THE UNIVERSITY OF LEEDS

SCHOOL OF EDUCATION

Report to QCA on an investigation of the construct relevance of sources of difficulty in the Key Stage 3 ICT tests

November 2006

Assessment and Evaluation Unit
School of Education
The University of Leeds
Leeds LS2 9JT

An investigation of the construct relevance of sources of difficulty in the Key Stage 3 ICT tests

Contents

	Page
Executive summary	**2**
Introduction	**7**
The KS3 ICT test	**7**
Issues for the project	**8**
Methodology	**10**
Initial framework for sources of difficulty	**13**
Background	**13**
Arriving at possible sources of difficulty	**14**
The initial list of possible sources of difficulty (SoDs) in the test	**15**
Data	**19**
Observation data	**19**
Quasi-protocol data	**22**
Analyses and findings	**23**
The nature of the evidence	**23**
Case studies	**25**
Support for the hypotheses	**30**
Final framework for sources of difficulty	**34**
Features of tasks	**35**
Software	**38**
Preparation	**39**
A note on the distinctiveness of listed 'sources of difficulty'	**41**
Pupil state of mind	**42**
Recommendations	**43**
Recommendations for test construction (a framework for quality assurance)	**44**
Recommendations for preparation by teachers	**47**
Discussion	**50**
Qualities of the pupil	**50**
Pupils experience with ICT outside school	**50**
Authenticity	**51**
A conundrum of the assessment	**53**
Conclusion	**54**
The potential of the tests	**54**
The risks arising from the assessment as written	**54**
The risks arising from the nature and form of the assessment	**55**
Issues for further research / investigation	**56**

Executive summary

This is the report by the Assessment and Evaluation Unit of the University of Leeds School of Education on an investigation into the construct relevance of sources of difficulty in the new Key Stage 3 ICT test.

The project
The KS3 ICT test is an on-screen e-assessment being developed by the Qualifications and Curriculum Authority (QCA) under contract to the Department for Education and Skills (DfES). The contractor responsible for overall project delivery is Research Machines (RM) plc. Work on the project began in 2004, with a feasibility study that gave rise to the design for the test. There is now a cycle of yearly pilots, with the aim of making the test statutory in 2008. This evaluation is of the 2006 pilot, and is just one of a number of projects investigating the tests.

The tests that were piloted in May 2006 differ in a number of important respects from other key stage assessments, and as a consequence the sources of difficulty for pupils taking the tests require cannot be presumed on the basis of parallels with tests in other subjects. It shares the central purpose of the assessments conducted at the end of key stages: to generate an NC level for each pupil. It also has a similar structure: two tests of 50 minutes containing a number of tasks to complete. However, the tests are conducted entirely onscreen. Further, the focus of the assessment is mostly on processes rather than outcomes – on how the tasks were completed, not the answers given – which means that the assessment is not made on the basis of 'marks' but on observed examples of 'opportunities' (sequences of recorded actions or configurations of saved documents that are taken to reveal the process by which the pupil was seeking to accomplish the task). Finally, the assessment is 'machine-scored', that is to say the computer logs all the actions taken by the pupil when taking the test, and the final state of documents worked on, and automatically identifies from that the examples of taken opportunities – which are then used to determine the final award.

This project was set up to try to determine what prevents the opportunities being taken – the 'sources of difficulty' in the test. In an ideal assessment, the only sources of difficulty in play are those which relate to the attribute being assessed – in this case the ICT capability of the pupil. In other words, the only thing that ought to prevent pupils from acting in a way that will trigger the opportunities is their lack of ability, and so the assessment will lead to an appropriate level award.

However, it was anticipated that in practice there would be other sources of difficulty that would lead to 'false negatives' in the assessment – where pupils with ICT capability at a certain level are prevented by some feature of the assessment situation from achieving that level. Such sources of difficulty apply to all assessments.

In investigating sources of difficulty (SoDs), therefore, the 'construct relevance' (whether the source was a legitimate one because it related to ICT capability) was a key element. The project was also to focus strongly on the SoDs that relate to aspects of the assessment that could be 'mitigated', that is could be addressed in future development of the tests. So, the SoDs that arise from the nature of the tasks were central. The project was conducted with a view to making suggestions for improvement in test construction, and towards giving advice about test preparation.

Methodology

Broadly, the approach adopted for the project was a qualitative examination of possibilities for sources of difficulty, using observation of pupils taking the tests to determine whether what might be sources of difficulty were actually sources of difficulty in practice.

There were three phases to this:

1. The research team used task analysis and informal observations of individuals taking the tests to generate a set of initial hypotheses about possible SoDs
2. Empirical work was undertaken in a number of schools, consisting of: observations of pupils taking the practice tests and the actual pilot tests; interviews with pupils and teachers; and quasi-protocol recordings of pupils trying tasks taken from the tests while giving a commentary on their thinking and actions.
3. The research team analysed the data against their initial framework of hypothesised sources of difficulty, and drew conclusions about the incidence of SoDs, the impact that they have, and what actions might be undertaken in future to reduce that and improve the tests.

The approach to data collection was qualitative, mostly because of purpose. The project was not intended as an account of what happened in the 2005 and 2006 tests, but to explore how pupils respond to the kind of demand that the tests place upon them. What was required was insight into what occurs when pupils take the tests, and this requires qualitative approaches such as direct observation and closely related questioning, informed to an extent by further information that could be provided by an interview with the teacher.

Observations of pupils taking the tests was done in three 'windows': the 'practice tests' – which were essentially revised versions of the 2005 tests; the pre-test, which was the final full run through of the 2006 tests, taken by a relatively small group of schools; and the actual 2006 tests. Data from each were collected for this project, with observers visiting schools as the tests were taking place, observing individual pupils and taking close notes on what was seen, then talking to the pupil afterwards, and their teacher at some time during the visit.

In addition pupils from four schools participated in 'quasi-protocol' observation – that is, pupils taking the (practice) test and when doing so 'thinking-aloud' to give a commentary on what they are doing and why, which was recorded on mini-disc and subsequently transcribed.

Data collection	Test	Number of schools	Number of pupils
Observation	Practice	10	20
Protocol data	Practice	4	20
Observation	Pre-test	4	12
Observation	Summative	7	14

Findings

Identified sources of difficulty

The project found sufficient evidence to conclude that several sources of difficulty that were not relevant to the construct being assessed had both occurred and had had a significant impact on the tests.

These are separated into: the difficulties that arose from characteristics of the tasks and software; and those which arose when pupils were inadequately prepared for the tests, for example were not familiar enough with the software interface or with the particular characteristics of the applications used.

Task and software-related sources of difficulty

- The task instructions are not explicit enough about what has to be done.
- The task is ambiguous, or vague.
- Limitations in subject knowledge (e.g. mathematics) constrain engagement.
- The requirement for sustained concentration is too great for many pupils, whatever their ICT capability.
- Pupils know enough to succeed in the tasks without using ICT for all the steps.
- The pupils (for whatever reason), get lost (visually or conceptually), and as a result waste time, or even give up on the task.
- The pupils (for whatever reason) are or become confused, frustrated, dispirited or de-motivated and do not engage in a way that leads to taking up the opportunities.
- The tasks require too many things to be remembered across a sequence of screens / actions.
- The need to keep switching between objects or parts of objects that cannot be seen at the same time mean that pupils become disorientated, and lose track of where they are in the task.
- The memory demands of the tasks are too high.
- There is not enough to enjoy in the assessment in order to sustain engagement.
- There is an inappropriately high demand for ICT knowledge and skill that is not relevant to the assessment.

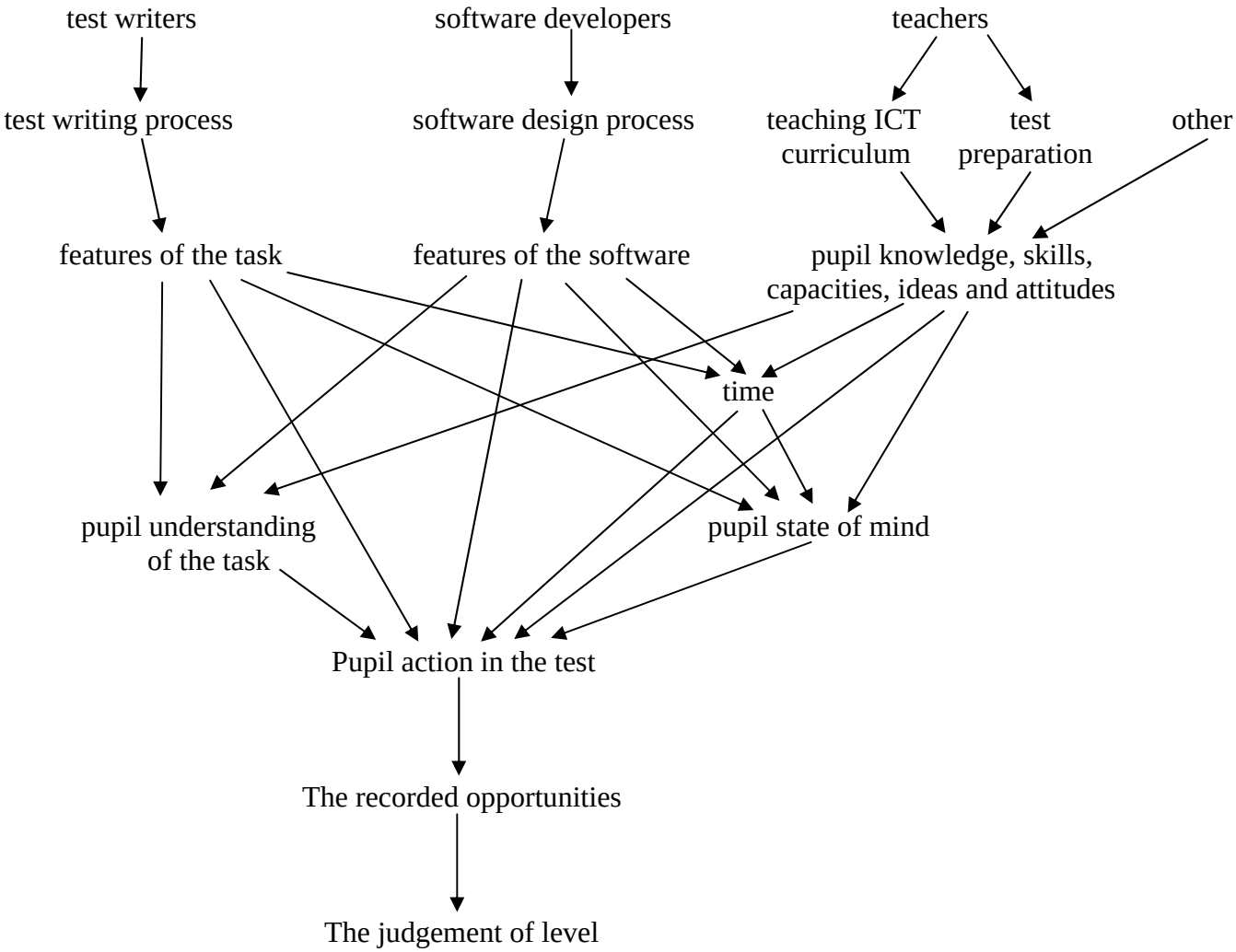
Preparation-related sources of difficulty

- The pupil has always used the same software, so her proficiency is context bound.
- The unfamiliarity of the software leads to multiple attempts to achieve a purpose.
- The pupil wants to undertake a particular action but does not know how.
- Pupils spend time making something nice rather than getting on with the test, e.g. trying several different fonts to choose the most appealing.
- Lack of knowledge of some specific actions to achieve certain things has disproportionate effects, such as precluding all progress with the task.
- Actions fail that were assumed to lead to a desired immediate objective.
- The cramped and busy screen is confusing.

A framework of inter-relationships within sources of difficulty

A number of the identified sources of difficulty relate to each other. Losing track leads to losing concentration; lack of knowledge impacts on enjoyment which de-motivates; and so on. The issue for impact is how the pupil's actions are affected by sources of difficulty, but there are often hidden variables, such as the effect on time, and the effect on pupil's state of mind, which in turn have effects on actions.

An examination of these inter-relationships, in the light of the experience of the project, led to the following diagram showing the complexity of the routes by which sources of difficulty have their ultimate effect on awarded levels.



Recommendations

For most of the sources of difficulty listed above, especially those concerned with the tasks and software, there is not a specific recommendation to 'deal' with it. The nature of the test construction process is such that one cannot say 'Do X to address Y', especially in the light of the inter-relationships considered in the diagram above. It is more the case of "Do these things together, and it should help all of those".

A. For the task-related sources of difficulty therefore, the recommendations of the project are:

1. To keep in mind when developing tasks the following five dimensions of potential sources of difficulty:
    - language;
    - memory and organisation;
    - irrelevant knowledge and skills;
    - too much to do / too high a demand on concentration;
    - idiosyncratic interpretations of the task by pupils

2. To use the following criteria in task reviews:
   - Will the overall task make sense to the pupil?
   - Will there be sufficient appeal and interest in the task?
   - Is there ambiguity in the instructions, or other language in the task?
   - Is there too much repetition, such that it brings a risk to focus and motivation?
   - Does the task have a clear goal, and enable a sense of purpose to be retained throughout?
   - Is there a relatively easy beginning that will get the pupils going on the task?
   - Is there sufficient structured support in the task?
   - Is the task do-able by most pupils in the allotted time (typically 16 minutes)?
   - Does the task assess what it is supposed to assess?
   - Will the actions that pupils are likely to do in response to the task be the ones that the task is designed to provoke, and which will trigger the appropriate assessment opportunities?
3. To apply the consideration of potential SoDs and the review criteria to the following five aspects of tasks:
   i. Context – the nature of the activity in the task
   ii. Content – the language, numbers, pictures, etc. used in the task
   iii. Task instructions
   iv. Visual representation onscreen
   v. Scenario, or 'story' of how the pupil is expected to engage with the task.
4. Wherever possible (unless the quantity of reading would become excessive) to make the tasks and task instructions explicit about what is to be done, and how responses are to be given.
5. To pay keen attention to the possible effects of features of tasks and the software interface on pupils' emerging states of mind during the assessment.

Related to that, the project makes two further recommendations, which are about the test-development process.

6. To try the finished tasks out with pupils, in order to gain information about pupils' interpretations, and the effect of those on their engagement.
7. To give opportunities within test development for the task writer to ask for changes to the digitised question.

B. For the software-related sources of difficulty, the recommendations of the project are:

1. To reduce as far as possible by appropriate labelling of tools and functions the level of particular knowledge required to use the software.
2. To work to make the software as smooth-running as possible with respect to functions such as highlighting and pasting across applications.
3. To add an onscreen scribble pad.
4. To develop a means of keeping track of where one is in a task and the test.

C. For the issue of preparation for the tests, the recommendations of the project are:

1. To raise pupils' awareness of the test, and of themselves as test takers.
2. To improve pupils' familiarity with the interface and applications.
3. To develop ICT capability in a way that is commensurate with the approach of the tests – for example by using applications together to achieve a purpose.

1. Introduction

The KS3 ICT test

In common with other National Curriculum testing, the KS3 ICT tests that were piloted in May 2006 are designed to generate an NC level for each pupil. A full test comprises two computer-based sessions of 50 minutes, during which the pupils address a number of tasks which are presented to them via the interface. The assessment package operates behind this, recording the actions that the pupil takes and subsequently evaluating them. The KS3 ICT tests differ in a number of important respects from other key stage assessments. In particular, the conception, development and implementation of these tests means that the focus of assessment is on process rather than outcome. Partly in consequence of this, the test is not comprised of items in the conventional sense – questions with attached marks. In general, the ICT tests are not designed or implemented with restricted, finite outcomes as the scoreable objectives. In these tests, the stimulus is for the pupils to engage in some activity which has an explicit goal, but in which the goal itself is subsidiary to the means by which that goal is pursued. Instead of items and marks, the points of assessment are called 'opportunities' and represent machine scored sequences of actions that (in the context of the task the pupil is engaged upon) may be inferred to reveal the process by which the pupil is seeking to accomplish the task. That said, however, the first task in the first session does contains a number of more direct questions with restricted, finite elements that are scored on the basis of the correctness of the answers, and so do behave in a way that is close to conventional test items.

Opportunities are either triggered or not, and while triggering an opportunity provides evidence of capability at a given level, not triggering it cannot be construed as anything other than the absence of such evidence. This absence of evidence is essentially neutral in assessment terms. The fact that an un-triggered opportunity is not in any sense 'wrong' is the principal reason that opportunities do not (and could never) operate in the same way as conventional items. An opportunity may be triggered infrequently simply because it represents an unusual (though legitimate) procedural strategy. Significantly, the procedure might also represent a relatively low level approach to the problem. Thus, a low 'hit' rate for an opportunity cannot be regarded as necessarily providing any useful information about ability. Put another way, conventional test models operate on the interpretation of 'didn't do x' as 'can't do x', but this interpretation cannot be made of un-triggered opportunities.

As opportunities cannot be regarded as items in the usual sense, they cannot simply be combined additively according to the classical reliability model or used to make ability estimates under item response theory models. A different approach is required. The model used combines the elements of evidence provided by triggered opportunities by pooling the evidence at each of the three NC levels targeted by the test tier. A judgment is then made as to whether the evidence at each level is sufficient to justify awarding that level. Evidence from higher levels is allowed to contribute to the evidence base for lower level decisions, but not the other way around.

The targets of the assessment are the processes of purposeful action in ICT, which are the thrust of curriculum ambition as summarised in the National Curriculum. The point of ICT is to be useful, to draw on skills to address purposes. So to have ICT capability at a level it is not enough to have the skills associated with that level in isolation, or to be able to engage with the purposes specified for that level, it requires the use of one in pursuit of the other.

The effective use of tools to achieve purposes beyond ICT skill / tool use is often interpreted as the open use of self-chosen tools towards general purposes invoked or interpreted by the individual. However, the constraints of a timed test are such that neither the purposes nor the skills can be assessed fully when in combination, and this test assesses the use of particular skills in tool use in the context of a particular set of computer applications to address a set of restricted purposes (the purpose of doing what told to do in completing a task to achieve a purpose that is externally defined, often related to completing something others have started, usually in the context of a work-related broader purpose). The anticipated actions necessary to achieve these limited purposes can be captured by the computer, and so can be the basis of the assessment, using a presumed relationship with the National Curriculum.

A critique of the KS3 ICT test can be made at two levels: one being the validity of the relationship, with commentary on what might improve the likelihood of a correct assessment outcome, the other being the practicalities of the relationship, and what might improve the quality of the evidence.

This evaluation is wholly focused on the latter, treating as unproblematic the issue of the validity of using the kind of evidence that the test generates to assess the National Curriculum in ICT. We consider only what might in the assessment context prevent the pertinent actions of a pupil occurring or being captured – referred to henceforth as ‘sources of difficulty’ – with a view to making suggestions for improvement in test construction and advice about test preparation.

Issues for the project

As the identification of sources of difficulty in the test is aimed at constructive improvement of the tests, it is very important to the project to distinguish between construct relevant and construct irrelevant sources of difficulty, that is between sources of difficulty that are in the test by design – the sources of difficulty that are related to ICT capability – and those that were not intended to be in the test, but nevertheless are.

There are a number of issues pertinent to that consideration.

1. Testing ICT capability

The KS3 ICT test is not intended to be a skills test, but a test of capability. It is to assess more than the competence to do simple tasks using commonly used software applications; it is to assess the technical and cognitive proficiency to access, use and communicate information using technological tools. Meaningful descriptions of such capability at different levels are given in the National Curriculum Attainment target

level descriptions for ICT, and the KS3 ICT test offers as an outcome a judgement of the level at which pupils are assessed as operating.

The ICT National Curriculum levels are almost exclusively 'best-fit' process statements about the quality of the use of tools, rather than descriptions of knowledge and skills. As such, determining the level of each pupil accurately is not appropriately achieved through the examination of outcomes, and this poses a challenge for tests, which are usually scored on the basis of outcomes. The approach adopted for these tests involves a novel use of technology both to track the use of particular tools and to subsequently evaluate the quality of that use.

The assessment presumes that process can be captured in a timed test, but if fundamentally it cannot, some of the sources of difficulty observed will have arisen from the underlying issue, rather than from particular characteristics of these tests.

2. On-screen assessment

Whatever is being assessed (ICT capability or something else entirely) the fact of the assessment being on-screen has its effects. For example, in a paper-based assessment, it is possible to scan and flick back over previous material more easily – whereas in a computer-based assessment it is more feasible to re-arrange material, and to try out answers. The fact that this assessment is on-screen is a factor in some sources of difficulty, but these should not be confounded with the sources of difficulty that arise from task writing or software design.

3. Legitimacy of sources of difficulty

Pupils may find difficulty with the test questions just because they do not have the capability being assessed. The questions may be too difficult for them because they are not at the level being assessed. Although this may raise questions about the entry decisions of the school staff, it is always a possibility where pupils are struggling.

4. Preparation

Pupils may struggle because they are under-prepared for the assessment, either because of insufficient time to become familiar with the software interface and the particular form of its applications, or because the purposes, structure and demands of the test have been insufficiently explained. In both cases, pupils could struggle with the test, despite having the capability, and the source of difficulty for them lies outside the main focus of the project. However, it is an important secondary focus to identify where preparation might most helpfully be targeted.

Summary

The sources of difficulty that this project was attempting to identify were the sources of difficulty that do not arise because of the problem of using timed tests to assess capability, nor from the fact of it being an on-screen assessment, nor from pupils lacking the ICT capability being assessed, nor through lack of preparation.

Teasing these strands out from observations was one of the major challenges of the project.

Methodology

Specification

The specification for the project was to "investigate the construct relevance of sources of difficulty in the test" by exploring:

- which SoDs in the test can be demonstrated to have a substantial impact on pupils' performances;
- whether any such SoDs have a disproportionate impact on any identifiable group of pupils taking the test;
- if any SoDs in the test amount to implicit or hidden criteria in the assessment, and if so what should be done – e.g. make the assessment criteria explicit, or remove the SoD?
- how SoDs in the test might be amended or ameliorated.

It is important to note that the specification also gave a number of specific examples of what the project was *not* to focus on:

- other aspects of validity ('face validity'; reliability; concurrent evidence of validity; discriminant evidence of validity; content evidence of validity; fairness for all pupils);
- formative assessment or assessment for learning;
- level awarding;
- an overall evaluation of the test's validity;
- the decision to use a non-proprietary desktop environment;
- the levels that pupils achieved in 2005.

The narrow focus of the project was because this investigation into the construct relevance of sources of difficulty in the tests was just one part of a broader attempt to engage with the question of validity in the tests.

As a result, the project was not required to engage in a statistical demonstration of sources of difficulty using national data. What was looked for was more detailed and textured work with a smaller sample – in other words the work would be hypothesis generating rather than hypothesis testing.

It was also the case, as indicated above, that the main focus of the project was not on the role of the teachers in preparing pupils for the test, but on the interaction between pupils and the test – with a primary aim of the project to provide advice to task writers.

Further, while examples would be necessary to illustrate the meaning of generic issues, the focus would not be on task-specific difficulties.

Approach

Essentially the project fell into three phases:
Phase 1: Task analysis and familiarisation with the test by the research team to generate hypotheses about possible SoDs.

Phase 2: Empirical work (observation / interviews) with pupils / teachers framed by hypothesised SoDs.

Phase 3: Analysis of the data leading to the generation of further hypotheses about the SoDs in the test.

It was acknowledged from the start that this would be a reflective project, with little reference to the existing literature, and a limited role for data. It would raise issues, and give suggestions about how the issues might be addressed.

Data collection and analysis

The approach to data collection adopted was strongly qualitative. To an extent, the interpretive approach was forced by the high number of variables in the test, with only a few core tasks, but variations within them – with different tiers, different variants and random elements in questions bringing the possibility of no two observed tests being exactly the same – limited any opportunity for comparative data or numerically based generalisability. The data needed to be mostly about how pupils respond in general to this kind of thing, rather than how particular elements of the tests were dealt with.

However, the approach to data collection was qualitative also because of purpose. The project was not about what happened in the 2005 and 2006 tests, as how pupils respond to the kind of demand that the tests place upon them. What was required was insight into what occurs when pupils engage in the tests, and this requires direct observation and closely related questioning, informed to an extent by further information that could be provided by an interview with the teacher.

There were three 'windows' for observation of pupils taking the tests: the 'practice tests' – which were essentially the 2005 tests, re-packaged; the pre-test, which was the final full run through of the 2006 tests, taken by a relatively small group of schools; and the actual 2006 tests. Data from each were collected for this project, with observers visiting schools as the tests were taking place, observing individual pupils and taking close notes on what was seen, then talking to the pupil afterwards, and their teacher at some time during the visit.

The interviews with pupils immediately afterwards were about both their experience with that set of tasks – how they felt, how they engaged – and also how the tasks contrasted with what they usually do in ICT lessons. The interviews with teachers attempted to gain information about teaching activities that may have been relevant to pupils' performance and engagement, for example by ascertaining the degree of support given by teachers in familiarising pupils with the package, and by exploring the extent to which teachers' usual provision enables pupils to explore different examples of ICT tools.

Where possible these interviews were recorded on mini-disc and later transcribed, but in other cases notes were made.

In addition four schools agreed to participate in 'quasi-protocol' observation – that is, pupils taking the (practice) test and when doing so 'thinking-aloud' to give a commentary on what they are doing and why. These were recorded on mini-disc and subsequently transcribed.

Samples

As the data collection was not designed for representative description, the schools to be used for the main data collection – while selected to provide a range of types with a full variety of pupils – were as local as possible, to improve access. Our school contacts in the six local LAs were the first port of call, although there were not sufficient local schools taking part in the pre-test, so schools from a wider geographical area were visited. The data collection involved:

Data collection	Test	Number of schools	Number of pupils
Observation	Practice	10	20
Protocol data	Practice	4	20
Observation	Pre-test	4	12
Observation	Summative	7	14

Interviews were undertaken with pupils and teachers in all contexts.

The observation notes and transcripts were subsequently examined and analysed in relation to a set of hypotheses derived from a framework for examining sources of difficulty, as described in the following section.

2. Initial framework for sources of difficulty

Background
In traditional tests, difficulty is directly related to items. A problem is 'difficult' if relatively few pupils are able to solve it. The correctness or otherwise of the response to an item, or more usually the pattern of responses to a large number of items, is used as the measure of competence. In a second and independent step, the measure of competence is used to infer something about cognitive ability in the pupils. The quality of these two steps is judged differently, reliability relates to the consistent generation of response patterns while validity is concerned with the extent to which the measure thus obtained relates to the desired construct. In essence, the two questions are whether or not the test is a good one, and whether or not the test is the right one.

In the KS3 ICT assessments a fairly radical rethink of what is meant by difficulty is required. In the present context, the outcome of the test is a classification to NC level without the intervening consideration of traditional item responses. Instead, the measurement model considers the extent to which pupil actions provide enough evidence to support the classification of a pupil at a particular level.

The steps in the process of assessment, from the initial stimuli to the final outcome, might be viewed in the following way. On each row of the table below, the process indicates what happens when moving from the current step to the next.

Step	Process	Source
Stimulus	The task as perceived by the pupil	Task writers
Cognition	What action to take?	Pupil (Target measure)
Action	Logging of activity	System
Judgement 1	Is this evidence?	Opportunity coding
Judgement 2	If so, at what level?	Rules Base
Evidence	How is the evidence pooled?	Sufficient Evidence Model
Level Classification		

The third column represents the locus of the process. This is also the source of difficulty in that it determines (or at least strongly influences) the outcome for that link in the chain.

The second process – "What action to take?" – is where the pupil contributes to the overall structure. It is here that the perceived task is turned into recordable activity. This then is the focus of the assessment since it is the level at which this happens that the final outcome should reflect.

The last four processes are in essence the level awarding process, and not therefore a focus of the research, which is focused on the first and second steps in the table – although other sources of difficulty are found in the areas of pupil motivation, pupil capability and the relationship between pupil thinking and the tasks.

Arriving at possible sources of difficulty

As a way of beginning the project, a 'brainstorming' process was used to arrive at a set of possible SoDs that could be used as potential hypotheses. This was done firstly by collectively working through the 2005 test and reflecting on what factors seemed to be at work, and also by 'pilot' activity – observing others as they worked through the tests, and reflecting on what could be seen, and also what underlying factors may have affected their actions.

This collection of possible SoDs was then organised into a structure, as described in the following section, followed by the structured set of possible sources of difficulty that arose from this process. This became the initial framework for the research.

A set of hypotheses to examine

Each of the steps in the process of assessment outlined above gives rise to possible SoDs:

1. SoDs that affect how the task is perceived by the pupil
2. SoDs that affect the actions that are taken
3. SoDs that affect how the activity is logged
4. SoDs that affect how the logged activity is judged as evidence
5. SoDs that affect how the evidence is levelled
6. SoDs that affect how the evidence is pooled to arrive at a level judgement

but only the first and second of these are considered in the research. A structure was added to these, because the effects on perception and action arise from different kinds of sources, at different levels of description.

1. SoDs that affect how the task is perceived by the pupil
    - 1a. SoDs in the task that affect how it is understood
    - 1b. SoDs in the pupil's knowledge and ideas that affect how it is understood
    - 1c. SoDs in the pupil's experience that affect their knowledge and ideas
2. SoDs that affect the actions that are taken
    - 2a. SoDs in the tasks that affect the actions that are taken
    - 2b. SoDs in the pupil's thinking that affect the actions that are taken
    - 2c. SoDs in the pupil's experience that affect their thinking

There is then further structure in 2c, because 'experience' includes both the experience when doing the tasks and the experience which preceded doing the test. Within the more immediate experience, this splits into those that arise from the tasks themselves, and those that arise from acting in the general software environment. Further, it seems helpful to isolate the SoDs in each of these that may affect states of mind, such as confusion, frustration and motivation, so there are further sub-categories.

The initial list of possible sources of difficulty (SoDs) in the test.

1. SoDs that affect how the task is perceived by the pupil (and hence the actions that are taken).

1a. SoDs in the task that affect how it is understood

- The task instructions are not explicit enough about what has to be done.
- The language of the task is difficult.
- The task is very complex.
- The reading demand of the task is too high.
- The task is ambiguous.

1b. SoDs in the pupil's knowledge and ideas that affect how it is understood

- The pupil's attitude to tests leads to focusing on completing the tasks in as short a time as possible;
- The test situation leads to less sophisticated approaches being used by some pupils, as they prioritise on being correct. Clunky but reliable approaches are perceived to be better because they lead to right answers.
- The test situations lead to attempts by some pupils to be as sophisticated as possible, even though the software is unfamiliar – and their rehearsed skills do not fit.
- The pupils are unaware of the short cuts they use, the habits they have developed, and so are not sufficiently in control of them to recognise the need for an alternative.
- Limitations in subject knowledge (e.g. mathematics) constrain engagement. For example, to operate with a spreadsheet in a worthwhile way there is a high threshold of mathematical understanding, not just ICT capability. Pupils have to be able to model the relationships of the situation in their own understanding in order to be able to know what cells to change and in what way.

1c. SoDs in the pupil's experience that affect their knowledge and ideas

- The pupil has always used the same software, so her proficiency is context bound;
- The time taken to become familiar with the different environment is insufficient to enable the level of familiarity that would be needed to become comfortable enough to use the approaches that the software is looking for as evidence of higher level performance.

2. SoDs that affect the actions that are taken (and hence the activity that is logged).

2a. SoDs in the tasks and software environment that affect the actions that are taken

- Some of the opportunities are not in line with typical pupil thinking about the kinds of tasks being undertaken, so that the available opportunities are not taken.
- The unfamiliarity of the software leads to multiple attempts to achieve a purpose.
- The demands in the interaction between tasks and software on short-term memory and organisational skills are inappropriate to the level of ICT capability that is being assessed.

- The scaffolding in the tasks to support pupils in showing what they can do is inadequate, either because it is insufficient, or inappropriate (does not chime with pupil thinking), or because it is unfamiliar – the kinds of support offered by conventional assessments takes time to learn, as is part of exam preparation.
- The assessment is not 'ramped' successfully – with relatively easy steps at first. The first thing that is to be done is quite tricky, so some pupils are put off.
- Tasks require actions that call on knowledge that is not there, and which it is unreasonable to expect.
- The tasks call on knowledge of the world that is outside the reasonable expectations for the age group.
- There is too much to do in the time available.
- The necessary systematic exploration to find out how to do things requires a level of knowledge that is at a higher level than the ICT capability being assessed.
- The requirement for sustained concentration is too great for many pupils, whatever their ICT capability.

2b. SoDs in the pupil's thinking that affect the actions that are taken (see also 1 above – how the task is perceived)

- The pupil wants to undertake a particular action but does not know how.
- High levels of subject knowledge make the use of ICT unnecessary to achieve the specified ends. For example, able mathematicians may not need to get the spreadsheet to do the calculations in order to know the answer to a question that is, for others, answered by changing the variables in a spreadsheet.
- The time spent on the tasks is more getting a feel for how the software works than using it to do things.
- The motivation to un-learn familiar techniques for the sake of this test is not high enough.
- The build up of pressure because of time constraints reduces the inclination to do things thoroughly.
- The actions that are desirable for aesthetic reasons may be awkward to do on the software and lead to more time being taken than was intended (e.g. retaining the proportions of images when resizing them).
- A focus on quality of product (possibly as an interpretation of the test, but also as an expression of personal values) leads to time being spent that does not enhance the assessment outcome at all.
- The relationship between the real task (the avowed purpose of the activity, what has to be achieved) and the ICT task (the means to achieve the purpose, using ICT) is such that being good at the real task might not correlate with being good at ICT e.g. pupils who are good at spelling do not need to use ICT facilities to help them get words right, and those who have poor language skills may not recognise the correct word among the list offered by the spell checker.
- For each pupil taking the test the doing of one thing rather than another to achieve a task is matter of habit / personal preference / perceived convenience (easy to remember / easy to do) rather than being a reflection of their ICT capability.

- The task is confusing, and the pupil gets lost (visually or conceptually), and as a result wastes time, or even gives up on the task.
- The pupils (for whatever reason) are or become confused frustrated, dispirited or de-motivated and do not engage in a way that leads to taking up the opportunities.
- Lack of knowledge of some specific actions to achieve certain things have disproportionate effects, such as precluding all progress with the task.

2c. SoDs in the pupil's experience that affect their thinking, and hence their actions.

SoDs in the tasks that affect the pupil's thinking

- The pupil's thinking has to keep switching between organising the environment and deciding what to do.
- The tasks require too many things to be remembered across a sequence of screen / actions.
- The need to keep switching between objects or parts of objects that cannot be seen at the same time mean that pupils become disorientated, and lose track of where they are in the task.
- The complexity of the tasks mean that it is hard for pupils to have a sense of progress and what they have done and what there is to do.

SoDs in the tasks that create confusion

- Ambiguity in the task, because there is not yet enough familiarity with the new format to make an 'educated guess' about what is meant.
- The memory demands of the tasks (e.g. to hold information from one screen to another; or to remember where to go back to having completed a sub-task).

SoDs in the tasks that generate frustration

- The tasks ask questions that cannot be answered unless assumptions are made that not everyone would be willing to make, e.g. about the reliability of web-sites, when there is no opportunity to validate the sources in the usual ways.
- The action required to do something is not obvious, is not known from other software, and was not covered in (or remembered from) training in the software.

SoDs in the task that undermine motivation

- There isn't a systematic progression towards a clear goal.
- The tasks are confusing.
- There is too much (apparently pointless) repetition.
- There is not enough to enjoy in the assessment in order to sustain engagement.
- The authenticity in the task is at the expense of lack of clarity in the instructions. The potentially improved motivation from the former is overwhelmed by the loss of it through frustration arising from the latter.

SoDs in the software environment that affect the pupil's thinking

- The unfamiliarity of the environment makes pupils want to do the minimum to achieve the task.
- The demands of tasks in the context of the structure of the software leads to too great a call on memory when switching between screens, or when trying to get back to an earlier state.

- Mistaken actions are taken that lead to disastrous consequences (the screen disappears, the data is lost) and there is no recovery, either because the software has no facility for it, or the pupil just does not know it.
- The pupil takes up the opportunity of the rich environment to explore without purpose.
- The pupil is overwhelmed by the complexity of the rich environment, and is paralysed into inaction, waiting for something to come to mind that will say what to do.
- The number of different things that can be done at any time, the alternatives for action, require a much stronger goal orientation / task focus than is reasonable for pupils of the age.
- There is an inappropriately high demand for ICT knowledge and skill that is not relevant to the assessment.
- The short-cuts that work in the software are not those that the pupil knows.

SoDs in the software environment that create confusion

- The cramped and busy screen leads to confusion.
- When sub-tasks are on different screens, deciding how to do a sub-task can remove from the mind the overall purpose of the task, leading to a need to keep going back to check out what the task is – and lose track of what had been done in the sub-task, resulting in confusion.

SoDs in the software environment that generate frustration

- Actions fail that were assumed to lead to a desired immediate objective.
- In the use of drop-down menus, there is no option for "other", leading to frustration when none of the given options seems right.
- The software does not enable the pupil to prepare something to the level of their usual standards in the time available.
- The most capable pupils feel the limitations of the software most strongly, because of the higher expectations they come in with.

SoDs in the software environment that undermine motivation

- Pupils' known short-cuts / idiosyncratic approaches don't work.
- Pupils who are used to achieving their purposes quickly and efficiently may be disappointed and frustrated by the limitations of the software.

SoDs in earlier experience that affect the pupil's thinking

- The pupil has always used conventional tests, so has an expectation of tests being opportunities to show off knowledge.
- Inexperience in the kinds of tasks being done means that pupils do not have an organisational strategy for doing the task, so the ICT activity that would fit into this schema has no-where to go, so does not occur.
- The more sophisticated methods of approaching the tasks are not easy to achieve with the available tools, and require a purpose beyond completing the task well.

These possible sources of difficulty became the 'hypotheses' to be examined by reflecting on the observation and interview data.

3. Data

Observation data
The observations gave rise to 65 sets of observation notes, each covering a test session, together with notes from the interviews with pupils and teachers, including 26 recorded pupil interviews and 20 recorded teacher interviews.

Examples of each follow:

Example of observation notes (part)

Time remaining	
50	Opens and reads Task 1 – leaflet
48	Enlarges screen and splits it. [Says later this was learned from earlier experience with software] Competes the task without difficulty. Is using control-c and control-v to copy and paste [later says "A friend showed me."] Uses right click to change fonts.
46	Uses object location to place the picture. [Says later this was discovered in earlier practice sessions – "I like to try things."]
44	Moves around the screen, opening and closing files confidently.
40	Knows to save the completed task in the appropriate location.
39	Goes back to the email to check each job in the task has been done.
37	Moves on to Task 2 – Spreadsheet
36	Maximises screen again. Uses care when reading
34	Scrolls up and down. On third one keeps scrolling up and down [Later says "Kept double checking it. Not concentrating."] Eventually arrives at answer.
31	To fill in a formula, puts = and pauses. Does "= sum (C9-600)"
29	Is not happy. Reads again, highlighting some parts of the text. [Later says "Not sure if the formula is right."]
28	Moves on to next part of the task. Goes up and down, checking.
26	Looks at fifth, then sixth.
24	Completes the task, and goes to the email to remember who to send it to.

Example of pupil interview (part)

R: So given that was the first time you did it, I thought you really did brilliantly with it actually

P: *Thank you*

R: But how much is it like the sort of things you normally do in school?

P: *A bit, well we do, do a bit of like the emails and attaching things but not that much, but we do a lot of moving things round and copying things and pasting them in new places so that, making the advert to start with was easy and we have done the Excel formulas but it was in Year 7 so it was two years ago and I can't really remember much of that stuff but.*

R: So when you were trying to do the, let me have a look back to all my sheets, when you were trying to do the adverts and you were, were you having a bit of trouble finding the map?

P: *Yeah, well no because it was just I didn't read the question properly and it wanted the colour map and I was looking at them and there was lots of them so.*

R: And what was happening when you were doing the spreadsheet, there was a lot of going backwards and forwards wasn't there, reading up and down the page?

P: *Yeah I just had to keep reading it because I didn't understand that the £1.75 per customer, because I thought I had to times the amount of customers by £1.75 but that put me like in a big loss of money so I just kept reading it but I'd finished so I don't have to worry about that bit until next time.*

R: And when you were doing Question 2, I remember a lad came over and gave you some help with it and he suggested you do something

P: *Yeah*

R: What did he tell you to do?

P: *He said um, I can't remember, when we were doing the spreadsheet?*

R: Yeah with Question 2, you were having a bit of trouble trying to work out what to do with it weren't you?

P: *Yeah, well I can't remember which was Question 2?*

R: When someone came over, the one about changing costs I think.

P: *Oh right yeah. I didn't really understand the table very well because but, um there was two totals but one was for June so I didn't understand how to do it but then I do now.*

R: But you do now?

P: *Yeah*

R: So what would you have to do with it?

P: *Um I had to change the amount of customers and it's just like trial and error and I'll have thought how to work it out somehow.*

R: Oh right so that's what he suggested you do?

P: *Yeah*

R: He suggested you change it until you got it right?

P: *Yeah*

Example of teacher interview (part)

R: How much time have you spent with the software, you and the students, the Year 9 students

T: *We didn't do the familiarisation one, the very first one. We went straight into the Practice 1 and Practice 2. So that's two hours, well two 50mins sessions that we spent, and this one now is the mock exam.*

R: So how have they adapted to the practice tests?

T: *Pretty well really, yeah pretty well*
R: How does it fit in with things that they normally get, how similar or different is it?
T: *The interface is obviously different, quite a bit different, and that throws them a little bit at first but once they get familiar with clicking on the icons and they can see what's coming in and reading the emails, because the email is different as well and attachments are slightly different, but it's just getting them familiar really with what they need to do to be able to see what's going on. The database one's a little bit, a bit iffy as well, but that's not…*
R: Do you mean it's more different to what they usually use?
T: *It's slightly different but we haven't done an awful lot of database work.*
R: Because it's quite like Access isn't it?
T: *It is yeah, but we haven't done an awful lot of Access this year which is something that we need to address really for the future that, we will do that. I did change the syllabus around in view of last year's tasks and we used to do Access right at the very end. I've now brought it more to the beginning but I think what we need to do is maybe put it in the middle because all the stuff they did in September they've forgotten. So, and it is quite a nice part of it, the things like the presentation and the leaflets and things like that, they don't seem to have too much trouble with.*
R: Are the tasks and the nature of the tasks, similar to how you would have been working?
T: *They're not too bad. Because I'd seen them last year, we were aware of what's required. We do something here called Dig IT, the Digit software we use in Year 9, well right through actually, 7, 8 and 9 and we find that that does help them to produce things that are together, you know using the email, to posting an email to friends, getting it to adapt and different things like that. So it's not something that's completely alien to them which it was last year, but that's because I know what I need to do now a bit more. So I think really in many ways it's a, it's a learning curve for us all and if we were taking an exam, we would want to know what the format was and what we needed to do.*
R: Oh yeah
T: *And I think that's the difference really is that nobody's known what it's going to be like.*
R: From your point of view of working in charge of IT, does the way in which the software work and the way in which the tasks work address issues that are important in ICT
T: *Oh I think so, I think so yeah. I mean somebody actually said to me the other day that this, the tasks are far more real world and maybe a little bit divorced from some of the things that we actually do in school. So what we knew to do, is to become more integrated with the software that we use. I think in this school we're quite at the forefront of teaching ICT in the way we go about things. I know a lot of schools don't even allow their kids to have emails, so how can they possibly do this test if they've never used email and it's a really big issue, I was talking to a guy last week at one of our Head of Departments Meetings and he said well we don't have email, and he's had to set them all up with an account and had to run through it very quickly. Well you know, it's ridiculous in this day and age that kids aren't allowed email, but it's a security issue for them and they don't want to go down that road. But yes I would say the tests are designed far more real world and maybe far more real world than some teachers at some schools would, we're doing. But I think once we know that this is what we've got to do, then it's going to become far easier, far more easier, and certainly we do Goal tests as well which, I don't know if you've heard but Goal*

software, G O A L, it tests their levels, very much I'd say a multi choice type scenario and interface, but even they are changing that now to reflect more the testing that's going to be required, so it's going to be rather than just click on a number or a box to say this is the answer, this is the answer, they're having to do things, they have to draw that, draw a face that's smiling and it works out the fact that they can use the mouse properly and things like that. So very simple things but very important.

Quasi-protocol data

In addition to the observation visits, there were visits to school to undertake quasi-protocol sessions, and these gave rise to 33 recordings of pupils talking aloud when taking a task.

Think-aloud is not always revealing. Some of the reported activity is mundane, such as reading. Also what is being done is not always reported. One kind of situation where this particularly applies is when talking about an action interferes with doing it – for example when pupils were scrolling up and down trying to retain in memory the information that is needed, for example in a spreadsheet task they became silent. Another kind of situation where there was limited commentary was when the actions of the pupil when undertaking the test were exploratory – trying things out to see what happens, and looking for something to trigger an idea about what to do to achieve a purpose – it happens too quickly and is too embedded to be talked aloud, and pupils of this age are unlikely to have the self awareness and confidence to continue doing it and give commentary on it along the lines of "I am just playing around to see if I can figure out what to do" – so again they became silent.

As a result, the quasi-protocol recordings could not be said to give a full account of the activity of task completion – and were in many ways a thinner account of it than the observation notes. However, their difference from observation notes offered scope for different insights into some of the underlying sources of difficulty in the tests.

Example of quasi-protocol transcription (part)

Ok. So I'm reading the email and I'm just looking for what it tells me to do and it's telling me to fill in the empty cells in the table that are on the attachment, so that's No. 1 and there's two things to do, you have to find the height. So I'm looking on the internet and trying to find the things. [Pause] *And I'm splitting the screen just so I know what, I know that I have to get from the internet.* [Pause]

What are you doing now?

I'm trying to find a site that'll tell me the information that I need to fill in the table and I've gone on this website 'cos it looks quite interesting and it might tell me the things that I need, but the website doesn't tell me what I need to find so I'm going to look for another one. I've clicked on another one 'cos it had the words that I need to find. [Long pause]

What are you doing now?

I'm trying to find the information to fill the table in but I can't find it so I'm going to look back on the internet and see if I can find it. I'm on the website now trying to find the information so I can find out and fill in the table. [Long pause] *But I still can't find the screen that I'm looking for. I can't find anything on that website so I'm looking at this website.* [Long pause] *I'm still trying to look for the website but I can't find anything that'll tell me the information to fill in the table.*

4. Analyses and findings

The observation data were coded in relation to the hypothesised SoDs. This involved a judgement by the observers, informed by their interviews with teachers and pupils. This is the evidence that was used in analysis.

Given the design and overall approach of the work, there is only limited scope for quantitative analysis of the data in this project. There is variation in pupils, schools and tests observed – the last because the randomisation of task context, content, and order for different pupils means that even within the same session and tier of a test administration there was considerable variation in the activities actually observed. As a result, a very much larger scale project would have been necessary to establish statistically significant patterns of behaviour related to generalisable sources of difficulty. Nevertheless, it is reasonable to take a general overview of the observation sessions in terms of consistency in the kinds of SoDs recorded. We have to be careful not to over generalise in doing this – since the fact that a particular SoD didn't happen to be observed at all by the project does not necessarily mean that it would not appear as a significant source of difficulty in a larger scale study.

It seems self evident though, that if a theoretically derived source of difficulty were recorded by many different observers and across many sessions, then this might reasonably be regarded as a real, and not simply theoretical, phenomenon.

The nature of the evidence

Evidence for particular sources of difficulty is rarely direct – the direct evidence usually only shows that there is difficulty, and although one can presume that it has a source, it is not generally clear what it is. Pupils can be seen to be having difficulty, for example, by failing to engage with or successfully complete a task, perhaps taking multiple attempts to achieve a purpose, becoming visibly frustrated or confused, or trying things that do not achieve their purpose. Occasionally also pupils express (in different ways) their own view that they are having difficulties.

We can then speculate about the source of the difficulty, but ultimately how does anyone know what causes what? One can observe pupils being confused, but what caused that? You can see a potential cause of confusion at the same time, and can infer that one caused the other, and this can be 'confirmed' later in interview with the pupil – but you cannot actually say that the evidence shows that one caused the other. The evidence is indicative, the process more a matter of warranted judgement.

When a pupil does not do what the assessment is looking for, it may be because there is a barrier in the question such as language that is not understood or too great a demand on working memory, or it could be that they did not think to do it because of limitations in preparation for the test, or perhaps they are unable to do it as they lack the necessary skills and knowledge. One cannot tell from the direct evidence – which is just an observation of them not doing it – it has to be a judgement, requiring inference that is not justified by reference solely to describable behaviour.

The evidence for SoDs is therefore the judgements of the observers – some made straight away, during the data collection, but also some that were made later, informed by contextual information, including what the teacher has said, or what was seen to occur at another time. However, there were also many cases of difficulties being

observed without a SoD being attached to them, because there was insufficient indirect evidence for them to be able to come to a judgement about it.

Nevertheless, it needs to be accepted that in some cases the amount of inference involved in evidence 'for' a SoD will be relatively high. This is justified, in this project, by the constructive purpose of the enterprise.

In the observation context, there is access only to pupil behaviour, observed actions and postures, verbal reports and so on, and the process of inferring from these the sources of difficulty in the test relies fundamentally on judgement by the observer – but observer consistency is not something that can be guaranteed. Each observer brings to the process of observing a set of understandings, presuppositions, ideas and so on which affect the experience from their point of view, so that what they notice and record is going to be different for each individual. In a project with a longer time frame, early observations can establish a framework for observation which is used in subsequent observations, often in the form of a checklist or set of agreed categories. In this project, operating as it did in an extremely narrow window of opportunity as the 2006 tests were being trialled, there was no opportunity to use this approach.

For any observer it is perfectly possible to be mistaken in the attribution of a source to the difficulty. There could always be several possible sources for any observed difficulty. For example, using inefficient approaches could be sourced in the task, arising say from a misleading instruction, or it could come out of the software, if for instance the more efficient approaches are difficult to realise in the particular application, or it could be sourced in the pupil, a manifestation of their view that it is more important to get the correct answer than to follow a high level process.

So, in the absence of corroborative evidence of some kind, no observer can say with full confidence that a given pupil had a particular difficulty as a result of a specific source. However, we can say that we thought they did, and if a number of observers are each persuaded of a particular source for difficulties across different observations, then our confidence can be high that that source of difficulty for the tests is real.

The issue of observer consistency is thus being addressed in this project by the principle of inclusion – noting when all or nearly all of the observers felt that they were observing an example of a described source of difficulty, presuming that this means that there is something substantial behind that, and attempting from the examples observed by the different observers what the underlying issue might be.

Where fewer observers feel that they have seen examples of that described SoD, the reporting will be very cautious, accepting the possibility that it may have arisen from particular pre-dispositions in those individuals.

The range of observed actions that was felt to arise from a particular source also allows us to say something about the scope of it (whether it is in different tasks or just one kind; whether it seems to affect a range of pupils or just one kind; whether it manifests itself in a number of ways or just one way; and so on) and we can also make judgements about the degree of impact that a source of difficulty seems to have in terms of depressing the assessed performance of pupils.

This judgement is complicated by the fact that the effect of the difficulty on pupil action can be indirect, through its effect on understanding, on state of mind, or on the time available in the test – which itself can have an effect on state of mind.

It should be remembered in all this, however, that some of the possible sources of difficulty are legitimate, are part of the correct operation of the assessment. Tasks do need to make demands on pupils, the software does have to have features that need to be known, and the pupils do need to have ICT capability. If pupils are not succeeding because they lack the qualities that the assessment is trying to assess, then no further comment is needed. This cannot be determined of itself, because there is no independent measure of the qualities beyond the test.

Case studies

There now follow three case studies of pupils engaging with the tests, to illustrate the nature of the observed behaviours, and inferences about motives and thinking that could be made – prior to judgements about sources of difficulty.

The first two are studies of individuals, which follow a pupil through a test session.

1. A pupil with moderate skills – pre-test, higher tier, first session (tasks 1-3)

Pupil clicks 'proceed' and begins the test.

Pupil receives an email containing task 1, which comprises a set of questions. He reads the instructions for the first question, which is about a database, but he is not sure what to do. He goes to the database and covers all the icons with the cursor, as if looking for a connection that will trigger an idea of what to do. He goes back to the email, and hovers over "best data type for each of the fields" but is unsure what this means.

Pupil goes back to the database and opens fields, but is still not sure. He tries different buttons hoping to find an answer in the software.

Pupil goes to 'file manager', looking for a different 'Task 1' because the other one makes no sense to him. He goes back to the original, and looks at the fields. He calls the teacher over, who gives support without telling him what to do. Pupil makes a choice for the first and second, but is not at all sure what the others might be.

Pupil finishes the task, but is reluctant to leave it, because he is unsure that he has finished. Eventually he moves on to the second part of task 1. He is twelve minutes into the test.

Question 2 is about 'Logo' type control software. Pupil scrolls up and down, up and down because he can never see all that he needs to see. He does not enlarge the visible screen. He chooses an 'answer' and goes back to question 1, then back to 2 then 1 again.

Pupil moves on to question 3: a spreadsheet. He is unsure how to add a row so he opens and closes menus – very quickly – to try to remember. He cuts the bottom two rows, and pastes them in one row lower. He types "Trousers" in one cell but hovers over the other cells. Pupil restores the font size of the two moved rows so that the spreadsheet looks all the same because he likes it to look right. Pupil scrolls down and finds the values for trousers and types them into the row.

Finally, Pupil changes the totals to what they add up to, without bothering about formulae. This part of the test has taken seven minutes to complete.

Pupil moves on to question 4 of task 1, which is a multiple choice item. He chooses his answers very quickly, spending less than a minute on this task

The next task (1/5) is about a database. Pupil reads just some of the instructions, and then finds the file easily. He scrolls up and down the data looking for the error that he has been instructed to find. He does not find anything. He goes back to email then to the data file then to the email then back to the data file then back to the email and again to the data file where he deletes a row. He clicks repeatedly on the empty row. Pupil scrolls up and down the data and changes "Aberdeen" to "9" because he thinks that the values have to be in order, but then changes "9" to "12". This task has taken five minutes.

The final part of task 1 consists of multiple choice questions. In little over a minute he chooses options 2, 5, 2 and 1 and then changes the first to 4.

Task 1 has taken almost 30 minutes to complete, out of a total time of 50 minutes for the whole test.

Pupil goes to his inbox, opens the email for task 2, goes into the presentation, looks at each slide, then goes back to the email and deletes the task 1 message because he finds it confusing.

Pupil opens the presentation again. The screen is getting very cramped, so he has to keep scrolling to see the whole of even one slide. He repeatedly clicks on things and looks at what comes up. Pupil goes into the web search page, types in the whole question that the task is asking him to answer, and receives a list of 287 sites. He opens the first web page then closes it and briefly scrolls down the list. He has spent four minutes on this task so far.

Pupil goes back to the presentation. He scrolls around on a few slides then switches to slide sorter view in order to see whole set of slides. Pupil cannot read the print in slide sorter view, so goes back to slide view, and scrolls around again. Pupil goes up to the earlier slides and tries to cut and paste the text that he needs but it doesn't work. He is working with the content of the previous slide rather than with what the question is supposed to be about. He uses the action button instruction and types in text, linked to a slide. He repeats this with the other elements (from the wrong slide). Pupil takes time sizing the box with the information in it so that it looks right. Pupil looks at the next slide and wonders what to do. Pupil reviews the information again. He then uses the action button instruction to get text into the slide. Pupil tries to copy and paste a graph into a slide. It does not work at first, but then it does. He has spent nine minutes on this part of the activity.

Pupil changes the labels on the graph, and the axes. Pupil tries to save the presentation, but fails, first because he does not change the name, and then because of trying to save it in 'my computer'. He finally succeeds in saving the presentation and goes on to the email for task 3, but there is hardly any time left, so he just waits for the end of the test.

2. A pupil with modest skills – pre-test, lower tier, second session (tasks 4-5)

Pupil starts the test with task 4 – a database about a survey. He immediately goes to the toilet so the test is paused. When he returns from the toilet he clicks to cancel the pause (thus losing the time – seven minutes) and opens the email about the database task.

Pupil opens the presentation and then opens the database. He goes back to the presentation to check the question. He goes back to database and scrolls across to the appropriate column. He scrolls down the column (counting). He goes back to presentation, and enters an answer. He looks at next slide.

Pupil goes back to the database and opens the 'edit fields' view. He chooses 'list' because he is looking for an automatic count. He goes back to data sheet, and scrolls down counting the cells. He goes back to 'edit fields' view, opens and closes elements, but he does not change anything. He then goes back to the datasheet and scrolls down, counting again. He goes back to the slide, pauses, takes a guess at the answer and enters it. He goes into the database, opens 'edit fields' view again, pokes around again, but again does nothing. His neighbour whispers something, and he finds the 'Sort' button, but doesn't do anything. Pupil goes back to the data sheet, then back to 'Sort' and applies a criterion. He goes back to the slide and changes his previous guessed answer. Since returning from the toilet he has spent nine minutes on the task.

Pupil goes to slide 3, and opens 'insert', apparently looking for 'insert graph'. Pupil is told by his neighbour to look in 'chart' in the database. He opens and goes into 'design chart' and adds labels to what is there. He pauses over 'chart title' but moves on. He tries to save the chart, but cannot, and does not know why so he puts his hand up. The teacher tells him to use copy and paste. He says "Oh, yeah." He copies it and tries 'insert' again until his neighbour suggests "right click paste". "Oh, yeah", he says. Pupil re-sizes the graph and makes space for it among the text. This has taken him five minutes.

Pupil looks at the next question. He puts his hand up. The chart label does not include all labels, so he can't see the information he needs. The teacher cannot help. Pupil goes back into the database, and uses sort again, but to no avail. Pupil answers on the basis of the biggest one that does have a label, even though he knows it is wrong. This has taken two minutes.

Pupil goes to next slide. The connection with the server fails and the test crashes. When the problem is sorted, no time has elapsed on the test (The teacher is relieved). Pupil reads the question and goes to the database where he goes into 'edit fields' again, pointlessly. Pupil goes to the data sheet, into 'sort' where he sorts by one criterion, and scrolls down; but the list is mixed in relation to the second relevant criterion. He sorts by the second, but that re-sets the first, so it remains mixed. Pupil asks the teacher but the teacher cannot help. Pupil goes back into 'edit fields' and tries 'list' again but it is fruitless. Pupil sorts by the first criterion again then by the second again, and again it does not work. Someone (at another computer) tells him what to do, and he does it, getting a properly sorted list. Pupil scrolls down the list and counts. He goes back to the slide and enters a figure. He has spent nine minutes on this slide.

Pupil reads the next question in the task, but is unsettled and is unwilling to do more of the same so he comes out of the presentation without saving it. Altogether, he has spent 36 minutes on the first task of the test.

Pupil looks at next email with a task involving creating a leaflet. He reads it half-heartedly, looking around him. He chooses three of the listed options in the first part of the task, and saves it – there is a reminder to do so at the bottom. Pupil finds the templates file and looks at template options; he needs to scroll to see them. Pupil chooses one option. He pauses over the title and writes nothing.

Pupil goes back to the email, and follows the 'web' link. He hovers over things. He tries to save but fails. Someone suggests that he uses 'save as', which he does. Pupil changes the title and tells his neighbour about it. He has spent eight minutes on this task so far.

Pupil finds an image in 'clip art' and pastes it in. He gets the another image in 'clip art' and pastes that in too. He complains there is no other suitable clip-art. Someone suggests looking in 'images'. Pupil finds an image and pastes it in, then another. He hovers over a text space, but does not write anything. He is unsettled because the class is being dismissed (he is 5 minutes behind the rest of the class because of the time he lost when his computer lost contact with the server.)

Pupil does not save what he has done, just leaves with the rest of the class. The test session ends but he isn't there any more.

3. The 'story' of the leaflet design task: summative test, second session.

This is a composite of the actions of several pupils observed undertaking the leaflet design task in the second session of the summative test. It demonstrates a range of actions observed during the task. The records of several different observers were used.

Pupil opens the email which tells her to create a leaflet. The email contains several instructions about the leaflet: style guide, headings, target audience, images and web sources of text.

Pupil opens the attachment. She returns to the email to check the instructions. She goes to the word processor, clicks on 'open document' and opens the sample leaflet.

The email tells her to open 'File Manager' to look for a template. She expects to find 'File Manager' on a 'file' menu in the word processor or web browser. She opens file manager from the toolbar, searches for the style guide and opens the file.

Pupil goes back to the email to re-read the instructions. She goes to the file manager and opens all the template files (there are six). She goes back to the style guide and re-reads it. She chooses one of the templates.

Pupil splits the screen so that she can see both the email and the template leaflet. She scrolls up and down the email.

Pupil goes to the file manager and looks for images. She finds several image files. The names do not provide much information about the content so she opens and

closes each file in turn. She opens some of the files more than once. She chooses and image and tries to insert into the leaflet. She fails because she unknowingly tried to insert the image into a text box.

Pupil tries again to insert the image, this time successfully. She returns to the email and scrolls as she reads the instructions in the narrow split screen. She decides that she has the wrong image and looks for a more appropriate one. She opens all the image files. She opens some of the image files again. She receives an error message saying that she has too many windows open. She returns to the email and re-reads the instructions.

Pupil closes all the image files. She opens an image and closes it. She opens another image and inserts it into the leaflet.

Pupil goes to file manager and reads the style guide again. She re-reads the instructions in the email.

Pupil types a title and selects the text. Pupil opens the font dialogue and selects a font size and makes the title larger. She changes the colour of the font.

Pupil compares the new leaflet with the original sample because she needs to make the formatting the same on the new leaflet as on the sample. Because of the spilt screen, she finds it difficult to see both at once.

Pupil needs to find text for the leaflet. She goes back to the email and clicks on a website address. She reads the website, moving her mouse along the text as she reads and uses the 'stovetop' button to navigate back to the leaflet. She adds some information to the leaflet, typing from the website (she does not copy and paste the text). She selects text, goes to the font dialogue and changes the size of the font. Not all the text had been selected to she repeats the font change with the remaining text. She moves the text so that it is under an image. She swaps the image and text. She adjusts the image position. She moves text into a new position and changes the font.

Pupil clicks 'print preview'. At first, she can't close the 'print preview' window but eventually she succeeds.

Pupil selects a text box and types in a bullet point list, copying the text from the sample leaflet. Pupil tries to save her work using 'save as' but can't as the file is protected. She tries again with a different name but the name she chooses contains forbidden characters (apostrophe). She tries a third filename and is successful.

Pupil goes back to the email and writes down the email address. Going back to the leaflet, she clicks on 'send as an attachment' and when the new email page appears she types in the email address and the subject and sends the email.

Pupil goes back to the original email again and opens the 'leaflet evaluation form' She ticks three items. She goes back to the email and sees the instruction to save the leaflet. She saves it and opens the email for the next task.

Support for the hypotheses

In the process of examining the hypothesised SoDs through observation and reflection there were new SoDs added to the framework which had not been anticipated. Also, some of the hypotheses were re-phrased, some were split, and others were combined, where it was not possible to distinguish between them.

A distinction was also made between sources of difficulty which related to preparation for the tests (or to be precise, lack of it) and others. In the lists below, the 'preparation SoDs' are listed second.

An initial analysis was made to establish which of the revised hypothesised SoDs were observed relatively often. This was based on whether or not each of the SODs in the framework appeared at least once in an observation session. It should be noted that this ignores the prevalence or persistence of a source of difficulty within any particular observation since, although information on these can sometimes be inferred from the observers' notes, these were not the main observational focus.

The most commonly occurring, the top third of all the hypothesised SoDs, and all being observed in at least one fifth of all the observed sessions, were:

- The task instructions are not explicit enough about what has to be done.
- The task is ambiguous, or vague.
- Limitations in subject knowledge (e.g. mathematics) constrain engagement.
- There is too much (apparently pointless) repetition.
- The requirement for sustained concentration is too great for many pupils, whatever their ICT capability.
- Pupils know enough to succeed in the tasks without using ICT for all the steps.
- The pupils (for whatever reason), get lost (visually or conceptually), and as a result waste time, or even give up on the task.
- The pupils (for whatever reason) are or become confused, frustrated, dispirited or de-motivated and do not engage in a way that leads to taking up the opportunities.
- The tasks require too many things to be remembered across a sequence of screens / actions.
- The need to keep switching between objects or parts of objects that cannot be seen at the same time mean that pupils become disorientated, and lose track of where they are in the task.
- The memory demands of the tasks are too high.
- There is not enough to enjoy in the assessment in order to sustain engagement.
- There is an inappropriately high demand for ICT knowledge and skill that is not relevant to the assessment.

(Preparation SoDs)

- The pupil has always used the same software, so her proficiency is context bound.
- The unfamiliarity of the software leads to multiple attempts to achieve a purpose.
- The pupil wants to undertake a particular action but does not know how.
- Pupils spend time making something nice rather than getting on with the test, e.g. trying several different fonts to choose the most appealing.
- Lack of knowledge of some specific actions to achieve certain things has disproportionate effects, such as precluding all progress with the task.
- Actions fail that were assumed to lead to a desired immediate objective.

- The cramped and busy screen is confusing.

A different perspective on the evidence is to take as significant those SoDs that were observed by a majority of the observers (at least 5) at some time.
These were:

- The task instructions are not explicit enough about what has to be done.
- The language of the task is difficult.
- The task is ambiguous, or vague.
- Limitations in subject knowledge (e.g. mathematics) constrain engagement.
- The demands in the interaction between tasks and software on short-term memory and organisational skills are inappropriate to the level of ICT capability that is being assessed.
- There is too much to do in each task for the time available to do it.
- The requirement for sustained concentration is too great for many pupils, whatever their ICT capability.
- Pupils know enough to succeed in the tasks without using ICT for all the steps.
- The pupils (for whatever reason), get lost (visually or conceptually), and as a result waste time, or even give up on the task.
- The pupils (for whatever reason) are or become confused, frustrated, dispirited or de-motivated (see below) and do not engage in a way that leads to taking up the opportunities.
- The tasks require too many things to be remembered across a sequence of screens / actions.
- The need to keep switching between objects or parts of objects that cannot be seen at the same time mean that pupils become disorientated, and lose track of where they are in the task
- The memory demands of the tasks are too high.
- There is not enough to enjoy in the assessment in order to sustain engagement.
- There is an inappropriately high demand for ICT knowledge and skill that is not relevant to the assessment.

(Preparation SoDs)

- The pupil has always used the same software, so her proficiency is context bound.
- The unfamiliarity of the software leads to multiple attempts to achieve a purpose.
- The pupil wants to undertake a particular action but does not know how.
- Pupils do not think to use ICT to do tasks that they can do without ICT.
- Pupils spend time making something nice rather than getting on with the test, e.g. trying several different fonts to choose the most appealing.
- Lack of knowledge of some specific actions to achieve certain things has disproportionate effects, such as precluding all progress with the task.
- Actions fail that were assumed to lead to a desired immediate objective.
- The pupil is overwhelmed by the complexity of the rich environment, and is paralysed into inaction, waiting for something to come to mind that will say what to do.
- The cramped and busy screen is confusing.

It will be noted that the list of commonly observed SoDs is only a little different from the list of SoDs observed by many observers.

In drawing conclusions about the SoDs that the project has found evidence for, it seems sensible to focus mainly on those that are in both lists, about which there can be greater confidence.
This corresponds to the first list, but excluding "There is too much (apparently pointless) repetition" – which was only noted as a source of difficulty by four observers.

The remaining hypothesised SoDs were noted by fewer than five observers over a total of fewer than 20% of the observed sessions, and so were considered by the project to be insufficiently supported by the evidence.

Of course this does not mean that they are not factors in the test, just that they were not observed or judged to be present to a sufficient extent to be sure about it. Yet even factors with no observations at all relating to them may be there without anyone noticing, and others could occur only under certain circumstances that happen not to have been operational during the observations made. The hypothesised SoDs least supported by evidence were:

- The work-based nature of the tasks is alien and disorientating.
- The software does not enable the pupil to prepare something to the level of their usual standards in the time available.
- The tasks favour convergent thinkers, who take one task at a time, successively.
- In the use of drop-down menus, there is no option for "other", leading to frustration when none of the given options seems right.

It is always desirable to continue to hold possibilities in mind if they seem plausible, even if not so far directly evidenced. However, one can only report on what was seen, and it is sensible to act first on those aspects that are well supported by the evidence.

Further consideration within the project will therefore focus on the implications of the most commonly observed sources of difficulty that were seen by a majority of observers (at least five different observers) across a range of question contexts, and a range of different pupils.

Further analysis

Analysis was undertaken to explore possible differences according to gender, tier, ICT level, academic ability in English and mathematics, as well as across the different tests that were observed. In relation to this analysis, it should be noted that, as indicated in the introduction, the forms of data collected were not designed with numerical analysis in mind, and so absence of significance is not strong evidence for the absence of an underlying difference.

In analysis of differences by gender, two of the featured (preparation-related) SoDs showed significance – in both cases occurring more often for girls:

- The pupil wants to undertake a particular action but does not know how.
- The cramped and busy screen is confusing.

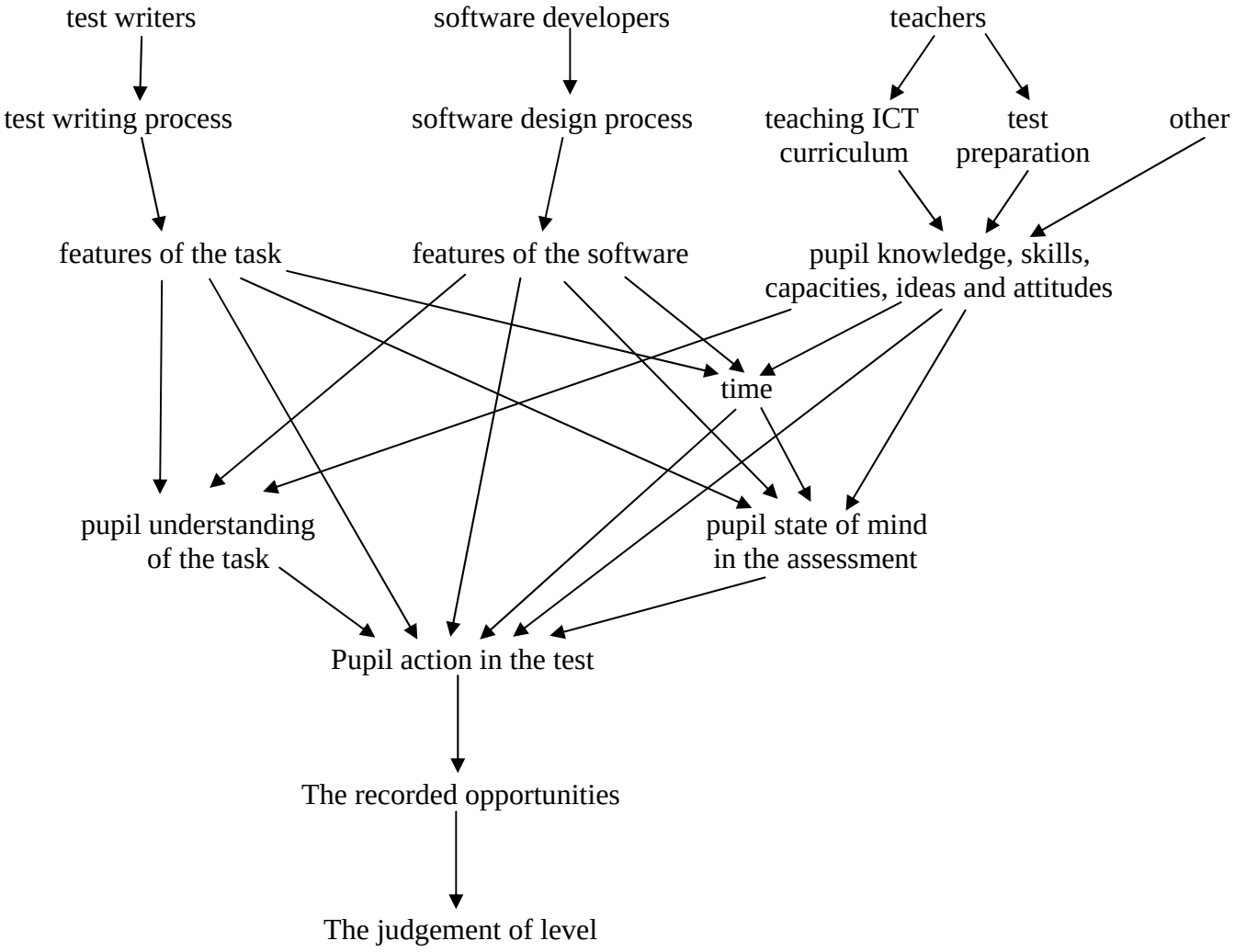
There were no significant differences according to tier of entry, or the level of the pupil in relation to ICT, English or mathematics.

In relation to the test, by and large the SoDs were judged to be present to roughly the same degree in the practice test (the 2005), the summative (2006) test, and the 2006 pre-test. However, the hypothesised SoD "Limitations in subject knowledge (e.g. mathematics) constrain engagement" was barely observed in the practice test, yet featured strongly in the observations of the pre-test (noted in 36% of observations) and summative test (28%). On the other hand, the hypothesised SoD "The memory demands of the tasks are too high", which was observed extensively in the practice test (in 52% of all observations), was not noted at all in the summative test observations.

5. Final framework for sources of difficulty.

In addition to changes to hypothesised SoDs arising from experience, referred to above, the nature of the evidence arising in the project led to a re-consideration of the framework for sources of difficulty. Rather than having essentially the same source of difficulty listed in a number places because its action was routed differently (for example, a SoD about memory issues affecting pupil confidence, and another about them affecting their understanding of the task, and yet another about them affecting action) it seemed more sensible to organise the sources of difficulty first in relation to their context, and only then add the nature of the effects – so that for example memory limitation as a source of difficulty occurs only once.

The starting point for this re-framing was the following diagram showing the interrelationships between the different elements and agents.



The purpose of the SoDs analysis is to try to maximise the likelihood that failure on the test arises from legitimate sources of difficulty, by identifying the most likely sources of inappropriate difficulty – the features of tasks and software that make disproportionate demands, or raise unnecessary barriers, and the forms of knowledge that are irrelevant to the assessment but which nevertheless prevent success.

The intended outcome of the test is that each pupil is awarded the right level of ICT capability, the one that matches their true ability in terms of the national curriculum. In a context in which most of the risk of inaccuracy of assessment is in the direction of 'false negatives' – that is pupils being awarded a level that is lower than it should be – an appropriate analysis to undertake is of what makes the assessment 'harder' than it should be. The concept of 'sources of difficulty' can be introduced as a set of possible reasons for under-performance.

In some ways, and with this interpretation, sources of difficulty can be identified in any part of this diagram. One can say on one occasion that the source of difficulty is in the pupil's understanding of the task, and on another that it is in the test writing process, or in test preparation, or in the judgement of a level.

Thinking about the implications of this diagram for action to improve the tests leads to a reorganisation of plausible sources of difficulty into those in the tasks that relate to the preceding test writing process, those in the software that relate to the preceding design, and those in the pupil that relate to preceding preparation for the tests. Any of these will in practice work through pupil understanding and effects on time and pupil state of mind into actions and these might be observed as different – but have a common root.

Features of tasks
These are clustered under five main headings – language; memory and organisation; irrelevant knowledge and skills; too much to do / too high a demand on concentration; unanticipated interpretation.

Language
Many pupils struggled to carry out the detailed instructions as given. There are a number of possible reasons for this which are difficult to distinguish, even after talking to the pupils about it. For example reading was definitely an issue for some. But do you say that this is because their reading is not up to the task, or that the task instructions are not clear enough? Sophisticated readers pick up the nuance of detailed intention more readily than do other readers.
Nevertheless, it was possible on other occasions to identify two main elements in the main language related sources of difficulty:

- Lack of explicitness either about what to do, or what to do with the 'answer'. For example, pupils could be unsure sure about whether a task was finished, or not know where to save material, or not have a basis on which to make required choices, or know where to look for information that is needed.
- Ambiguous, vague or misleading language, such as what it might mean to "complete a sentence" in this context, or what it might mean to "make similar", as well as specific uses of words, such as the use of "graph" to mean chart (which was also a mismatch between language in the instructions and that used in the software).

Memory / organisation
Of course no task can be completed without some call on memory and organisational skills, so again there is in every particular case an uncertainty about whether the pupil was having difficulty because of an unusually low level of organisation of the

environment and their own cognition (their memory strategies) or because the task was too demanding. However the frequency of this SoD, and the distribution of it across observers, seems enough to say that by and large the tasks make demands that are out of keeping with the object of the assessment. The contexts identified by observers were:

- when information is in one place, but has to be used in another, especially when the cut and paste options are not available or not trusted – e.g. output from queries; information going into a spreadsheet; statistics into a slide;
- when having to hold three criteria in mind in order to construct a query;
- when having to scroll to collect up information from within one application, or to switch between screens to collect up information from more than one;
- when it is not possible to see multiple choice options and the question at the same time, especially if they cannot see all the options at once (to remain clear about which is which); and
- getting lost in a sequence – "goal stacking" (for example, he needs to save then send, but goes to email first, where he cannot send attachment until saved, so he goes out to save, but then forgets to send).

Irrelevant knowledge and skills – i.e. beyond what is actually being assessed. Again it is inevitable that a question will call on knowledge and skills other that what is being assessed, and there has to be an accepted level of 'basic' knowledge and skills below which the source of difficulty is not an issue for the test. However, the observers felt that the demands of the test on knowledge and skills in English and mathematics were greater than it is reasonable to expect.

- Pupils who are reluctant to read (e.g. including but not exclusively pupils for whom English is not their first language) will not do well on the tests whatever their ICT capability, since close reading and following specific instructions is a pre-requisite to success on these questions. The barrier to pupils with low literacy is exacerbated by the fact that there is more to read in the lower tier tasks (possibly from trying to make them clearer). There is also a need for good spelling and accurate transcription in order to ensure that met opportunities are recognised by the scoring algorithm of the computer.
- In mathematics, there are calculation demands in spreadsheet questions and a demand for spatial orientation in logo-type questions that act as a barrier to success.

Of course there is in this the subjective nature of reasonableness. A test that requires level 5 mathematics to score at level 4 in ICT would clearly be unreasonable, but how much could be included and it be considered fair?

Too much to do in the time / too high a demand for sustained concentration
Very few observed pupils completed all the tasks in a steady way over the duration of the tests. Most struggled to sustain good concentration for the full 50 minutes of the test. Many showed signs of fatigue after about 35-40 minutes, all with some questions incomplete or poorly done. Some of these seemed to get 'second wind', especially in the summative test, where they had no option but to see it out. Perhaps as a result of the patchy engagement, most pupils did not complete all aspects of all tasks – at least not properly. Also, some rushed to complete at the end of the 50 minutes, sending anything off, putting anything in. Some just stopped, without completing.

This may, of course, be a by-product of the particular 'legitimate' source of difficulty which is pupils' lacking the competence that would enable the smooth completion of all the tasks well within the time limit.

It could also arise from lack of familiarity, because even pupils with the potential to do well on the test may as a result of unfamiliarity use inefficient approaches, or take multiple attempts to do something, or need time to find out how to use the particular software.

There are also potential effects on time of other sources of difficulty, noted by the observers, particularly:

- misleading instructions;
- complexity;
- unhelpful aspects of the software;
- making unnecessary adjustments, e.g. to give the answer visual appeal.

With all those elements in play, it is very difficult to say whether or not the tests actually do have too much to do in the time. There are not many tasks, but each of them has a number of parts, and some of the parts have parts, so there is plenty to do – and it was too much under the circumstances that prevailed in the 2006 trial. Whether or not that amount will remain too much in the future is not possible to predict. However, it might legitimately be asked whether the assessment actually requires so much in it in order to be able to determine levels.

It is also notable that in some contexts slow hardware had an effect on how long tasks took. Despite the specification requirement stated in guidance to the schools, some were unable to provide a suite of computers suitable for the test that were at that level, and so were slow. Others appeared to operate slowly as a result of their location in a network.

Completing the task in a way that is different to that anticipated in the assessment.
This is a source of difficulty not in the sense of making progress through the tasks – indeed it can actually help in the apparent smooth running of the test – but in the sense of whether the action taken leads to 'opportunities' being identified as being taken, and an accurate assessment judgement at the end.
As the computer can only reward what is anticipated, if pupils do other things, their worthwhile actions may not be recognised as worthwhile, leading to a lower assessment outcome than they deserve. All observers saw examples of this.

Some of the examples of this are really related to preparation – for example where pupils count rather than using tools on the application, or re-type rather than cutting and pasting. Other examples arise from other sources of difficulty, with language issues leading many pupils to interpret the task differently from intended, the most common of which was for pupils to improve in the way they felt to be best, rather than improving in exactly the way specified.

However, there were also specific examples of alternative ways of achieving ends: cutting and pasting rather than using the spreadsheet fill command; moving rows down in a database rather than using the add row command.

Aspects of tasks

A further dimension to the consideration of sources of difficulty in relation to tasks is that many of them can be located in different facets of a task:

(i) Context – the nature of the activity
(ii) Content – the language, numbers, pictures, etc.
(iii) Task instructions
(iv) Visual representation onscreen
(v) Scenario, or 'story' of how the pupil is expected to engage with the task.

For example vague or ambiguous language may be found in: the way the context is described in the email, leading to mistaken presuppositions; in the content, leading to inappropriate values being entered; in the task instructions, leading to the wrong task being undertaken.

Similarly, the demands on organisation and short term memory might apply to any of the five aspects:

1. Context: where the activity has different information in different places, which needs to be collected up and co-ordinated;
2. Content: where the items that need to be held in memory are relatively abstract or complex;
3. Task instructions: where they are in different places, so that elements in one place need to be held in mind and added to with further instructions elsewhere;
4. Visual representation: where the items which need to be held in memory are visually inaccessible (e.g. dense text) or inevitably on different screens;
5. Scenario: where the sequence of actions to be done requires things to be remembered while taking unrelated actions.

For that reason, it will probably helpful to consider the different aspects when considering task design

Software

Two main areas here which arose in the observations are: the level of particular knowledge required; and the unhelpful operation of the software.

The level of particular knowledge required

Knowledge of the interface and of the applications might be thought of as a preparation issue, and familiarity did seem to be a big factor in how pupils responded to the test, but as it is impossible for any pupil to be familiar with all of the features of every application, it is inevitable there are will be things that they do not know how to do that are beyond what it is reasonable to expect. For example in:

- the conventions used in this software, such as b/w;
- how formatting is set up / works;
- how to write a formula in this spreadsheet;
- how to access the header section on the word processor software;
- the syntax of database query fields, which have to be 'just so' e.g. "f" for girl / woman, "m" for boy / man;
- how to retrieve a deleted asset.

The question of what level of knowledge *is* reasonable is hard to resolve, of course, and is all part of the debate about what it means to be capable in ICT. One cannot use

ICT effectively to meet a range of purposes without knowing some software quite well, but how well, and which software, is less easy to resolve.

The unhelpful operation of the software

Some modern software has astonishingly high standards of user-friendliness and reliability. This software falls short of those standards, and has had features that have presented difficulties to users.

- Highlighting text in text boxes can be unreliable and fiddly, e.g. in the presentation.
- Pasting does not always work, especially into text boxes and spreadsheets. Sometimes information has to be saved and copied again from a different source before it can be pasted.
- Sometimes a box can highlight without allowing text entry.
- Text boxes in the presentation can move around for no clear reason.
- Instruction emails can be deleted.
- There are unforgiving requirements about where objects are placed.
- There is no mechanism for keeping track of how much there is to do and where the pupils are in doing it. As a result, when they get stuck they are not in a position to make a good decision about what to do (go on to the next bit or persevere).

Preparation

As can be seen from the diagram at the start of this section of the report, the influences on test-relevant pupil knowledge, skills, capacities, ideas and states of mind going into the test come from a range of sources, including the teaching of ICT and the pupils' experience of computers outside school. Indeed much of what would be listed as pupil knowledge etc. is the ICT capability that the test is trying to assess. Yet as the observations (and common sense) have indicated, this test cannot assess ICT capability without presuming some pupil knowledge, skills, capacities, ideas and states of mind that are not in themselves part of the capability but which have to be dealt with in preparation for the tests.

The sources of difficulty related to preparation that were most strongly evidenced were in three main groups: inappropriate ambitions and consequent actions in the test; persistence with methods and approaches developed on other software; and spending time finding out how to do things rather than completing the tasks in the test.

Inappropriate ambitions and consequent actions in the test

Either from the influence of other tests, or from the influence of 'normal' ICT lessons, a number of pupils seemed to take into the test an unhelpful notion of what they should do.

- Some seemed to assume that this test would like the other KS3 assessment being taken at the same time, and so expected to do well by giving correct answers, by neatness, by showing off their knowledge etc.
- Many pupils took time to do things in what they seemed to think was the 'right' way, such as neat presentation, giving the answers a good visual appearance, or being polite in emails. Sometimes this arose from their notions about ICT, sometimes it was about what they thought is important in tests, sometimes it was just a reflection of their own personal aesthetic.

- Many pupils used non-ICT means to get the required answers, such as counting manually, or re-typing, rather than using the tools in the software. In talking to them it seemed that some do this because their ICT capability is low, in that they didn't know that there are tools that they could use, or felt more secure doing it otherwise. In other cases they didn't know how to do it in this software. In yet other cases they just didn't seem to think that it mattered, focusing on getting answers right by any means possible.
- Some pupils – including pupils who had learned to operate with the software, and were quite skilled in using the tools to achieve purposes – did not when taking the tests do as the questions asked. They looked briefly at the emails, but then interpreted the tasks using 'common sense' based partly on what they had been taught in ICT lessons concerned with that kind of application (to "make it more eye catching"). They improved the leaflet and the presentation with a view to an imagined audience, rather than making changes in line with the unread specific instructions. In doing the improvements they moved smoothly around the application, and the confidence on the computer was quite impressive, but they did not do the specific actions that the software would be looking to count. They did not really under-perform, but the outcome of the test will be a low level.

These problems could presumably be addressed in preparation by giving pupils a better sense of what the test is for, and what characteristics it assesses.

The pupils persist with methods and approaches developed on other software

This general characteristic presumably arose from insufficient time and guidance to get used to the fact that the software will not behave in exactly the same way as the software the pupil is used to. It usually manifested itself in surprise and expressions of frustration when what is tried does not work, subsequently confirmed in interview with the pupil. Observed specific examples, in relation to the spreadsheet, were:

trying to point and click when formula building;
right clicking on a highlighted row to add a row;
double clicking on a cell to allow an insert;
trying to test out a formula to see if it does what it is intended to.

It will probably be corrected by more time spent in becoming familiar with the interface and applications – although there may be attitude issues as well. This will be explored more in the 'advice' section below.

There were quick typists who re-type rather than using cutting and pasting, just because it is faster. They would presumably change if suitably prepared by developing their awareness of what the test is assessing.

There were also pupils for whom the persistence seemed more deeply rooted. For example, many pupils tend to have particular ways of doing things, which they use across a range of settings, rather than a repertoire of approaches that are deployed strategically as befits the situation. Some tried to cut and paste to do everything. This will not be addressed by familiarisation or awareness training. It is more to do with teaching of ICT capability in the first place.

The pupils spend time finding out how to do things.

When needing to use a tool in pursuit of a goal, some pupils just knew what to do. Others tried to remember what to do, and could not, so stopped (and moved on to the next task). However, some pupils who did not know what to do tried things, and since

playing about has benefits, it can suggest ways forward, presumably stimulating by association some link with activity done before, it did in some cases lead to progress with tasks.

It is a source of difficulty because it has a strong impact on the use of time productively in relation to the way the test is marked. Exploratory activity does not trigger any of the opportunities in the test, even though it can be indicative of good ICT capability. Indeed it is perhaps ironic that the pupil who finds such a way to deal with a requirement when he or she does not happen to remember the software command is showing quite good ICT capability, but is not rewarded, whereas the pupil who happens to know and remember the command, which seems relatively low level (knowledge rather than problem solving), is much more likely to score an opportunity.

When pupils know more, they will presumably need to spend less time in the test exploring – although it was noted in some observations that pupil seemed to like to explore computer software, and may do so just to find out more, whether or not it was needed to complete a task.

It might also be noted that, since there is too much to know in the software to avoid exploratory activity within the test to find something out, the pupils with that aspect of ICT capability may be benefited within the test as a whole – so long as there is not so much to do (see above).

A note on the distinctiveness of listed 'sources of difficulty'.
In looking at pupil actions, and finding difficulties being manifested, and trying to interpret that, it is possible at one and the same time to recognise different possible sources of difficulty through the supposed 'path' to that behaviour. For example, one can observe a pupil aimlessly clicking on buttons in the task, which will mean that, whatever their actual ICT capability, they are unlikely to be triggering opportunities that will give rise to an outcome that reflects them. In interpreting that behaviour in terms of SoDs, one might say that the Source of Difficulty (why they under-perform) is not being purposive in the task. But one might also note the pupil's state of mind, and say that the source of difficulty was that they were frustrated, or observe that they felt time pressure, and say that the source of difficulty was not having enough time to engage purposefully with the task, or see that the time pressure arose from ambiguous instructions that had led to having wasted time, so that source of difficulty was the ambiguity, or perhaps the limited language skills of the pupil.

A range of possible interpretations are possible for all observed behaviours, at different points of 'causal distance' (often characterised as a dichotomy between 'proximal' and 'distal' causes). In any actual observation, however, it is likely that one will have been noticed by the observer more clearly than the others, and is recorded as 'the' source of difficulty for that occasion. This should not obscure the reality that difficulties arise from nested factors, each of which needs to be considered when attempting to reduce the likelihood of pupils scoring inappropriately on the test.

It is after all rare that aimless behaviour is entirely the result of frustration, or that frustration is entirely because of time pressure, or that time pressure is entirely the result of task ambiguity, or limited language skill, or the interaction between them.

Pupil state of mind

Within this re-organisation of SoDs some of the most commonly observed sources of difficulty were not directly represented, as they were pupil states of mind in the test, which cannot be addressed directly in test design or preparation. However, it is very important to be aware of them as the key to a successful test. Whatever is done to write tasks, adjust software and prepare pupils for the test, keeping half an eye on their possible impact on pupil state of mind in the test is essential.

In the tests observed, many pupils seemed to develop an unhelpful state of mind within the assessment – all observers reported pupils seeming to be rushed, lost, confused, anxious, confused, frustrated, dispirited or de-motivated. Perhaps some of this (especially anxiety, which can trigger some of the others) is to be expected in high status tests, and may be a direct consequence of the pressure put upon pupils by teachers and others over time and in the immediate run up to the experience, and may be connected to the desire to do well. The issue for this project is whether the less helpful states of mind may have been exacerbated by features of the tasks or software that could be improved in test preparation.

Among possible contributory factors to pupil state of mind, the novelty and challenge of the test is unavoidable, and the ability of the pupil is also a given, with pupil experience and attitudes variables that are under the control of the teacher through preparation (see next section), but it is possible to identify, within the SoD framework, a number of features of the tasks and software that were felt to have an impact on pupil state of mind.
The examples of this which were highlighted by a majority of observers are:

- vague or misleading language
- a lot to read on screen
- a level of complexity that requires split attention and makes a heavy demand on organisational abilities
- a lot to do, especially when at a high level
- demands on other subject knowledge, especially mathematics
- when actions that pupils assumed would lead to particular consequences did not do so
- no easy way to know where you are in the task / test

Beyond this, the absence of enough sources of enjoyment in the task to sustain pupil engagement (sometimes manifested in the form of the outward signs of ‘boredom’) was judged as a major source of difficulty in the test. There are many different possible causes of this, and these include ‘legitimate’ ones such as lack of knowledge of what is being tested, but there may also be features of tests that could be looked at in an attempt to improve this aspect, such as the number of repeated opportunities to take up essentially the same opportunities (the repetition ‘SoD’ that was not quite in itself sufficiently widely observed to be in the task listing above).

6. Recommendations

The recommendations for what to do to try to improve the test derive directly from the revised framework for sources of difficulty, since each of the sources of difficulty described above has been evident in observations made in this project, and since their organisation was done with remedial action in mind. However, it is important also to consider which of the sources of difficulty to give most attention to in order to bring about improvement, and this needs to consider the level of impact that each source of difficulty may have. A less common source of difficulty that has very strong effects may be more important to try to deal with than a commonplace one which has only minor effects. For this purpose, impact was judged by the observers as a combination of two factors:

- Scale – which sources of difficulty affect just parts of tasks, and which whole tasks, and which the whole test;
- Strength – which sources of difficulty have major effects and which only minor effects.

The sources of difficulty that are well evidenced by this project, ordered by degree of impact as judged within the project, are as follows (with highest impact first)

- The requirement for sustained concentration is too great for many pupils, whatever their ICT capability.
- The pupils (for whatever reason) are or become confused, frustrated, dispirited or de-motivated and do not engage in a way that leads to taking up the opportunities.
- The pupils (for whatever reason), get lost (visually or conceptually), and as a result waste time, or even give up on the task.
- The cramped and busy screen is confusing.
- The tasks require too many things to be remembered across a sequence of screens / actions.
- The memory demands of the tasks are too high.
- There is not enough to enjoy in the assessment in order to sustain engagement.
- Limitations in subject knowledge (e.g. mathematics) constrain engagement.
- The task instructions are not explicit enough about what has to be done.
- The task is ambiguous, or vague.
- The need to keep switching between objects or parts of objects that cannot be seen at the same time mean that pupils become disorientated, and lose track of where they are in the task
- Lack of knowledge of some specific actions to achieve certain things have disproportionate effects, such as precluding all progress with the task.
- Actions fail that were assumed to lead to a desired immediate objective.
- There is an inappropriately high demand for ICT knowledge and skill that is not relevant to the assessment.
- The pupil wants to undertake a particular action but does not know how.
- Pupils spend time making something nice rather than getting on with the test, e.g. trying several different fonts to choose the most appealing.
- The pupil has always used the same software, so her proficiency is context bound.
- Pupils know enough to succeed in the tasks without using ICT for all the steps.
- The unfamiliarity of the software leads to multiple attempts to achieve a purpose.

Although all the featured SoDs were recognised as having significant impact, it can be seen from this ordering that sources of difficulty that relate to pupils' state of mind were felt to have the highest impact, with memory and organisation issues also high, just ahead of sources of difficulty relating to language matters and pupil knowledge and skills. Relatively less impact was judged to arise from sources of difficulty relating to particular ICT knowledge, and pupils' approach to the tests. Generally, preparation-related sources of difficulty were judged to have lower impact, with the exception of the difficulties arising from a cramped and busy screen.

A further factor which is no doubt relevant is the ease with which remedial action may be taken, or the likelihood of success. However, this project is not in a position to comment on this, and it is left to others to decide which to pursue because of viability.

Recommendations for test construction (a framework for quality assurance)

Tasks

To a large degree, writing a high quality assessment task is not likely to benefit from 'recommendations' about tasks derived from the apparent limitations of past tasks. Removing ambiguity from the language used is not going to be a novel consideration, and is likely to be already present in the ambitions of task writers. Similarly, writing to ensure that memory demands are not too high, or to make the activities engaging and viable for the audience are not really recommendations so much as focused common sense. It seems more appropriate therefore to offer a set of criteria that can be applied both by reviewers and by the writers themselves.

Reflection on the implications of the project led to the following possible criteria when looking at tasks, put in the form of questions to ask, at increasing levels of specificity:

1. Will the overall task make sense to the pupil?
2. Will there be sufficient appeal and interest in the task?
3. Is there ambiguity in the instructions, or other language in the task?
4. Is there too much repetition, such that it brings a risk to focus and motivation?
5. Does the task have a clear goal, and retain a sense of purpose throughout?
6. Is there a relatively easy beginning that will get the pupils going on the task?
7. Is there sufficient structured support in the task?
8. Is the task do-able by most pupils in the allotted time (typically 16 minutes)?
9. Does the task assess what it is supposed to assess?
    - Are the language demands reasonable?
    - Does the task call on knowledge of the world that it is unreasonable to expect from a pupil of this age and level?
    - Does the task require memory skills beyond what should be expected?
    - Are the organisational demands of the task too great relative to the ICT being assessed?

- Does the pupil have to know other subjects to a high level, e.g. mathematics, in order to score on the task?

10. Will the actions that pupils are likely to do in response to the task be the ones that the task is designed to provoke, and which will trigger the appropriate assessment opportunities?
    - Might pupils interpret the task incorrectly (e.g. by supposing that the purpose is to improve in any way, rather than in the particular ways specified).
    - Could elements of the task be readily done without using ICT much?
    - Is a low-level way of doing the tasks feasible (relative to the level of the assessment)?

However, there are two suggestions that do take the form of recommendations:
1. To keep in mind when designing and reviewing tasks the five facets:
- Context – the nature of the activity;
- Content – the language, numbers, pictures, etc.;
- Task instructions;
- Visual representation onscreen;
- Scenario, or 'story' of how the pupil is expected to engage with the task.

2. To make the tasks and task instructions more explicit.
This recommendation does conflict a little with the 'appeal and interest' criterion, and the tension between them is explored further in the authenticity discussion in the following section. It can also conflict with the issue of too much to read in the test, as making text more explicit does generally make it longer.

Test development process
There are also some recommendations about the test development process that relate to meeting the criteria – recommendations about mechanisms to increase the likelihood of delivering a test with tasks that meet the criteria, given the available resources in terms of personnel. The first of these is about information, and the other is about the review points in the development process.

1. To try the finished tasks out with pupils, in order to gain information about the understandings they have of what is being asked of them, and the effect of that on their engagement.
The pupils' state of mind came out as the category of source of difficulty with the highest impact, but cannot be addressed directly. There is no recommendation for what to put into tasks to improve it, and while the list of criteria above will help, it is possible to imagine a task that appears to score well on all of them but still confuses and alienates the target audience. In order to address this vital aspect of task development, there really has to be detailed feedback about pupil responses built in to the writing process.

2. To give opportunities within test development for the task writer to ask for changes to the digitised question.
When an task which has by and large been conceived statically (as a sequence of steps) is realised dynamically (as an interactive computer activity) there can be surprises about how it operates in practice.

It can for example be more confusing than had been expected when all the elements in the task (intended to be dealt with one at a time) are open and active all at once.

A test item in an interactive context has the complication of chains of reactions. In a non-interactive test, there is a stimulus, and the assessment response, which is evaluated. In an interactive context (which includes practical assessments, as well as ICT based ones) there is a stimulus, and a response, but the pupil response is then responded to, and there is then a response to that, and so on. The basic activity is not 'show what you can do', but to make the computer do what it is supposed to.

What is assessed is bound up with the relationship of the person to the context. In interactive contexts there is a much higher risk of 'interference' as a challenge to accurate assessment. The relationship of the person to the static context is simple, and usually relatively general, a matter of attitude and expectation. In the interactive context, however, there is how the agents in the context respond to what is done, and what the person makes of that. ICT is a passive agent, but has been programmed, and responds as if thinking – even if more predictably. If the person is not familiar enough with it to anticipate the 'thinking' of the computer application, they do not know what to do to get the response they require. This is often the trigger for an extended activity of 'finding out how the thing works so as to be able to get it to do what it is supposed to' – which in a timed test is damaging to overall success, as it uses time that might have been more effectively spent on other activity.

After the task has been digitised, therefore, the task writer – who has the vision about how it is supposed to work – needs to spend time to ensure that it works as it was intended to, and to make amendments if necessary.

Software

Amending software is not a minor matter, and any recommendations here do need to be based on strong confidence that the change is needed and will make a significant difference. The sources of difficulty that the amendments to software would address are those related to memory and to becoming lost, respectively.

1. To add an onscreen scribble pad.

If it is accepted that digital solutions to memory issues do not represent good practice (an artificial solution adopted by a small minority largely to make a point) and yet paper and pencil support is excluded, the use of on-screen scribble pads should be considered. It can even be argued that ICT capability includes as an essential part of itself the use of paper or on-screen scribble pads, as it enables the organisation that sophisticated ICT requires – people do not develop additional cognitive capability (better memory) when becoming more capable at something, they develop techniques within an organisational structure. In terms of memory of the applications etc, it is about automatisation, but in relation to storage of external information which has to be dealt with at any one time, it is about organising the information outside of working memory.

2. To offer a means of keeping track of where one is in a task and the test.

The initial guidance (three tasks on one test and two on the other) which is currently offered is nothing like enough to enable a successful orientation or awareness in the pupil of what there is to do relative to what has been done. What other mechanism

might help? Might there, for example, be a tick list of elements on a floating memo window?

Pupil state of mind is certainly another aspect of the test that is affected by features of the software, and the listing above of the elements of the two main areas of software SoDs – the level of particular knowledge required, and the unhelpful operation of the software – might well be improved by amendments to the software.
However, it is not feasible to describe particular actions to take, and the positive effects of actions taken could not therefore be anticipated. Improving the functionality of highlighting and pasting, for example within spreadsheet cells and text boxes, and supporting knowledge about the applications in a more effective way than the current 'help' facility, are desirable but do not qualify as recommendations.

In a similar vein, it is expected to be helpful:
- to put plain language in the software menus, and increase the labels and descriptions attached to them;
- to make the software more tolerant of variation.

Recommendations for preparation by teachers
Preparation was a key aspect of sources of difficulty in the tests. Almost all observed pupils would have benefited from more or better preparation for the tests. However, it is not possible for any pupil to have perfect preparation for the test – to be both completely aware of the demands of the assessment so that they can shape their behaviour to maximise the outcome, and to be totally familiar with the software environment so that all the required actions are second nature to them.

In preparation the key question has to be how to make best use of the time that is given over to it. The amount of that time will vary across contexts, of course, because within the framework of a limited amount of time available for ICT instruction, strategic decisions will have to be taken, and these will vary.

These recommendations are based on our observations and analysis through judgements of what limitations in awareness and familiarity seemed to be the most damaging.

Test-awareness (including self-awareness as a test-taker)
Pupils need to know what they are facing – what kind of test it is, what will enable them to do well, and so on, or they will interpret the test in terms of precedent (other tests or other ICT lessons) and may well misapply their efforts. From the evidence of the project, the guidance to pupils might helpfully:

- Advise pupils to read instructions properly and not just act on the gist of them, and to do what they think it is likely to say, e.g. if it is about improving a leaflet, find out exactly in what way. They could not notice or bother with the exact instructions, and think they have done fine because they are sure that it is better than it was at the start, but this is not good in the test.
- Develop a sense of opportunistic assessment performance – focusing on doing what is being done in a way that will produce a good outcome, rather than 'getting through' the tasks as given. Be clear about what is wanted. In this respect clear

guidance from QCA and the test development agency about what the test is assessing and how it works will aid the teachers and pupils.

- Reflect on how to respond to difficulty in the light of what is being assessed.
- Advise quick typists to cut and paste rather than re-type, even though typing is faster.
- Advise all pupils to use ICT methods for searching and counting etc., even when alternatives may be faster.
- Address the issue of attitudes: for example there is a potential source of difficulty for some pupils related to contempt for the old-fashioned nature of some of the visual material being improved in the ways suggested. The presentations for example look to some pupils like they have been done by adults who are inexperienced with presentation software, as these pupils have been doing presentations for years, and some of them do know how to make much 'better' presentations. These pupils need to be pointed to attitudes that will enable them to succeed.
- Consider with the pupils the matter of what to do when they try an action that does not work. This can be for different reasons: (a) it is not possible in that interface / application; (b) it is blocked in that task; (c) the conditions are not right (e.g. something else needs to have been done first); (d) the action was performed incorrectly; (e) there is a 'bug' in the software. When pupils try an action that does not work, they respond in different ways, depending on what they judge to be the reason for the difficulty. Some try something else (presuming a or b), some make an adjustment and try again (presuming c), some just keep trying (presuming d), and some despair and give up (presuming e).
- Look at specifics in how the test works. For example, when an email attachment is opened and worked on, it has to be saved as a different name. But then the new file cannot be accessed from the email. Some pupils were observed to open the email attachment and think that their saved work was lost, so did it again (in some cases more than once). Pupils should be advised to access their work through the file manager system.

Familiarisation with the interface

The familiarisation SoD with the highest judged impact was "the cramped and busy screen is confusing". So beyond a general familiarisation with the interface, and general purpose navigation experience, it seems important to focus particularly on:

- Practising organisation of the visual environment on the software, split screens, maximising screens, etc. to enable better decisions about how to access information while avoiding cramped and busy screens.

Familiarisation with the applications

There is a great deal to learn about the particular applications used in the test, ranging from their formal names ("database"; "spreadsheet"; "presentation" etc.) which some pupils observed in the project did not recognise, to what conventions are used for various functions, which may well be different from what the pupils are used to.

In giving experience with the applications used in the test software, it is important not to do so just by running practice versions of the tests, as the pupils' attention will probably not be on learning how the applications work but on the content of the tasks, and the opportunity will not have led to improved familiarity.

It may be helpful to:

- look at the applications with purposes in mind – how it might be used to … ;
- compare the application in detail with the familiar application usually used – what is done in the same way; what is done differently and exactly how;
- use applications together, not just one application at once. Even when they use an application to achieve a real purpose (good practice in NC terms), it is not good preparation for this test if this is done one application at a time;
- consider the language used in the tasks of the test, such as “Model” to mean spreadsheet.

It may be important also to look across the other departments in the school to see how the applications are used (databases in science and geography, word processor in English, spreadsheet in mathematics, web-based work, leaflets and presentations in a number of subjects), and consider the impact.

Preparation in terms both of familiarisation and test awareness is clearly very important, and the two hours suggested for the 2006 test seems unlikely to be enough, given the range and scope of the descriptions above – although it is difficult to be precise about this from our data, since many of the schools we visited had not even done that much preparation, merely for example doing the ‘practice’ tests to prepare the pupils for the actual tests. Yet in the context of very limited ICT curriculum time an open ended ‘do as much as you can’ opens up the possibility of some schools spending more time preparing for the test than teaching the curriculum and developing the capability that the test is supposed to be assessing. How can schools be persuaded that the test is important and yet that they should only spend a certain amount of time preparing for it? Can there be preparation activity that also contributes to developing the ICT capability that the test is assessing – preparation that brings about worthwhile learning, not just test proficiency?

7. Discussion

In this section, there will be a brief consideration of a small number of broader perspectives that arose during the project.

Qualities of the pupil

What gives a task its difficulty should in theory be just the attribute or attributes that it assesses, but it is inevitable that other knowledge and skills will be drawn on in the assessment. These are generally a matter of shared understanding within the community of practice, established by precedence and by debate. In effect these are part of what is being assessed, but are often ignored as 'background' – unless it is shown that they introduce bias into the assessment. In ICT assessment, however, where this kind of assessment is a new enterprise, the 'background' qualities and skills arising from any particular form of assessment need to be discussed as to whether they are indeed acceptable as part of what is being assessed.

The pupils who did well on the test seemed to be the well organised, careful readers with good comprehension skills, who were reasonably confident and competent with ICT, had a good eye for literal detail, were methodical in task completion, and were willing to follow instructions. If all went well, these pupils seemed to have enough time on the test to do the tasks more or less as intended, and therefore presumably accrued the necessary evidence to score appropriately. However, the pupils who were disorganised, or were not good readers, or not methodical, or not good with detail, or who dislike following specific instructions, as well as those who are not competent or confident with ICT, these pupils did not seem to do well at all, in that they did not complete tasks, did them inappropriately, and often seemed to disengage. It seems unlikely that they would have been awarded a level, yet this was not in all cases because of poor ICT. It seems that the tests require that minimum standards of personal organisation, reading, and ability to engage with detailed requirements are met, in effect defining good ICT in terms of needing to have all of those qualities, as well as knowledge and skill in ICT as more conventionally defined. They amount to implicit or hidden criteria in the assessment, and this issue needs to be made more explicit. Statements about ICT capability currently do not refer directly to such qualities, but if is accepted that they do form part of what should be assessed, this needs to be made clear. If they are not accepted as part of ICT capability, then this test will be seen to have a bias in it against pupils with weaknesses in personal organisation, reading or compliance – which may well include many pupils who are characterised as having special needs.

Pupils' experience with ICT outside school

Experience with computers plays a large part in what pupils learn, and in ICT – above almost all other areas of learning addressed in school – experience outside school is extensive and influential. There are now few pupils who do not have a computer at home (but still some, and this should be borne in mind) but there is a large variation in the home-use of computers, with some homes focusing strongly on finding out, with others focusing on play and entertainment. It is a challenge to any test to be fair against that background.

For example, it may be argued that pupils' competence with games, websites, messaging and social software is not relevant to the demands of the tasks in this assessment which are assessing capability, because in those contexts they do not have to use what they retrieve in a cognitive way. But that means that some pupils will reap the benefits in the test of the use that they have made of computers outside school, but others will not. There will be pupils who will be widely considered to be highly competent in many aspects of computer use, but who not do well on this test.
In that respect 'capability' may be being defined in a particular way that may not chime well with common usage.

Authenticity
The first source of difficulty on the list developed by this project refers to the difficulties engendered by lack of explicitness in the tasks – and a recommendation is made to make them more explicit. But this avoids the question of why the tasks were not explicit about exactly what is required. The answer to this seems to lie in the attempt to make the tests 'authentic' – believable as a challenge.

Referring to the five aspects of tasks outlined above, the authenticity of a task might be said to refer to the same five dimensions:

1. the context – Is it the kind of thing that is asked for in office work that uses emails to set tasks?;
2. the content – Are the quantities and other 'real-world' references as might be expected?;
3. the instructions – Is that how jobs are given?;
4. the visual appearance – Does the software and what is shown on it look right?;
5. the scenario, the story of pupil engagement –Would that be how one would do the job?

A question to ask about authenticity is not whether it simulates reality, but whether the user believes that it does – and what effect that has on them.

Authentic assessment is generally approved of because of the motivational aspect – pupils are more likely to engage if the task is believable as a task that they might one day be required (even paid!) to do. It is contrasted with the artificial nature of so much school assessment, involving tasks that everyone knows will never be set again after the end of schooling.

However, authenticity is not so easily attained, and questions can be asked about just how authentic from the point of view of the user the 2005 and 2006 KS3 ICT tests have been.

For example, how many pupils will recognise the particular world of work that is being simulated in the test? Such recognition was not evident in any of the pupil interviews (although, to be fair, no specific question was asked about it). Also, a noticeable proportion of the pupils we have observed have said that they do not use email, some because their parents do not allow them to, and those who do use email are much more likely to use it for chat than to receive instructions, and many do not receive and send files as attachments. The authenticity of the context from the pupils' point of view would seem to be less secure than is being presumed.

In the absence of that overall recognition, the effort expended by task writers in getting the details right in the tasks and instructions would be wasted, because the whole setting does not seem authentic to the pupils.

The instructions usually include a general injunction to improve something, but as only the 'right' improvements contribute to scoring, there are also specific demands, often partly hidden to preserve a kind of authenticity. And indeed, this is exactly what does happen in office work, with employees continually having to interpret the manager's instructions in the light of what is known about them, so as to know what exactly to do. Where instructions are given by email, the vagueness is exacerbated by the informal nature of emails. It is an unusual (because unusually good) manager who distinguishes occasions in which employees are to be given discretion from those in which a particular solution is required, and for the latter spells out exactly what is to be done. Yet does this aid assessment? It is doubtful, as the ability to anticipate the thinking of an imaginary manager is not one of the qualities likely to be accepted as part of ICT capability.

A different aspect of the authenticity argument is about its supposed effect – to motivate pupils to engage in the tasks, and ICT lessons do emphasise motivation through interest with authenticity part of the means to achieve it. However, the bigger reality in the test situation is that it is a test. Pupils have to behave in a 'test' way, and have to be made aware of that, so motivation towards task engagement through interest arising from authenticity seems unlikely.

The key finding in this respect is that the test needs pupils to have a test awareness to behave in a way that will trigger 'opportunities' and generate an accurate test score. If the test could locate the creditable aspects of behaviour in the context of pupils being unaware of being tested (as an observing teacher might in an authentic practical assessment) then the benefits of authenticity – of pupils relaxing into a smooth performance to really show what they can do – would probably win the day. But in this test, if the pupils interpret the tasks in a way that is slightly differently than intended, they do not score as well as they should. The computer can only recognise what it has been programmed to recognise, so the pupils have to do the 'right' things. In this context, the authenticity does not achieve its major purpose, and it is no sacrifice to compromise on authenticity to get a better test. Motivation for engagement in the tasks in this test, as in all other tests, has to come from wanting to do well in the test, and the test should support pupils in this. Further, the authenticity is not only not serving its purpose, it is getting in the way of the success of the assessment. The authentic deliberate vagueness gets in the way of their understanding of what is required, so it is better perhaps just to tell them. More explicitness may mean less authentic instructions, but perhaps would help guide pupils to scoring responses.

It may be that there is an unavoidable tension between the authentic context chosen for the test, of tasks that in life reward by outcome, and the purpose of the test – the assessment of process. In the tasks, the authenticity suggests that what matters is producing outcomes that the imaginary client would like. In the test, it doesn't really matter what you achieve, so long as you do it in the right way. Raising the pupils' awareness of either is detrimental to the other, so there is an assessment reason for

undermining the authenticity aspect of the test. This is an example of the interesting relationship there is between authenticity and test reliability.

Trying to make the tests interesting through being authentic in some respects can be worthwhile – but recognising the limits of pupils both in terms of what they will recognise and of how far they will benefit is an important consideration, as it indicates where it is appropriate to compromise on authenticity to achieve the assessment purposes.

A conundrum of the assessment

The approach used in this test to the assessment of ICT capability is of using products with which the pupils are not familiar (although they will have used similar applications software) in order to demonstrate that they have developed capability beyond brand-specific skill. However, lack of familiarity has been shown to be a significant source of difficulty in the tests, and so either considerable time has to be taken within the assessment to explore how exactly the common software function operates in this version of the application, which prevents success, and undermines the assessment, or time has to be taken before the test to become familiar with the applications in the bespoke environment, which develops a new brand-specific skill that contributes to success, and undermines the assessment.

8. Conclusion

In many respects – smoothness, pupil engagement etc. – the test that was used for KS3 ICT assessment in 2006 seemed to work quite well, but this project tried to evaluate how well it assessed what it was supposed to assess, and raised a number of questions.

An abiding background issue throughout the project was that, although a very general account can be offered and accepted (the use of ICT tools to achieve relevant purposes), no-one really knows what it is to be good at ICT – it is a new area of capability, and for good or ill the test will help in the process of establishing perspectives on what it means to have ICT capability.

The potential of the tests

As a new enterprise, the assessment by a test of ICT capability, and also one which adopts a novel approach to a persistent problem – the reliable assessment of process – it would have been very surprising if the tests were wholly effective this early in their development. It is remarkable in fact that so much has been achieved in creating a viable pilot.

Equally, because it is all so new and so novel, and despite the remarkable start, it is not yet certain that the ambitions of the test can be realised. The rapidly changing nature of ICT, and the complexity of the characteristics of capability in it, may mean that the attempt to do ICT capability assessment in a timed test is a fool's errand. Yet one will not know without trying, and it is in a way reassuring that problems have been found through this project that can be addressed in a practical way. If there is something to improve, then there is the hope that it can be improved enough to serve at least some of its purposes. What cannot be said is that the improvements suggested by this project will make the test successful, and certainly not perfect (because no assessment is perfect). For example, some may feel that one of the essential elements of using skills in pursuit of purposes is the creative element. Common purposes such as 'capturing interest' are not achieved by a mechanical application of skills, but require a more open deployment of skills in a 'creative' way. This is not and could never be allowed for in a timed test with mechanical marking. That and some other aspects of ICT assessment are unlikely ever to be met successfully through a test of this kind.

Nevertheless, this test may prove to be sufficient for a number of aspects of ICT assessment, although establishing this may take time. The test is likely to be improved by following through on the recommendations listed earlier in the report, but when they have been addressed, others may become prominent (possibly some of the hypothesised SoDs that were not seen particularly frequently in this project).
The potential of the tests is therefore encapsulated in the fact that they are improvable – the project cannot speak about how far and how fast it can be improved.

The risks arising from the assessment as written.

These have been the main focus of the project, evaluated through observation of sources of difficulty in the tests. As a result of how the test is written:

- The pupil may become confused, frustrated, dispirited or de-motivated;
- The pupil may become lost in the task or test;

- The memory demands of a task could be too high;
- Limitations in subject knowledge might constrain engagement;
- The language of the task could be unclear;
- Too great a level of familiarity with the software may be required.

Recommendations for mitigating actions against these risks are given in section 6 of the report.

The risks arising from the nature and form of the assessment
All assessment carries risks, but in addition to the risks outlined above which arise from specifics in the tests, some risks reflect the nature and form of the approach to assessment adopted. In contrast to the 'constructed' risks, against which mitigating actions can be taken, the 'nature and form' risks may be seen as the fixed background risks, which constrain the extent to which the test can be improved by attention to tasks, software and preparation.

With this assessment, the following seem to be the risks that attach to it arising from its overall design.

1. The authenticity that is being striven for in the test may undermine the assessment.
Authenticity in the context could lead to inexplicit or ambiguous language.
Authenticity in the 'scenarios' could lead to very complex tasks.
Authenticity in the interface could lead to insufficient support to the pupil who is unsure how to achieve short term objectives.
Any of them may mean that achievement in coping with the authenticity is out of proportion to the requirements of the national curriculum that are being assessed.

2. The data capture approach can miss the evidence of capability.
The computer can only capture what it has been programmed to capture. This contrasts with ongoing assessment by teachers, who can respond with judgement to what pupils are seen to do, even if entirely novel and unexpected. The limitations of the marking algorithm need to be taken into account both in how the evidence is treated, and how the final judgement is used.

3. The complex tasks require sophisticated responses.
The assessment has a requirement for dispositions as part of the complex trait that is being assessed. This includes some degree of organisation of the visual environment and control of memory, language skills and so on, which have not previously been explicitly accepted as part of ICT capability.

4. The test situation may prevent sophisticated responses.
The complex trait that is ICT capability may be present in the person, and evident in self-organised activity, but fail to appear in a test situation, just because it is a test situation.

5. The test as a multi-level criterion-referenced test brings tension between the different purposes. Qualities that are wanted as part of the test to permit assessment at higher levels could prevent good assessment at lower levels – but the degree of support necessary to allow good assessment at lower levels could prevent the necessary independent action to give a meaningful award at higher levels.

6. The use of a bespoke interface could undermine the assessment.
The applications used in the interface are inevitably a limited set compared to the range of ICT tools used to pursue ICT related purposes elsewhere – they are those which are viable in the context of a timed test, and which work together within a package. However, if there is 'teaching to the test' and teaching ignores the wider range of ICT tools and purposes, then the assessment will be of a very constrained capability, without relevance to real usage.

There are also two 'novelty' risks:

7. The newness of the test may lead to standards being set too high.
In other areas of the curriculum there has been an adjustment over time in the demands of test situations relative to the levels that test performances are said to be indicative of. In other words, you are not required to do as much in a test to be awarded a level as you would be expected to show outside the test. In the absence of the precedent for making such decisions in ICT, the test may be set too hard.

8. The newness of the approach requires guidance that may undermine the assessment
Either the lack of visible 'mark-schemes' will lead to inappropriate preparation / action, or the presence of visible 'mark-schemes' will allow teaching to the test.

Issues for further research / investigation
Within the context of continual improvement of assessments, the matters relating to this test that may best reward further attention in the next cycle are:

- The effects of suggested changes: Where suggested improvements are neither obvious nor agreed to by all, to make changes that contrast in versions of tasks, and study the effects with groups of pupils.

- Reducing the size of the test: How far could tasks in the test be reduced in size, with fewer things to do, before the test could not generate enough opportunities to award levels?

- A study of how different schools / teachers prepare pupils for the tests, and which approaches seem to be the most effective (supporting or otherwise the speculations about this given above).

- A (later) close study of time taken to complete tasks and sub-tasks, and what takes up time, and how: Many pupils in the trial did all that they could in the time available, and wasting time on some aspects was not particularly damaging to their overall performance. However, when pupils are better prepared and are doing more, then distractions that waste time may have more impact, and new sources of difficulty may emerge.