Uncovering and Developing Student Statistical Competences via New Interfaces

Jim Ridgway, University of Durham, United Kingdom

Sean McCusker, University of Durham, United Kingdom

James Nicholson, Belfast Royal Academy, United Kingdom

Abstract

The paper reviews the nature of statistics in the UK National Curriculum. While there is great opportunity for statistics to inform thinking in a range of disciplines, there is little coherence in the planning of activities for children. Formal assessment of statistical ability focuses on procedural knowledge, applied to univariate and bivariate problems. These provisions are inadequate for informed citizenship. We report work on World Class Tests (WCT), where students aged 9 and 13 years are presented with both paper-based and IT-based problem solving tasks in mathematics and science. Many of these tasks require students to work with three or more variables, often non-linearly related to each other. Students perform rather well on these tasks. Tasks are shown, together with performance data disaggregated by school and sex. Features of the displays that make them well suited to the exploration of complex patterns are discussed, and our design principles are described. Our current activities set out to embed test items in curriculum materials suitable for use in mathematics and science classrooms. A good grounding in handling complex data is essential for informed citizenship. We are in the process of creating methods to help students develop appropriate skills.

Introduction

> Statistics ... 'the most important science in the whole world:' for upon it depends the practical application of every other science and of every art; the one science essential to all political and social administration, all education…
>
> Florence Nightingale

> …and the time may not be very remote when it will be understood that for complete initiation as an efficient citizen of one of the new great complex world-wide states that are now developing, it is as necessary to be able to compute, to think in averages and maxima and minima, as it is now to be able to read and write.
>
> H.G. Wells

The English National Curriculum sets out programmes of study for every curriculum subject for students aged 5 to 16 years. Statistics occurs as a topic for study under Data Handling, one of four strands in Mathematics (the other strands are: Using and Applying Mathematics; Number and Algebra; Shape, Space and Measure). There are rich opportunities across the curriculum to develop an understanding of complex issues with the help of statistical techniques (notably in Science; Geography; History; Personal, Social and Health Education; and Citizenship). Unfortunately, the National Curriculum was written subject by subject with no attempt to produce coherence across subjects. As a result, the same topic in statistics is to be presented to students of different ages in different subjects, and some important topics do not appear at all (for a detailed analysis see Holmes, 2000). Another consequence of the lack of coherence and continuity in the statistics curriculum is that teachers with no background in statistics education are making decisions about which concepts to introduce to students, with little appreciation of the inherent difficulties of those concepts.

Advanced-level subjects are available for students aged 16 to 18 years. Here students choose a range of more specialised courses. Mathematics is one of the three most popular subjects (along with English and General Studies). Data for the year 2003 from the UK's largest Examinations Authority show that more than 85 percent of students took at least one statistics module, and that more than 50 percent of students took two statistics modules. We set out to analyse the content of these modules. Two raters independently analysed the examination papers used by each of the six Examination Authorities at AS level (the first level of advanced study) to assess students' statistical knowledge. Raters were asked to categorise the questions according to whether they assessed computational or procedural knowledge, or could be broadly classified as interpretation items. The interpretation category of items included those that related to: identifying the assumptions underlying certain models (e.g. use of the Poisson distribution); making choices about which probability distribution to use in particular contexts; interpreting the conclusions of hypothesis tests in context; and all instances of reasoning from data.

Inter-rater agreement was high. On no paper did their judgements on the proportion of marks in each category differ by more than a single mark. When differences existed, they were resolved successfully by discussion. The proportion of marks awarded for the broad interpretative categorisation on the examination papers were: 3%; 12%; 22%; 25%; 25% and 28%. Raters were also asked to identify questions where students were required to work with three or more variables. Not a single example was found on any examination paper.

The Problem

An unsatisfactory pattern emerges when we analyse the national curriculum together with the detailed study of statistics examinations. Statistical education focuses on a rather narrow range of techniques applicable only to univariate and bivariate analyses. In addition, student exposure to practical applications of statistical ideas in the curriculum is poorly specified. It is likely that students in some programmes will enjoy a curriculum rich in statistical ideas, while the majority will not. If we consider the role of education to be preparing students for informed citizenship, one is hard pressed to show where in the National Curriculum students acquire the skills necessary to understand the data presented in government information, such as that intended to explain new evidence-informed policies on crime www.homeoffice.gov.uk/justice/sentencing/correctional/reducingcrime-changinglives.html). We suspect that neither Florence Nightingale nor H.G. Wells would be happy with our current situation.

Addressing the Problem

> It is the grand object of all theory to make these irreducible elements as simple and as few in number as possible, without having to renounce the adequate representation of any empirical content whatever.
>
> Everything should be made as simple as possible, but no simpler.
>
> Albert Einstein

For the democratic process to work effectively in a complex world, citizens need to be able to understand arguments based on complex evidence that uses both quantitative and qualitative data. This places an important duty upon educators world-wide. An extensive body of research literature highlights the statistical misconceptions held by adults (e.g. Tversky and Kahneman, 1983) and by young students (e.g., Batanero, Godino, Vallecillos, Green, & Holmes, 1994) that need to be overcome. Much of what is taught in schools focuses on univariate and bivariate analyses and on linear relationships. Such knowledge provides a useful grounding for statistical knowledge but has a rather limited application in the face of the data encountered outside school settings.

Technologies are emerging which make it possible to present complex data in new ways and to give users considerable control over how they interact with data sets. Some results have emerged from large-scale trials where informatin technology (IT) has been used to present exciting challenges to students, based on complex data sets. The results indicate that IT reduces the time spent doing the calculations and manipulations that characterise much of the current school statistics curriculum. However, apparent improved procedural competence may mask a lack of underlying conceptual understanding. Experience in classrooms (e.g., Nicholson, Mulhern, & Hunt, 2002) suggests that increased automation has reduced 'contact' with the actual data. As a consequence, a grasp of the meaning of the data and the sensibleness (or otherwise) of possible relationships is not well formed. Many software packages do not prevent the user from drawing inappropriate graphs or applying unjustified statistical procedures, nor do they provide either the depth of inferential insight or the background knowledge (i.e., context specific specialist knowledge) needed to make reasonable interpretations of the analysis.

Greer (2000) observes that access to data is increasing dramatically, but human skills in data handling lag far behind. Hawkins (1997) observed that having the vision to see what technology can, or might, do is not synonymous with knowing how to take advantage of this in a teaching context. Making effective use of technology in the context of teaching statistics has a number of inter-related dimensions. These relate to issues regarding the use and usability of software packages; the growth of opportunities to dynamically and interactively present ideas to students; and the need to develop robust models of new sorts of conceptual learning and understanding, which will underpin teaching in IT-rich environments.

One of the challenges of introducing statistics coursework into GCSE mathematics is that we have little subject-specific pedagogical knowledge to guide our teaching – such as a sensible hierarchical structure of statistical concepts that allows us to effectively develop interpretative skills in pupils. An additional challenge is the fact that pupils are expected to deal with complex data sets where multiple variables are to be considered (Nicholson, 2003), although the main emphasis of instruction has been on the mechanical skills of correctly constructing specified graphs or 'reading the graph' to extract detailed information. There has been little tradition of data interpretation within the mathematics curriculum. One way forward, in this regard, involves exposing students to large numbers of data sets and graphs that have something to say that is both accessible and of substance. The advent of new technology makes this an increasingly practical approach, provided that the necessary teaching resources are available when the hardware becomes available in classrooms. However, we must also recognise that IT can actually facilitate the development of new cognitive processes, so we need to develop a pedagogy that capitalises on a wide range of new affordances of IT (Ridgway and McCusker, 2003).

World Class Tests

Clues towards understanding what students understand and what needs to be done to develop their interpretative skills can be found in the results of assessments of their problem solving skills. Such assessments allow us to identify gaps between mechanical skills and conceptual understanding. Assessments of problem solving skills, given in the context of technology, provide even more clues as to how to best design an IT-based statistics education curriculum that helps students meet the GCSE standards for statistical competence.

One assessment that provides rich clues was administered in the context of a large-scale project funded by the UK Department for Education and Skills. This project set out to assess the problem solving skills of high attaining students in mathematics, science and technology. The Mathematics Assessment Resource Service (MARS) Group, based at Durham and Nottingham, has been responsible for developing the problem solving tests, and much of the testing has been computer-based. Our focus has been on students aged 9 and 13 years. So far, over 10,000 students in 20 countries have taken our tests. Each test has a paper component and a computer-based component. The computer-based component consists of five tasks, and students have 60 minutes (age 9 years) or 75 minutes (age 13 years) to work on these tasks.

Because of the focus on mathematics and science, many of the tasks relate to the category of 'reasoning from data'.

Our central concern in designing these tasks is assessing problem solving ability, not curriculum knowledge. A challenge for task designers is to create tasks that are intellectually rich, using interfaces that pose no technical problems during the testing sessions. A number of new task 'types' have been developed. In this paper we focus on only four of the interfaces. Our hope is to illustrate the scope for developing new interfaces that are easy to use and encourage an enthusiastic engagement with data exploration and data analysis. Three assessment tasks are shown below, and one curriculum task is shown later. Scoring rubrics for the assessment tasks are included in the Appendix.

Assessment Task 1: Oxygen

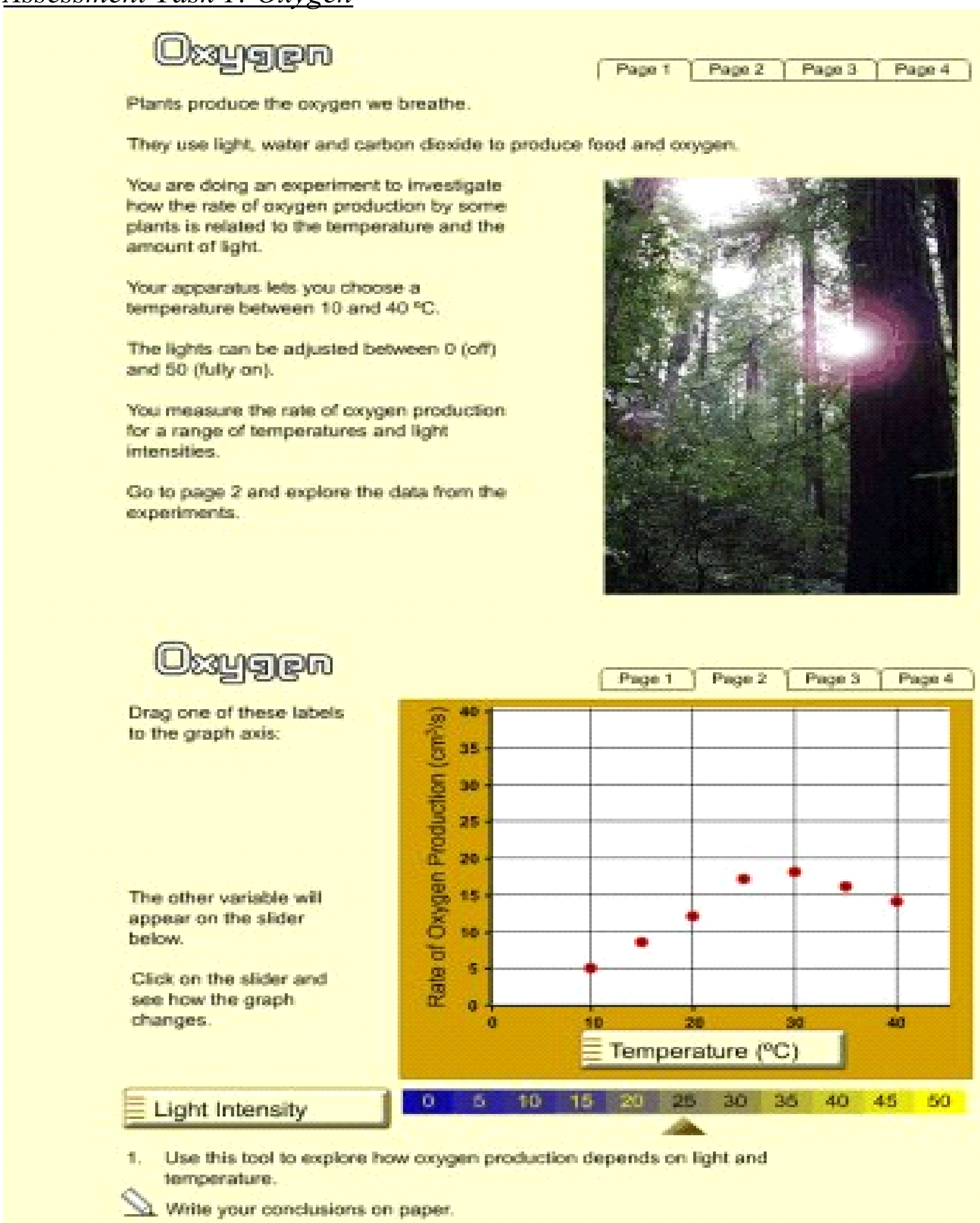


Figure 1. Screenshots from the Oxygen Task (age 13 years)

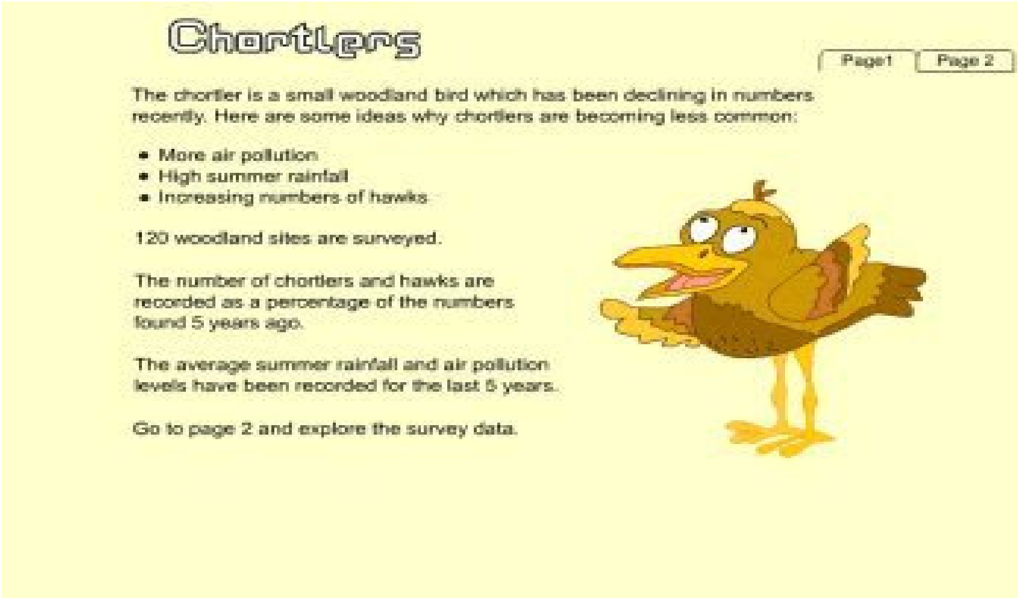
Oxygen is a task that provides data on oxygen production by plants as a function of light level (which is roughly linearly related to oxygen production within the range of variables used) and temperature (which shows a curvilinear relationship). At age 9 years, students are shown a graph of plant growth as a function of temperature, and have 'light level' as a slider. At age 13 years, they study oxygen production, light intensity and use a denser data set. Later the task allows students to choose the axes to be plotted (light level can be plotted, and temperature can be used as a slider). They are expected to use the slider to control the setting of the other variable. Students must actively engage with the display and make decisions about exactly which aspects of the data they wish to see. The task involves three variables within a real-world setting. The data used within the task are realistic, though somewhat 'cleaner' than real data, and the task is placed in a context such that students perceive that the exercise and results may be of use to someone (notably horticulturalists). A working version of the task can be viewed at www.worldclassarena.org/.

Students are required to explore relationships, some linear and some not, within a data set of three variables. Having understood the nature of the relationships, they are expected to find variable settings that produce the highest oxygen yield. A number of hypotheses are presented to the students that partially explain the features of the data and leave other obvious features unexplained. Students are asked to use their understanding of the data to evaluate and reconcile them.

An interesting informal observation while watching large numbers of students working with the task (and during development trials) was the extent to which students appeared to code the data enactively. If asked to describe the data, and often spontaneously when asked for explanations (or when observed solving the problem on their own), students would move their hands and arms in shapes that mapped out the 3D data surface. This map could be deduced from the sequential analysis of all the 2D data plots. Students used the kinetic code seemingly easily and unconsciously – they saw no need to explain what they were doing – the listener was expected to immediately understand the data surface being sketched in the air.

Continued on Next Page

Assessment Task 2: Chortlers



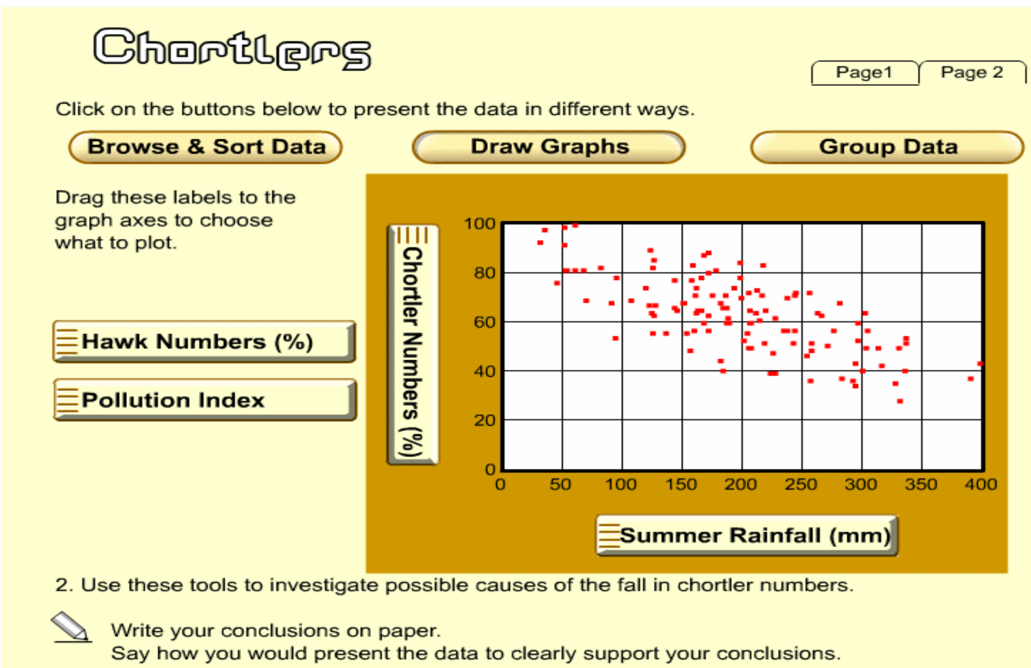


Figure 2. Screenshots from Chortlers Task

Chortlers is a task that provides data on the association between a bird population and three factors that might be associated with the populations' declining numbers. This task involves students in an active process in which a number of choices about data presentation need to be made. Students can choose to sort, group, or graph the data. Each of these choices have further options of choosing which variable to use as a primary discriminator and which variables to plot on a graph. Data can be explored in any way and in any order. Students are allowed to investigate as many data representations as they wish.

Students are required to explore a data set comprised of four variables. The task requires students to identify any factors associated with the decline in the bird population and to provide appropriate evidence, from the data, as justification for their conjectures. The data have been invented; however, they may be described as realistic in that they tell a credible story about factors influencing the decrease in a fictional bird population. Within this context, the task of exploring possible plausible factors can be seen as useful and productive.

<u>Assessment Task 3: Cowpats</u>

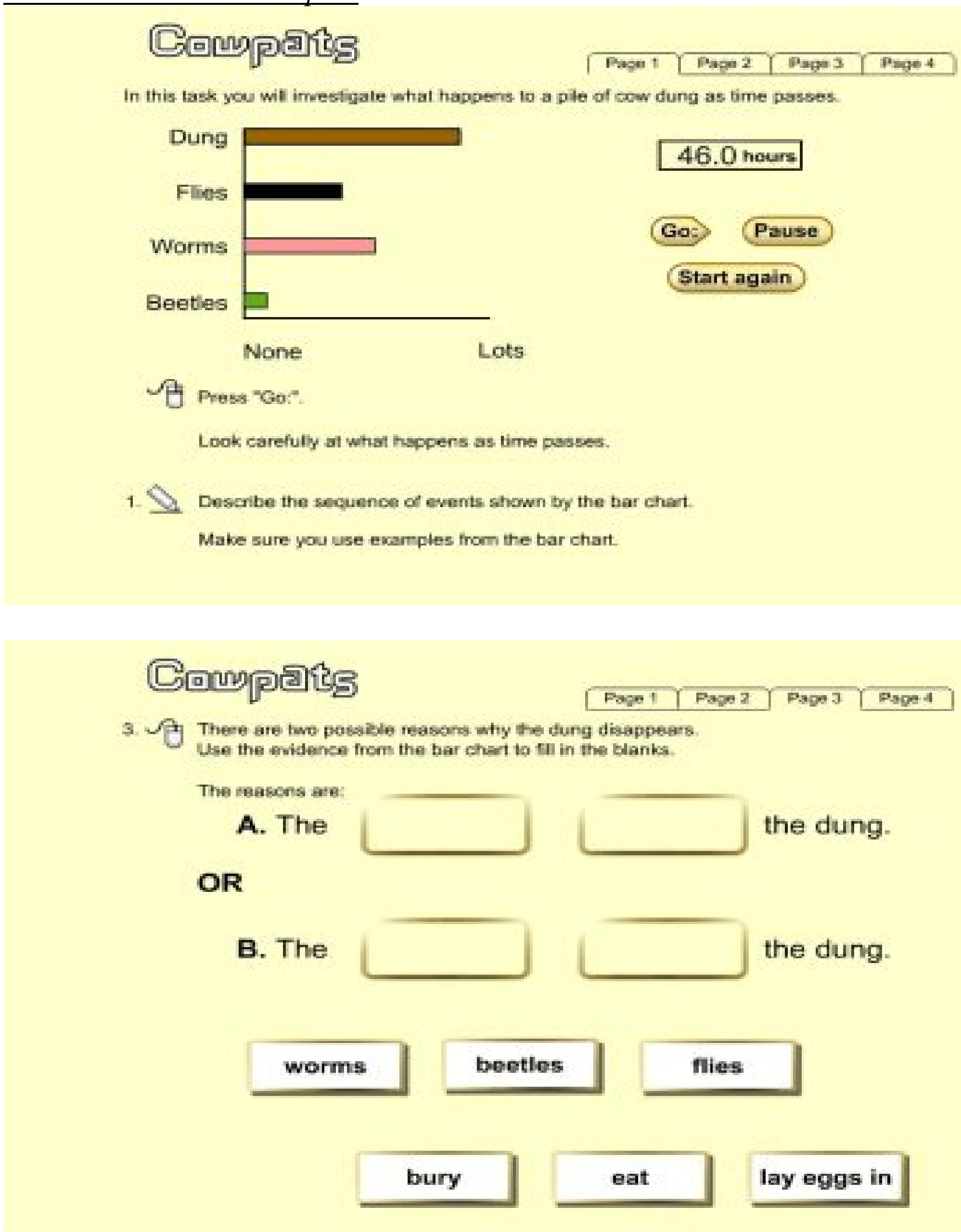


Figure 3. Screenshots from Cowpats Task

Cowpats is a task that presents data in a dynamic display of four variables. It asks students to reason about the causes of the removal of cowpats. The task shows the growth and decline in the populations of different creatures (worms, flies, and beetles) that might be the cause of the removal of the cowpat. They are also presented with data on the size of the cowpat itself.

The task is based on real-world data and the relevance of the process and results of the task are apparent. Students are asked to describe the data in general terms and then perform a sequencing task where they reorder 'snapshots' of the dynamic display. Students are asked to deduce the cause of the disappearance of cowpats based on the dynamic data and to suggest two hypotheses as to the cause. To achieve this, students must be able to identify plausible causal links by looking at the timing of different events in the data. They are also required to apply what they have learned from the data set to a real-world setting (the long life of cowpats on concrete) and from this to provide a better hypothesis than the one offered. Overall students need to view and understand the data, and from this conceive of a model that accounts for the phenomenon. They are then expected to apply that model to a real world setting.

Each of these three tasks was designed, then critiqued by an expert group, and refined in the light of small-scale trials with students. Among the criteria applied at these small-scale trials were that the task should be interesting, challenging, and easily engaged. Richardson, Baird, Ridgway, Ripley, Shorrocks-Taylor and Swan, (2002) presented a set of computer-based tasks to individual students and interviewed them on a number of task features. Students reported they were immediately interested in the tasks; they experienced no problems with the computer interfaces; and were happy to discuss the advantages of IT-presentation over paper-presentation. During the trialling process, students were asked about any interface problems they experienced. These were very rare, and students reported that they enjoyed doing the tests.

Results of Student Performance on the Specified Tasks

Tests were administered to hundreds of volunteer students who were entered for *World Class Tests* (WCT) in a number of test sittings. The sample is ill-defined, but entry was welcomed from students judged to be in the top 20 percent of the ability range. The data from these large-scale trials show that young students can work with multivariate data, even where the relationships are non-linear and they can make considerable progress on all of these tasks. These results show all the tasks pose challenges to students that they do not face in the English school curriculum. Nevertheless, many students performed well on these tasks. For example, on *Oxygen*, which is a 10-mark task, the mean student score was 4.12 (standard deviation = 2.85). On *Chortlers*—an 8-mark task—the mean score was 3.09 (standard deviation = 2.70). The best performance was on *Cowpats*—a 12-mark task, where the mean was 6.89 and the standard deviation was 2.03.

None of the tasks have been subjected to detailed analysis in order to determine the key design features that lead to student success. We recognise that more systematic work and further invention of task types is necessary if we are to understand and exploit the capabilities of IT to help students work with and interpret complex data. However, informal evidence from observing students during early trialling and during pre-tests suggests several distinct benefits from presenting tasks via these new interfaces:

- The user has control over the appearance of the data in ways that are impossible to achieve via paper and pencil – for example, in fast-changing dynamic displays.
- Data can be re-presented rapidly in a number of static forms under user control in ways that can also be achieved using paper and pencil – but only by considerable effort. This allows the students to use a more creative and playful approach to the investigation of the data relationships, because the time penalty for a null result investigation is minimal. As such, half-formed ideas or hypotheses can be tested and either rejected or investigated further.
- Three-dimensional representations of data ‘surfaces’ can be created.
- Hand movements can be used to encode complex data in ways that make it easy to understand and remember using kinaesthetic dynamic representations.

These data and observations promise much and raise some fascinating questions for further research. Some of these questions relate to whether, and how, task performance differs for various subgroups of students. We set out to look for differences in attainment between schools and between girls and boys. As our analysis progressed, we found that these judgements are not as easy to make as might first appear, especially since schools choose whether to enter students and which students to enter. Given this, it follows that a large school entering a small number of its highest attaining students might appear to out-perform a smaller school entering the majority of its students. A different outcome may have been noted had selection procedures been standardised across sites. Similarly, if schools put forward a far higher proportion of boys than girls and there are no sex differences on the test, the pattern of selection could give rise to an apparent sex difference in performance in favour of girls. There could also be an interaction between these two factors. For example, inclusion of students from a high attaining single sex school could distort the evidence on sex differences.

Results of Student Performance (by School)

The boxplot below shows data from five schools where at least 11 students took the test. As Figure 4 shows, the performance of students in the highest performing school is significantly higher than the performance of students in the two lowest performing schools.

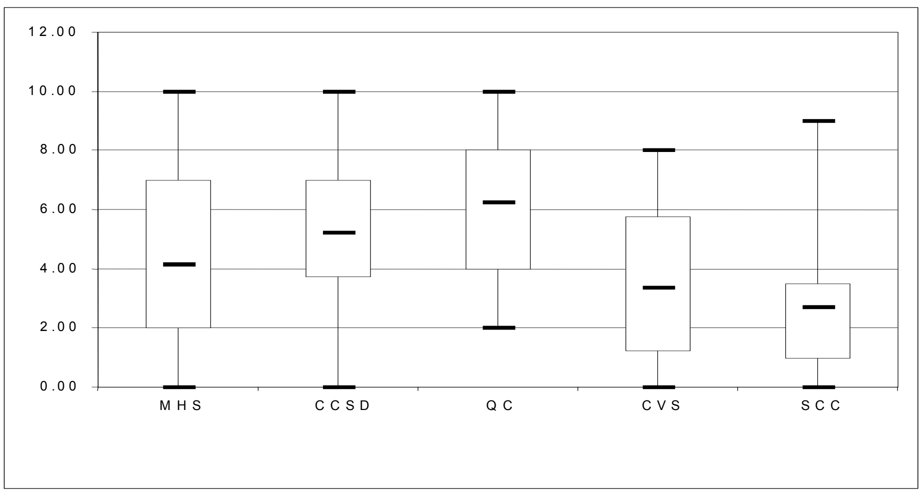


Figure 4. Performance on Oxygen in Different Schools

Results of Student Performance (by Sex)

There are no obvious surface features in the task which suggests that the *Oxygen* would be easier for boys than for girls. However, there are concerns that computer-based tasks in general might favour boys. On initial inspection it appears a statistically significant difference exists between the mean scores of girls and those of boys, with boys outperforming girls ($t = 2.67$, $df = 161$, $p < 0.01$). However, the highest performing school was a single sex boys' school. To address this confounding effect, two techniques were applied. First, student scores were converted to z-scores based on the performance in each school. The result of this approach showed that the differences in mean scores between the sexes was not statistically significant ($t = 0.67$, $df = 161$, $p = 0.51$). Second, to account for the fact that different proportions of girls and boys were entered by different schools, only certain schools were entered into the next analysis. For this analysis, we selected schools where the Male/Female ratio is near to 1. In this instance schools were only entered into the analysis if the ratio of girls to boys entered lay between 2/3 and 3/2. The results of this analysis showed no significant sex differences in the mean scores on the task ($t = 1.15$, $df = 47$, $p = 0.26$).

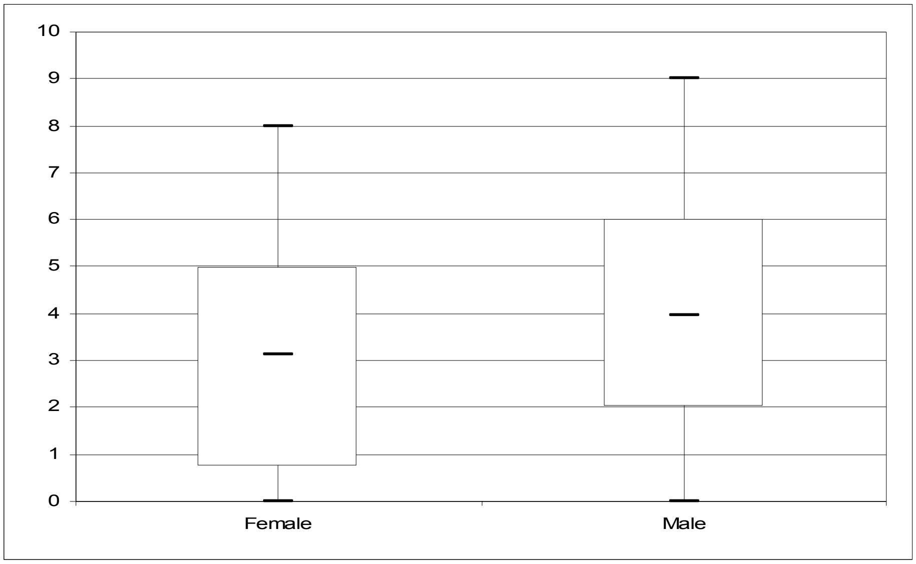


Figure 5. Performance on Oxygen by Sex in Schools with Male/Female ratio near 1

The same analyses have been carried out on *Chortlers* and *Cowpats*, and in neither case was there any indication that tasks were inequitable in terms of the sex of the students. On the basis of the data we have to hand (which also includes data from administrations of the whole test) there is little evidence of sex differences in performance on IT-based tasks.

Implications for Task Design

Curriculum tasks are designed to stimulate valuable learning experiences. Assessment tasks define curriculum ambitions, exemplify target behaviours, and show what students know, understand, and can do. Both kinds of tasks are central to the success of education. It is interesting, therefore, how little systematic research has been devoted to understanding the task design process (but see Ormerod & Ridgway, 1999; Ormerod, Fritz & Ridgway, 1999; and Samuda, Johnson & Ridgway, 2000). It is also interesting how few resources are devoted to the development of good assessment tasks. In the UK, this imbalance between the importance of high-stakes assessment and the resources devoted to the development of tasks is particularly marked. The costs of test design for some nationally administered tests are measured in hundreds, not thousands, of pounds, yet these tests are used to make decisions which affect the life-chances of students, the status of teachers and schools, and, indeed, the achievements of the education system as a whole. Such a system can work when tests each year are simply rather minor revisions of tests used in earlier years, but such a system is most unlikely to work when new sorts of test are required to assess new educational goals. Below, we sketch the processes used in the development of the new sorts of computer-based tasks produced for *WCT*.

The Task Development Process

Tests have a number of components, notably: tasks, a structure which defines test content, and an administrative system. Here, we focus on task development. Task development follows a particular sequence, which requires creative input and critical analysis. Experienced task designers often blend all these stages into a single process and have little insight into their own processes. Novice designers experience a multi-stage process characterised by setbacks, extensive revision, and often rejection, and wasted work.

First, an author (or authors) has an idea for a task, then creates a task. Often this will be discussed in a small group, and screen layouts will be sketched. The output from group meetings is often a set of connected ideas that are too big to form the basis of a single task. The author then produces a revised version of the task, which students should be able to complete in 10-15 minutes. At this point the task is in a state where critical analysis is useful. The author must describe the key activities, the substantive content, and develop ideas on how the task will be scored.

WCT tasks are written in FLASH by professional programmers. However, because the process of task design necessitates a great deal of revision, programmers are sent task specifications at a very late stage of development, in order to minimise change requests. The bulk of the task development process is done in-house, by our design team, working in either EXCEL or FLASH. Typically, once a task idea is well developed, the task can be created in about a day (excluding data recording). Either formally or implicitly using the criteria set out below (Figure 6), prototype tasks are discussed by the group.

Do the task in the way you would expect a good student in the target group to do it.

Is the purpose of the task clear to the students? Is the language clear?

Are the contexts natural and fully explained? Is the task really sensible in its context? Can the task length be justified in terms of the information gained about student knowledge?

How accessible is the task? Is there an easy entry point? Is there sufficient demand at the end?

Is there sufficient time for students to: complete the task, explore alternative strategies, check their work?

Describe the Task: What is being assessed? (Which problem solving strategies? Which concepts and connections? Which skills? What knowledge?) Is the task worth being able to do? (Does it stand as a worthwhile task in its own right? What class of worthwhile problems does it belong to?)

Have different solution methods been anticipated or encouraged? (Is there really just one solution method? If so, what is it? What are the alternative approaches?)

Evaluate the Marking Scheme: Is it unambiguous? Does it reward important skills appropriately? Does it reward a variety of solution methods?

Look at Student Work: Does the scoring scheme cover the variety of student work? Are there ambiguities? Does the task differentiate between students?

Figure 6. Criteria for Task Evaluation

Once the task is judged to be good enough, it is trialled with a small group of students – usually three individuals. Students are asked to take the role of ‘co-developers’ where they work on the task and are observed as they work. We ask them what they think the task assesses, about any aspects that are not clear, and about ways to improve the task. The task may be rewritten in the light of these trials.

After a task is deemed to be acceptable, it is sent for (re)programming. It is then trialled with a larger sample (set in a standard test administration ‘shell’) usually involving 20-30 students. The purpose of this level of trialling is to ensure that the task is accessible to students and that a range of student performances is obtained (i.e., that the task is neither impossible nor too easy). We use the range of student solutions to modify prototype marking schemes. In the final stage of trialling (i.e., the pre-test) tasks are assembled into tests in a way that reflects our test 'blueprint' or design specification. Here about 200 students take part. The pre-test provides information on the score distribution of tests and tasks, and allows us to fine tune the marking scheme.

Our primary ambition for task development is to produce a pre-test that is suitable for use as a live test. To achieve this, the design process has a number of important features:

- it is highly iterative;
- criteria for good tasks are considered from the outset;
- changes are made early, before extensive resources have been committed;
- outside groups are asked for well-defined semi-final products; their creative input is actively discouraged;
- tasks are trialled extensively before they are used in high-stakes situations.

Developing a Better Curriculum

We believe the current statistics curriculum is inadequate to equip students to become informed citizens in the 21st century where they will be required to deal with complex data sets. There is good evidence that appropriate IT support can make the understanding of data far easier for quite young students, for example by presenting new sorts of tasks where dynamic displays show changes in several variables over time. Interaction makes computers well suited to discovering rules and finding relationships. Furthermore students can work with complex realistic data sets, using methods similar to those used by professionals in the field (Ridgway & McCusker, 2003). We have only just begun to explore the qualitative aspects of this phenomenon, as we create new interfaces.

A Modular Approach

Our current research has two distinct features. First is the development of curriculum modules (at age 9 and 13 years) on data handling, which set out to capitalise on the enabling features of IT discovered so far (this complements two other modules at each age on other aspects of problem solving). Second is the development of new interfaces and tasks and research into the development of statistical competence.

Sample Module: Seeking Relationships in Data

A module consists of 4 one-hour units. In the case of *Seeking Relationships in Data* (MARS, 2004), which was designed for use with 13 year-old students, the components are:

- Lesson 1: Controlling variables
- Lesson 2: Finding relationships in situations with one response variable and two explanatory variables
- Lesson 3: Finding relationships in situations with one response variable and three explanatory variables
- Lesson 4: Finding relationships from graphs

In Lesson 1, two situations are presented. In the first, students explore the effects of two types of plant food on plant growth via a simulation, using (hopefully) a systematic search procedure. In the second task, students work with a system that calculates how long a hike will take, given the vertical and horizontal distances travelled. In Lesson 2, students work with simulations in mechanics and optics to discover inverse proportional relationships. In Lesson 3 the context is the design of bridges. Students can change the length, width and thickness of a plank bridge. They must discover linear, inverse and inverse squared relationships, then, combine them into a single equation to predict the strength of any given bridge. In Lesson 4, students are introduced to the power of exploratory data analysis through an interface where the axes are given some of the functionality of spreadsheets (see Figure 7). Given various biometric data, students use the computer to generate and study graphs in order to predict the speed at which a bird can fly. As part of the exploration, students can easily create new variables (such as 'weight/wing area').

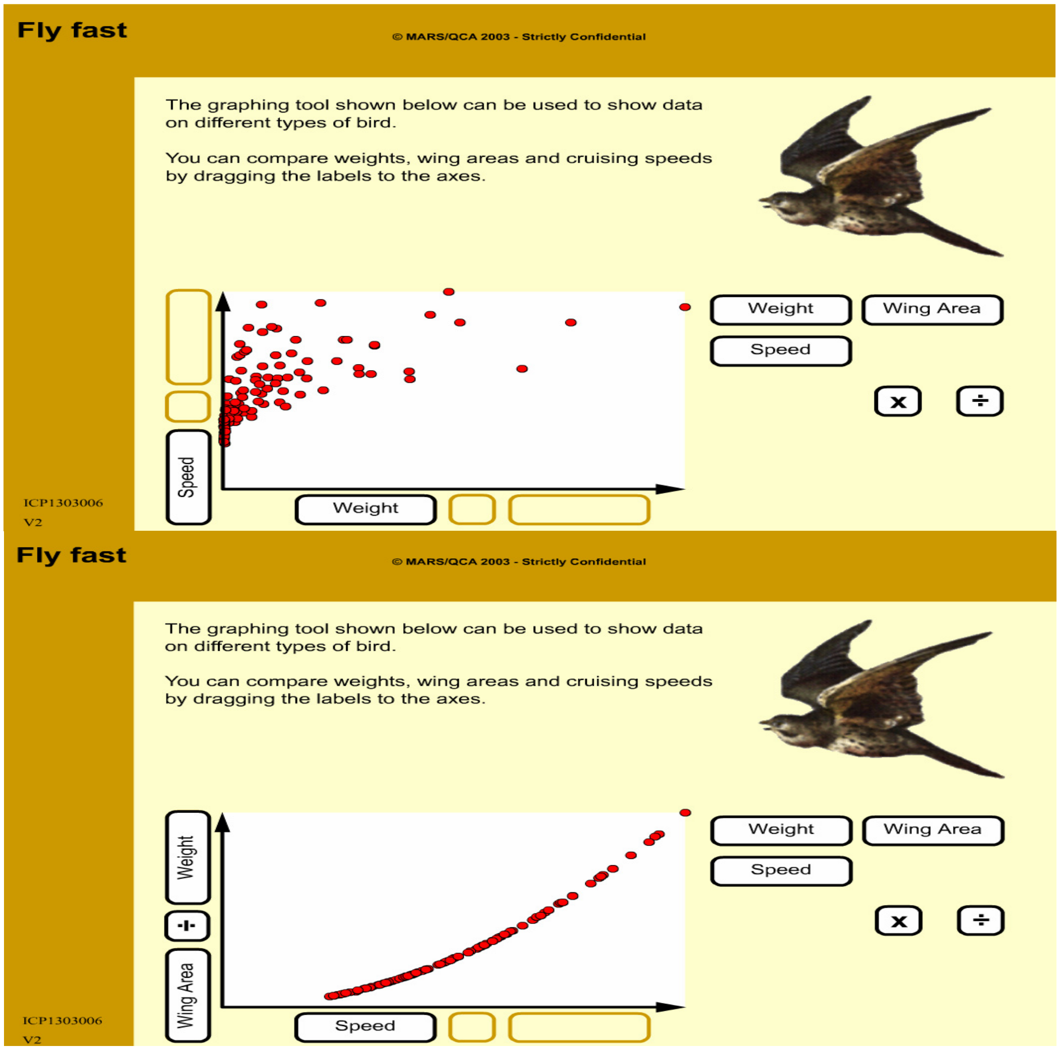


Figure 7. Screenshots of the Fly Fast Task using Functional Axes

Understanding the Applications of IT

We are continuing our explorations of the features of interfaces which actually make them effective. This work extends our test development into the classroom where we can develop more extensive task situations and simulations and can provide far more support for students in the form of teacher guidance and support materials (such as examples of other students' work, checklists, and hint sheets).

Conclusions

Statistics education in the UK falls far short of the ambitions of Nightingale and Wells. There is too little advocacy of the use of statistical techniques to inform thinking in key curriculum subjects, and the simplifications made in statistics courses actually go against Einstein's advice, making things too simple. Multivariate realistic problems (which students can actually solve) are ignored, and the formal statistics curriculum is reduced to practicing algorithms out of context, for which the range of applications is limited. IT can be used to present realistic, multidimensional data to students. Evidence from the *World Class Tests* shows that students—some as young as nine years old—can produce sensible conclusions after working with such data to reason about realistic situations.

IT has a number of unique features that make it particularly well suited to teaching and assessing skills in handling complex data. Evidence shows strong school effects (which suggests that data handling skills can be taught effectively) but few sex differences (which suggest that the format does not unfairly advantage or disadvantage males or females). The process of designing effective tasks has just begun, and some solid design principles are emerging. Tasks that are designed for assessment purposes are sufficiently rich to form the basis for valuable curriculum materials.

References

Batanero, C., Godino, J. D., Vallecillos, A., Green, D. & Holmes, P. (1994). Errors and difficulties in understanding elementary statistical concepts. *International Journal of Mathematics, Education, Science and Technology, 25*(4): 527-547.

Einstein, A (1933). On The Method Of Theoretical Physics - The Herbert Spencer lecture, delivered at Oxford, June 10, 1933*, Mein Weitbild*, Amsterdam: Querido Verlag, 1934 translated in *Ideas and opinions / by Albert Einstein ; based on Mein Weltbild*, edited by Carl Seelig, and other sources, new translations [from the German] and revisions by Sonja Bargmann.(1954). New York: WINGS Books.

Greer, B. (2000). Statistical Thinking and Learning. *Mathematical Thinking and Learning, 2*(1&2): 1-9.

Hawkins, A. (1997). Myth-Conceptions! In J.B.Garfield and G. Burrill (Eds.), *Research on the Role of Technology in Teaching and Learning Statistics.* Voorburg: International Statistical Institute.

Holmes, P. (2000). Statistics across the English National Curriculum. Nottingham, UK: Royal Statistical Society Centre for Statistical Education. Available online at: www.rsscse.org.uk/resources/natcur.html.

MARS (2004). Seeking Relationships in Data. London: Granada Learning.

Nicholson, J. R., Mulhern, G. & Hunt, D. N. (2002). Wizardry or Pedagogy? What is the driving force in the use of the new technology in teaching statistics? *Proceedings of ICOTS VI*. Voorburg, the Netherlands: International Statistical Institute.

Nicholson, J.R. (2003). GCSE statistics coursework, and dealing with multi-dimensional data. *Mathematics in Schools, 32*(3): 8-14.

Nightingale, F. cited in Agnew (1958). Florence Nightingale Statistician. *American Journal of Nursing, 58*(5), cited in Porter, S. (2001). Nightingale's realist philosophy of science. *Nursing Philosophy, 2*(1): 14.

Ormerod, T. C. & Ridgway, J. (1999). Developing task design guides through studies of expertise. In S. Bagnara (Ed.), *3rd European Conference on Cognitive Science* (ECCS, 1999), Siena: Consiglio Nazionale Delle Richerche. pp. 401-410.

Ormerod, T. C., Fritz, C. O. & Ridgway, J. (1999). From deep to superficial categorisation with increasing expertise. *Proceedings of the 21st conference of the Cognitive Science Society*, Vancouver, Canada, pp.502-506.

Ridgway, J. & McCusker, S. (2003). Using computers to assess new educational goals. *Assessment in Education*, *10*(3): 309-328.

Richardson, M., Baird, J-A., Ridgway, J., Ripley, M., Shorrocks-Taylor, D. & Swan, M. (2002). Challenging minds? Students' perceptions of computer-based World Class Tests of problem solving. *Computers in Human Behavior, 18*(6), 633-649.

Samuda, V., Johnson, K. & Ridgway, J. (2000). *Designing language learning tasks: A guide.* Working Papers on Task Design. Department of Linguistics and Modern English Language, University of Lancaster, UK.

Tversky, A. & Kahneman, D. (1983). Extension versus intuitive reason: The conjunction fallacy in probability judgement. *Psychological Review*, *90*: 293-315.

Wells, H.G. (1903). Mankind in the Making. London: Chapman and Hall: reproduced as an e-book downloaded via Kessinger Publishing, LLC. pp. 108. Available online at: http://etext.library.adelaide.edu.au/w/wells/hg/w45ma/.

World Class Arena (2005). To view 'Oxygen' go to www.worldclassarena.org/v5/default.htm <online challenge><problem solving 12-14><start test><problem 5 [Oxygen]>.

Paper Appendix: Scoring Rubrics for OXYGEN, CHORTLERS, & COWPATS Tasks

The scoring rubrics for the three assessment tasks are shown below. Experience of working through the tasks on a computer, and seeing some student responses, would be needed to make the application of the scoring rubric transparent. Nevertheless, we hope their inclusion will help to illuminate the tasks for the reader.

Q	**Oxygen: Age 13**		Points	Section points
	The core elements of performance required by this task are: • explore relationships among 3 variables in a scientific context; • identify optimal values. Based on these, credit for specific aspects of performance should be assigned as follows: *Mark entirely from paper*			
1	Oxygen production: • is not affected by light intensities below 5 • increases with light intensity above 5. *Part mark:* If pupil states that 'as light increases, more oxygen' : • increases as temperature rises up to 30°C • decreases as temperature rises over 30°C *Part mark:* If pupil states that 'best temperature is 30°C' :		1 1 (1) 1 1 (1)	4
2	They could both be right. *Reason:* They hadn't controlled the light intensity. or Ann was looking at low light intensity Jim was looking at high light intensity.		1 1	2
3	Ann 's idea is better than Jim's. Reason: e.g. Jim's idea might overheat the plants. A better idea still would be to have Temp = 30°C, Light intensity =50.		1 1 2	4
	Total Points		10	10

Q	box	PC/ Paper	**Chortlers: Age 13**	Points	Section Points
			Computer scoring*/Manual scoring/*Guidance		
1	1.1		*Words to the effect of:* "Chortler numbers are **only** affected by summer rainfall" **Part marks** *Award the marks shown for each of the following statements:* "Chortler numbers are not affected by hawk numbers" "Chortler numbers are not affected by pollution" "Chortler numbers appear to be affected by summer rain" (but doesn't eliminate other possibilities)	4 (1) (1) (2)	4
	1.2		*Suggests a presentation format –* ***either*** A graph of chortler numbers vs. summer rainfall (or vice versa) ***or*** "Group Data" by chortler numbers ***or*** "Group Data" by summer rainfall.	2	2
	1.3		*Describes how that presentation supports the hypothesis**, e.g.:*** "The graph shows that Chortler numbers go down as the rainfall goes up" or "Low chortler numbers (20-29%) are associated with high rain (329mm)".	2	2
				Total:	**8**

Q	box	PC/ Paper	**Cowpats: Age 13**	Points	Section Points
			***Computer scoring**/Manual scoring/*Guidance		
1	A		*Student should recognise sequence of events, i.e.:* Starts with lots of dung Flies start to arrive. Worms start to arrive Beetles start to arrive. Dung starts to disappear. Worms flies and beetles start to disappear. *Deduct 1 mark for each significant omission or clearly incorrect observation.*	3	3
2	B		**gorder=("g2","g3","g4","g1")** *i.e. students have to put events in the correct order*	2	2
3	C		**ans3nouns contains "T2" (Beetles)** *i.e. has recognised that only the beetles seem to affect the amount of dung.*	1	
	D		**ans3verbs contains only some combination of "T4" (Bury), "T5" (eat) and "_null"** **and is not (_null,_null)** *i.e. sees that only burying or eating the dung will make it go away, doesn't mention eggs.*	1	
	E		**ans3nouns=("T2","T2") and (ans3verbs=("T4","T5") or ans3verbs=("T5","T4")** *i.e.students choose the only plausible actors nbd events*	2	4
4	F		Cows drop dung everywhere, but the cowpats on the fields disappear quickly…	1	
	G		…because the beetles can't bury dung in a concrete drive. **Part mark** … because worms can't get to the cowpat if it's on concrete.	2 (1)	3
				Total:	**12**