korentz, Linn, Dunbar and Shepard (1991), The effects of high-stakes testing: Preliminary evidence about generalization across test: Symposium at The annual meeting of the American Educational Research assoc and the National council on Measurement in Education, cited in WILIAM, D. (2007). Comparative analysis c assessment practice and progress in the UK and USA. *Westminster Educational Forum Keynote Seminar.*

Kubicek, J.P. (2005). Inquiry-based learning, the nature of science, and computer technology: New possibilities in science education. *Canadian Journal of Learning and Technology*. Vol 31(1).

Kuhn, D. (1993). Science as Argument: Implications for Teaching and Learning Scientific Thinking. Science Education Vol 77, pp. 139-37.

Kuhn, T. S. (1970). *The Structure of Scientific Revolutions (2nd edn) Chicago, Il, University of Chicago* Press.

Lather, P. (2004). Scientific research in education in education: a critical perspective, *British Educational Research Journal, 31(2),* pp. 185-204.

Lawton,J. and Silver, H. (1973). A Social History of Education in England. London: Methuen.

Layton, D. (1973).Science for the people: the origins of school science. London: Allen and Unwin.

Leach, J.; Asoko, H.; Coles, J.; Jenkins, E.; Ryder, J. (2001). Keeping national curriculum science in step with the changing world of the 21st century. Final Report to QCA.

Levine, T. & Donitsa-Schmidt,S. (1998). Computer use, confidence, attitudes and knowledge: a causal analysis. Computers in Human Behaviour, 14(1), pp. 125-146.

Lewin, K. (1997). Test Development- Designing Tests and Presenting Results, in Bude,U. & Lewin, K. *Improving Test Design,* German foundation for International Development, Education, science and Documentation. Bonn: ZED. pp. 35-55

Lomax, P. and McLeman, P. (1984). The uses and abuses of nominal group technique in polytechnic course evaluation, *Studies in Higher Education.* 9 (2). pp. 183-190.

Lord, F. M. (1984). Statistical Errors of Measurement at Different Ability Levels. *Journal of educational Measurement,* 21, pp. 239-243.

Lord, F.M. and Novick, M.R. (1968). *Statistical Theories of Mental Test Scores,* London: Addison-Wesley.

Luecht, R. M., Hadadi, A., Swanson, D.B. & Case, S.M. (1998). A comparative study of a comprehensive basic sciences test using paper-and-pencil and computerized formats(testing the test). Academic Medicine, 73(10), S51-S53.

Lyle, J. (2003). Stimulated recall: a report on its use in naturalistic research. *British Educational Research Journal* 29, 6 pp. 861- 878.

MacCann, R.G. and Stanley,G. (2004). Estimating the standard error of the judging in a modified-Angoff standards setting procedure, *Practical Assessment Research and Evaluation.*

Mackenzie, I.S. (1988). Issues and methods in the microcomputer-based lab. Journal of Computers in Mathematics and Science Teaching. 1(4): pp. 18-20.

Magnusson,D. (1967) Test Theory. Reading. MA: Addison Wesley.

Massey, A. (1995). Evaluation and analysis of examination data: Some guidelines for reporting and interpretation, UCLES internal report, Cambridge.

Massey, A. J. (1995). Criterion-related test development and national test standards. *Assessment in Education,* 2, 2, pp. 187-203.

Mazzeo, J., and Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature* (ETS RR-88-21).Priceton, NJ: Educational Testing Service.

McClure,S. (1986). Educational Documents: England and Wales, 1816 to the present day (5th edition), London: Methuen.

McDonald, A.S. (2002). The impact of individual differences on the equivalence of computer-based and paper-and-pencil educational assessments. Computers & Education 39, pp. 299-312.

McNally, J. (1999). A Theory of Teaching Investigative Science. Paper presented at BERA.

Mead, A.D. & Drascow, F. (1993). Equivalence of computerized and paper cognitive ability tests: A meta-analysis. Psychological Bulletin, 114(3), pp. 449-458.

Merton, R.K. and Kendal, P.L. (1946) In Merton et al (1956). *The Focused Interview: a manual of problems and procedures.* Columbia University Bureau of Applied Social Research: Free Press.

Messick, S. (1980). Test validity and the ethics of assessments. *American Psychologist, 35, (11),* pp. *1012-1027.*

Messick, S. (1989). Validity, in R.L. Linn (Ed.) *Educational measurement* (3rd edn) pp. 13-103 (New York, American Council on Education/Macmillan).

Millar, R. (1996) *Towards a science curriculum for public understanding*. School Science Review, 77 (280),pp. 7-18.

Millar, R. and Driver, R. (1987). *Beyond processes*. Studies in Science Education, 14, pp. 33-62.

Millar,R. and Osborne,J (1999). (Eds) *Beyond 2000: Science Education for the future*. The report of a seminar series funded by the Nuffield Foundation. London. Kings College.

Morris, B. (1972). *Objectives and Perspectives in Education,* London, Routledge & Kegan Paul, pp. 60-61.

Murphy, P.J. (1986). Computer simulations in Biological education: Analogues or models? Journal of Computers in Mathematics and Science Teaching. 3(1): pp. 13-21.

Murphy, R. (1978). Reliability of marking in 8 GCE exams, *British Journal of Educational Psychology, 48,* pp. 196-200.

Murphy, R. (1982). A further Report of investigations into the reliability of marking of GCE examinations, *British Journal of Educational Psychology,* 52, pp. 28-63.

Murphy, P.K., Long, J., Hollerton, T. and Esterly, E. (2000). Pursuasion online or on paper: A new take on an old issue. Paper presented at the annual meeting of the American Psychological Association, Washington, DC.

Nash, R. (2005). Explanation and quantification in educational research: the arguments of critical and scientific realism, British *Educational Research Journal, 16,* pp. 237-248*.*

Negroponte, N. (1995). Being digital. New York: Vintage.

Newton, P.E. (2003). The defensibility of national curriculum assessment in England*. Research Papers in Education, 18,2,* pp. 101-127.

Newton, P. E. (2004). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31, 4*, pp. 419 – 442*.*

Newton, P. E. (2005). The public understanding of measurement inaccuracy. *British Educational Research Journal, 31, 4,* pp. 419 – 442*.*

Newton, P., Driver, R. and Osborne, J. (1999). The place of argumentation in the pedagogy of school science. International Journal of Science Education Vol 21 no 5, pp 553-576.

Newton, P.E. (2007). Assessment- A system fit for purpose? *Westminster Education Forum Keynote Seminar.*

Nichols,L.M. (1996). Paper and pencil versus word processing: a comparative study of creative writing in the elementary school. Journal of Research on Computing in Education, 29(2), 159-166.

Noyes, J. and Garland, K. (2004) Computer Experience: a poor predictor of computer attitudes. Computers in Human Behaviour. Vol 20, 6, pp 823- 840.

Nunnally, J.C. & Bernstein, I.H. (1993). *Psychometric theory.* Third Edition. New York: McGraw-Hill.

Nuttall, D.L. and Goldstein, H. (1984). Profiles and graded tests: The technical issues, in *Profiles* *in Action,* London, Further Education Unit.

Nuttall, D. L. (1987). The validity of assessments, *European Journal of Psychology of Education,* 2**,** 2, pp. 109-118.

Nuttall, D. L. and Willmott (1972). *Reliability, in British Examinations: Techniques of Analysis,* Slough: NFER.

O'Malley, J., Kirkpatrick, R., Sherwood,W., Burdick,H.J., Hsieh, M.C., Sanford, E.E. (2005). Comparability of a paper Based and Computer Based Reading Test in Early elementary grades. Paper presented at the AERA Division D Graduate Student Seminar, Montreal, Canada.

Oates, T. (2007). Protecting the innocent: The ethics of mass innovation in education & training. In Saunders,L.(ed) *Educational Research & Policy Making. Exploring the border country between research and policy.* Routledge.

Oates, T. (2009). Quoted in: Essays to be marked by 'robots', TES, 25/09/09.

Ofqual (2009) Annual qualifications market report. QCA.

Ofsted Science Subject Reports (1999- 2005). HMI. HMSO.

Oppenheim, A.N. (1992).*Questionnaire Design, Interviewing and Attitude Measurement,* Pinter.

Orlansky, J. and String, J. (1979). Cost-effectiveness of computer-based instructions in military training, IDA Paper P-1375, Institute for Defense Analysis, Alexandria, Virginia.

Osborne, R.J. and Wittrock, M.C. (1983). Learning Science: a generative process. Science Education, 67 (4) pp. 489-508.

Osborne,J, Collins,S. and Millar,R (June 2000). Keeping School Science in Step with the Changing world: A Review of Arguments and Evidence. Report commissioned for QCA.

Pollitt, A., Ahmed, A., Baird, J., Tognolini, J. and Davidson, M. (2007). *Improving the Quality of GCSE Assessment.* Final Report to QCA.

Pollitt, A., Hughes, S., Ahmed, A., Fisher-Hoch, H. and Bramley, T. (1999). The effects of structure on the demands in GCSE and A level questions. Final Research Report.(UCLES).

Pollitt, A., Hutchinson, C., Entwistle, N. & De Luca, C. (1985). What makes Exam Questions Difficult, Edinburgh, Scottish Academic Press.

Pommerich,M. (2004). Developing computerized versions of paper tests: mode effects for passage-based tests. The journal of Technology, Learning and measurement, 2(6), pp. 1-44.

Poplun, M., Frey, S. & Becker, D.F. (2002). The score equivalence of paper and computerized versions of a speeded test of reading comprehension. Educational and Psychological Measurement, 62(2), pp. 337-354.

Potter, J. and Whetherall, M (1986) *Discourse and social psychology: Beyond attitudes and behaviour*. London: Sage.

QCA (2002, 2003, 2004). *Evaluation of end of key stage tests.* London, QCA.

QCA (2009) Consensus statement. London, QCA.

QCA Science National Curriculum. London, QCA.

Quinlan, M. & Scharaschkin, A. (1999). National Curriculum Testing: Problems and Practicalities. Paper presented at 1999 BERA Conference.

Ravetz, J.R. (1997). Simple Scientific Truths and Uncertain Policy Realities: Implications for Science. Studies in Science Education, 30, pp. 5-18.

Reason, P. and Rowan, J. (Eds) (1981) *Human Enquiry.* New York: Wiley.

Ripley, M. (2004). Expert Technologies Seminar on e-Assessment: The e-Assessment vision. Presentation at Becta Expert Technology Seminar.

Roberts, R. and Gott, R. (2006). Assessment of performance in practical science and pupil attributes. Assessment in Education, Vol 13, (1), pp. 45-67.

Robson, C. (2002). *Real World Research.* Second edition: Blackwell.

Rokeach, M. (1973). *The Nature of Human Values (* New York, The Free Press).

Ross, A (1999). Curriculum: Construction and Critique. London, Falmer Press.

Ruddock, J. (1999). *Compound Fractures,* Seminar presentation, QCA, London.

Russell, M. (1999). Testing on computers: A follow-up study comparing performance on computer and on paper. Education Policy Analysis Archives, 7, (20).

Russell, M. and Hanley, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper. Educational Policy Analysis Archives, 5(3).

Russell, M. & Hanley, W. (2000). *Bridging the gap between testing and technology in schools.* Education Policy Analysis Archives, 8(19).

Russell,M. and Plati, T. (2001*). Effects of computer versus paper administration of a state-mandated writing assessment*. Teachers College.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science* 18**:** pp**.** 119-144.

Sax, G. (1989). *Principles of educational and psychological measurement and evaluation (*3rd ed.). Belmont, CA: Wadsworth.

Schagan, I. & Hutchinson, D. (1994). Measuring the reliability of national curriculum assessment. *Educational Research, 36(3),* pp. 211-221*.*

Schagan, I. (1994). Graphical representation of the reliability of national curriculum assessment, in Hutchison, D and Schagen, I. (eds) *How Reliable is National Curriculum Assessment.* Slough: NFER. pp 71-90.

Schagan, I.P. (1993). Problems in measuring in measuring the reliability of national curriculum assessment in England and Wales. *Educational Studies,* 19 (1), pp. 41-54.

Schagan, I. (1999). Testing, testing, testing. *Managing Schools Today,* **8,** 4, pp.28-9.

Schrok, J.R. (1984). Computers in science education: Can they go far enough? Have we gone too far? The American Biology Teacher, 46, 252-256.

Shepard, L. A. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14.

Shepard, L.A. (1993). Evaluating test validity, *Review of Research in Education,* 19, pp. 405-450.

Shorrocks-Taylor, D. (1999). *National Testing: Past, Present and Future (*Issues in Assessment and Testing). Leicester: BPS Books.

Singleton, C., Horne, J. & Thomas, K. (1999). Computerized baseline assessment of literacy. Journal of Research in Reading, 22(1),pp. 67-80.

Solomon, J. (1999). Meta-scientific criticisms, curriculum innovation and the propagation of scientific culture. Journal of Curriculum Studies, Vol.31, No1, pp. 1-15.

Stobart, G. & Gipps, C. (1998). The underachievement debate: fairness and equity in assessment, *British Journal of curriculum and Assessment,* 8, 3, pp. 43-49.

Stobart, G. (1999). *The validity of national curriculum assessment*. Paper presented at the 1999 BERA conference.

Stobart, G. (2000). What is fair assessment in a multicultural society? IAEA Conference Paper, Jerusalem.

Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education,* 12, 3, pp. 275-287.

Stobart, G. (2008). Testing Times, The uses and abuses of assessment. London, Routledge.

Sturman, L. (2003). Teaching to the test: science or intuition? Educational Research Vol. 45 no. 3 pp 261-273.

Sutton, R.E. (1997). Equity and high stakes testing: implications for computerized testing. Equity and Excellence in Education, 30(1) pp. 5-15.

Tapscot,D. (1996). Growing up Digital. New York: McGraw-Hill.

Tatteshall, K. (2009) quoted in: Time for pupils to take screen tests, says watchdog, *Times Online, December 17th.*

Taylor, C., Kirsch, I., Eignor, D. & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. Language Learning, 49(2), pp.219-274.

Thomas, R. & Hooper, E. (1991). Simulation: An opportunity we are missing. Journal of Research on Computing in Education, 23(4), pp. 497-513.

Thompson, B., & Levitov, J.E. (1985). Using microcomputers to score and evaluate test items. *Collegiate Microcomputer,* 3, pp 163-168.

Thorndike, R.M., Cunningham, G.K. Thorndike, R.L., & Hagen, E.P. (1991). *Measurement and Evaluation in psychology and education* (5th ed). New York: MacMillan.

Threlfall, J.; Pool, P.; Homer, M. & Swinnerton, B. (2007). Implicit aspects of paper and pencil mathematics assessment that come to light through the use of the computer. Educational studies in Mathematics. 66, pp 335-348.

Tooley, T. & Darby, D. (1998). *Educational Research: a critique. A Survey of Published Educational Research (*London, Office for Standards in Education).

Tymms, P. (2004). Are standards rising in English primary schools? *British Educational Research Journal* 30: 4, pp. 477-494.

Tymms, P. (2007). Evidence given at the DCSF Select Committee on Testing and Assessment.
U.S. Department of Commerce. (2002). A nation on-line: How Americans are expanding their use of the Internet. Washington, DC.

U.S. Department of Commerce. (2002). A nation on-line: How Americans are expanding their use of the Internet. Washington, DC.

Walker, R. (1985). *Doing Research: A handbook for teachers.* Methuen: London.

Wang, S. (2004). Online or paper: does delivery affect results? Administration mode comparability study for Stanford diagnostic Reading and Mathematics tests. San Antonio, Texas: Harcourt.

Watson, J.R., Goldworthy,A. and Wood-Robinson, V (1999). Practical Science Investigations in England and Wales: a National Survey. Paper presented at ESERA conference, Kiel.

Wheadon, C. & Adams, C. (2007). *The comparability of onscreen and paper and pencil tests: no further research required? Paper presented at the IAEA, Azerbejan.*

Wiersma, W. & Jurs, S.G. (1990). *Educational measurement and testing(*2nd ed). Boston, MA: Allyn and Bacon.

Wiggins, G. (1993). Assessment: Authenticity, Context, and Validity. *Phi Delta Kappan* 75, 3, pp. 200-214.

Wiliam, D. (1993), Validity, dependability and reliability in National Curriculum Assessment, *The Curriculum Journal,* **4,** 4, pp. 335-350.

Wiliam, D. (1994). Reconceptualising Validity, Dependability and Reliability for National Curriculum Assessment, in Hutchison, D and Schagen, I. (eds) *How Reliable is National Curriculum Assessment.* Slough: NFER, pp. 11-34.

Wiliam, D. (1995a). Combination, aggregation and reconciliation: evidential and consequential bases. *Assessment in Education,* 2, 1, pp. 53-73.

Wiliam, D. (1995b). It'll all end in tiers, *British journal of Curriculum and Assessment,* **5,** 3, pp. 21-24.

Wiliam, D. (2000a). The meaning and consequences of educational assessments . *Critical Quarterly 42(1),* pp. 105-127.

Wiliam, D. (2000b). Integrating Summative and Formative functions of Assessment. Keynote address to the European Association for Educational Assessment, Prague, Czech Republic.

Wiliam, D. (2001a). Reliability, validity and all that jazz. *Education 29,3,* pp. 17-21.

Wiliam, D. (2001b). What is wrong with our educational assessments and what can be done about it. *Education Review,* 15, 1.

Wiliam, D. (2001c). *Level Best? Levels of attainment in National Curriculum Assessment.* London: ATL.

Wiliam, D. (2007). Comparative analysis of assessment practice and progress in the UK and USA. *Westminster Educational Forum Keynote Seminar.*

Wiliam, D. (2008). Six degrees of integration: an agenda for joined-up assessment. *Annual Conference of the Chartered Institute of educational Assessors,* London, 23rd April.

Willmott, A. & Nuttall, D. (1975). *The Reliability of Examinations at 16+.* Slough, NFER.

Wilson, N. and McLean, S. (1994). *Questionnaire Design: A Practical Introduction.* Newtown Abbey, Co. Antrim: University of Ulster Press.

Wolf, A. & Silver, R. (1993). The Reliability of Test Candidates and the Implications for One-Shot Testing. *Educational Review,* 45, 3.

Wolf, A. (1991). Assessing core skills: wisdom or wild goose chase? *Cambridge Journal of Education,* 21, pp. 189-201.

Wolpert, L. (1992). The unnatural Nature of Science. Cambridge, MA: Harvard University Press).

Wood, D.A. (1960) *Test Construction.* Merrill.

Wood, R. (1991). *Assessment and Testing,* Cambridge University Press.

Woodhead, C. (1998). Foreward, in Tooley,J. & Darby (Eds) *Educational Researcher, 22*, pp. 15-23.