International Electronic Journal of
Mathematics Education

Volume 2, Number 3, October 2007 www.iejme.com

REASONING WITH MULTIVARIATE EVIDENCE

Jim Ridgway

James Nicholson

Sean McCusker

ABSTRACT. We report a study where 195 students aged 12 to 15 years were presented with computer-based tasks that require reasoning with multivariate data, together with paper-based tasks from a well established scale of statistical literacy. The computer tasks were cognitively more complex, but were only slightly more difficult than paper tasks. All the tasks fitted well onto a single Rasch scale. Implications for the curriculum, and public presentations of data are discussed.

KEYWORDS. Multivariariate Data, Curriculum, Rasch Scaling, Hierarchies, Statistical Reasoning, Statistical Literacy, Assessment.

INTRODUCTION

Reasoning with evidence about realistic situations - education, health, crime, social change, climate change - is inherently problematic. Evidence is often multivariate; relationships between variables are rarely linear; variables interact, and show effects of different sizes, over different timelines. If people are to use evidence to inform and understand public debate, to function in complex, fast-moving commercial environments, and to make important personal decisions, these sophisticated statistical notions need to be part of ‘common sense’. For this to happen, we need to understand just how difficult it is to acquire these sophisticated notions, we need curriculum activities that develop these ideas, and we need some ways to assess the acquisition of these skills, if they are to be taken seriously by students and teachers. Here, we focus on the difficulty of reasoning using multivariate data, and explore possible consequences in the development of reasoning from multivariate data.

Statistical Literacy

A variety of educational communities have emphasised the role of education as an agent in social reform. Condorcet (1994/1792) discusses *le savoir liberateur*. His vision was that

ISSN: 1306-3030

education should improve the lives of all citizens, and would provide the basis for a vibrant democracy. Dewey (1999) refers to 'liberating literacy', 'popular enlightenment', and the ability to 'judge independently', in part by being able to critique the quality of the evidence presented. Orrill (2001) refers to Cremin's (1988) useful distinction between 'inert' and 'liberating' literacy. The former refers to the ability to understand instructions, and to be able to act out the rituals of a society by performing routine acts in one's daily life. The latter refers to the ability to find and critique information, and to use it when making decisions.

The core concept underlying 'literacy' is the ability to function effectively using the everyday communication tools embedded in social functioning; 'numeracy' is a neologism coined to capture the idea of a set of mathematical skills analogous to literacy. The key performances that constitute effective literacy and numeracy are located in place and time. In a pre-literate society, one might survive perfectly well without reading (although possessing an effective repertoire of mnemonics might convey considerable advantages). In modern societies, illiteracy and non-numeracy can pose a serious barrier to personal effectiveness. However, there are important problems for society posed by non-numeracy and illiteracy at the level of citizenship and personal engagement in the political process, that correspond to an absence of *le savoir liberateur* and of 'liberating literacy'. The movement towards 'evidence informed policy' requires that citizens must be able to understand evidence and arguments based on evidence in order to engage with contemporary political debates.

Statistical theory emerged relatively recently as a branch of applied mathematics, in response to practical problems (e.g. Gigerenzer *et al*., 1989). Early advances in the invention of statistics were made before the invention of computers, and techniques were created in response to severe constraints on computational power. Commonly used techniques make strong assumptions about the nature of the data, and were designed to model data that were considerably less complex than the data in the social sciences.

Silva (2006) offered an heroic survey of the emerging nature of the discipline of statistics, and mapped out changes in the tools available to statisticians and for the training of statisticians. He argued that the dominant tool in the 1970s was the desk calculator, and that teaching emphasized the mastery of statistical technique. The1980s was the era of powerful general purpose packages (e.g. SPSS, SAS, BMPD, then EDA, GLIM and CART), and teaching emphasized understanding a wider range of statistical models. The 1990s saw the introduction of graphical displays, spreadsheets, and simulations, and greater emphasis on representing and understanding data, and statistical modelling. The discipline of statistics is seeing the development of specific statistical tools and models targeted towards particular problem areas, such as biometrics, demography, and econometrics. Overall, according to Silva's analysis, the discipline of statistics is becoming increasingly data and problem driven, and is less focused on generally applicable techniques. Saying this another way, the emphasis of professional statisticians is now more focused on the creation of statistical models to fit interesting problems, than on fitting interesting problems into standard statistical models.

How have these developments in professional practice trickled down to universities and schools? Chance (2002) argued that undergraduate programmes are still dominated by the mastery of statistical technique to the exclusion of interpretation, and reasoning with evidence. Ridgway, Nicholson and McCusker (2007a) analysed statistics examinations used to select students for university places in the UK, in terms of content (never more than two variables; relationships are always linear) and in terms of the proportion of assessment devoted to technical mastery or modelling and interpretation skills (technical mastery dominates)[1]. They concluded that the UK education system does very little to develop 'statistical literacy' in the sense of understanding the role of complex data relationships in informing policy or for other decision-making.

It may be the case that reasoning with multivariate data is actually very difficult; certainly, students have traditionally struggled to master relatively simple statistical concepts (e.g. Batanero, Godino, Vallecillos, Green, & Holmes, 1994). Schield (2006) identifies problems that undergraduates, college teachers, and professional data analysts have in drawing conclusions from simple data displays. In one example (see http://www.statlit.org/Survey/), respondents are shown a pie chart of smokers, 20% of whom are Catholics, and 40% are Protestants. Over 60% of students and data analysts agreed with the incorrect statement "Protestants (40%) are twice as likely to be smokers as are Catholics (20%)". (A correct statement would be "Protestants are twice as likely among smokers as are Catholics"). Other research dealing with students' or teachers' difficulties in understanding statistical graphs are included in this issue (Aoyama, 2007; Monteiro & Ainley, 2007).

THE RESEARCH PROBLEM

New technology offers opportunities for change. The World Class Arena (WCA) project (http://www.worldclassarena.org/) was set out to assess *inter alia* the problem solving skills of high attaining students in science, mathematics and technology, via computer. Tests were designed for students aged 9 and 13 years. A wide range of novel tasks was designed; most relevant for the current discussion are the tasks which required students to reason from realistic, multivariate data. We were able to show that high-attaining students aged 9 years can work effectively with multivariate data, when presented via new computer interfaces which gave students control over the ways data were presented, and which often displayed data dynamically (Ridgway & McCusker, 2003). This work provides some evidence of what high-attaining students can do, and raises some important questions for educational practice in general.

The focus of this paper is an investigation of the difficulty that representative students (as opposed to high attaining students) aged around 13 years have in reasoning with multivariate

[1] An analysis of 11 exemplar questions can be found at www.durham.ac.uk/smart.centre/other_publications.

data, in comparison with the difficulties they experience handling simpler statistical challenges. We also explore the extent to which reasoning from multivariate data relates to other components of statistical literacy, and outline some ideas on the ways that reasoning from multivariate data develops.These issues are important for a number of reasons. First is the urgent need for citizens to engage effectively with realistic data, in order to understand arguments about social policy and to make informed decisions about their personal affairs. Students leaving school or college will be working in environments where complexity is a fact of life. Almost all the information presented in large databases on the internet (e.g. World Health Organisation; National Census data) is presented in a static, tabular form. We present evidence to suggest that better interfaces (notably interfaces which present multivariate data dynamically, under user control) will make these data intelligible to far more people. A second set of reasons relates directly to education. Evidence presented here and elsewhere (Ridgway, Nicholson & McCusker, 2007b, 2007c) suggests that the curriculum (in geography, citizenship, science, psychology etc.) could be made more realistic and more relevant to students, so that they could leave school with a far greater understanding of ways to work with multivariate data.

Understanding the structure of statistical literacy

Watson and Callingham (2003) and Callingham and Watson (2005) have done some interesting and important work to understand the structure and logical development of statistical reasoning (and literacy). They developed a number of paper-based tasks designed to assess different components (sampling, chance, variation and average) and different levels of statistical literacy, gave the tasks to a large sample of students with a broad spread of abilities, and analysed the data using a Partial Credit Rasch model. (for a discussion of why Rasch scaling is appropriate for the study of development, see Ridgway, 1997; for further discussion and practical examples, see Bond & Fox, 2001). Watson and Callingham described a hierarchy of statistical literacy evident in the data, which they proposed as a developmental model for statistical literacy. Their six-level model is shown in Table 1.

Table 1. A hierarchy of statistical literacy skills (Watson & Callingham, 2003).

Level 6. Mathematical Critical: questioning engagement with context, using proportional reasoning particularly in media or chance contexts, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language.

Level 5. Critical: questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, qualitative interpretation of chance, and appreciation of variation.

Level 4. Consistent Non-Critical: appropriate but non-critical engagement with context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities, and graph characteristics.

Level 3. Inconsistent: selective engagement with context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.

Level 2. Informal: only colloquial or informal engagement with context often reflecting intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph, and chance calculations.

Level 1. Idiosyncratic: idiosyncratic engagement with context, tautological use of terminology, and demonstrating basic mathematical skills associated with one-to-one counting and reading cell values in tables.

The fact that tasks can be located on a single scale does not in itself mean that a single attribute (e.g. ‘statistical literacy’) is being measured. A single scale is consistent with this view, but so too is the idea that the single scale represents student performances on different cognitive dimensions that are highly correlated. Conceptual analysis is an essential component in understanding evidence, and must not be subjugated by naïve empiricism. The identification of levels is done via a mixture of observation and professional judgment. For example, inspection of the distribution of task difficulties can reveal ‘notches’ – locations along the difficulty continuum where there are very few tasks. These would occur if these places corresponded to boundaries between cognitive stages – tasks above the notch require a higher level of cognitive functioning than do tasks below the notch. A notch could also occur if tasks taken by students had an inadequate range of difficulties. It follows that notches highlight places to look; the analysis of tasks and student responses is essential to any judgments about the nature of task demands and of student performances around possible boundaries between different performance levels.

The Callingham and Watson studies are based entirely on paper-based tasks that require students to show simple statistical skills and appropriate use of terminology, qualitative and quantitative interpretation of chance, some understanding of variation and the need for uncertainty in making predictions, skill in understanding data representations and in drawing inferences from simple data sets. Here, we set out to see how more complex tasks might fit into their hierarchy, with two key questions in mind:

- Are the more cognitively complex computer-based tasks actually more difficult (and perhaps too difficult) when scaled against paper tasks?
- Can reasoning from multivariate data be seen as an integral part of statistical literacy, or does it assess other cognitive skills?

THE STUDY

195 students from an academically selective school in Northern Ireland and a comprehensive school (with no academic selection at entry) in the North East of England took part in the study. Student ages ranged from 12 to 15 years. Tests were assembled that included paper-based tasks from the Watson and Callingham studies, a new paper-based task that required reasoning with multivariate data, and computer-based tasks from the WCA work. These were administered by the students' teachers, in the presence of an observer from the research team. Tests lasted approximately 70 minutes. A brief description of the five computer-based tasks is given in Figure 1 (sample working tasks can be found at http://www.worldclassarena.org/), followed by a brief description of the seven paper-based tasks (Figure 2).

Waterfleas (WF): Students can vary conditions of pollutant and temperature and see the effect that this has on water fleas in a beaker. The students are then presented with a series of claims about the effects of temperature and pollutant and have to judge the accuracy of these claims.

Rare Fish (RF): Students are presented with data representing a population of rare fish, over time. Alongside these data they have access to data concerning rainfall, seagull numbers and temperature. A number of scenarios are described which might account for the decline in the number of rare fish, and students are expected to place these scenarios in order of plausibility, based on the data.

Oxygen (Oxy): Students are presented with a graphical representation of the oxygen production of plants, based on ambient lighting and heating conditions. Students are required to evaluate a range of statements, resolve an apparent paradox of experimental design, choose a set of conditions to maximise oxygen production and to explain their reasoning.

Bingo (B): Bingo numbers are called according to the product of 2 random dice. Students are asked to evaluate the likelihood of winning of a series of bingo cards and are then asked to create their own 'best' card.

Big Wheel (BW): Students are required to interpret the data from a sinusoidal curve representing the height of a point on a fairground Big Wheel, describe the information represented by each of 3 parameters and modify these parameters to match a fixed specification.

Figure 1. Computer tasks

Smoking (T2X2): Students are presented with a 2 by 2 Table of representing the survey results of smokers and non-smokers with and without Lung Disease. They are then asked if the results for the occurrence lung disease indicated an association with smoking and to explain their reasoning.

School (TRV): Students are presented with a pictograph of the methods by which children arrive at school. They are then asked a series of questions involving ideas about variability and uncertainty.

House Prices (HSE): Students are presented with an excerpt from a media article about house prices and are asked to interpret the statistical terminology used within the article.

Mobile Phone (MPDif): Students are required to interpret multivariate data concerning mobile phone ownership.

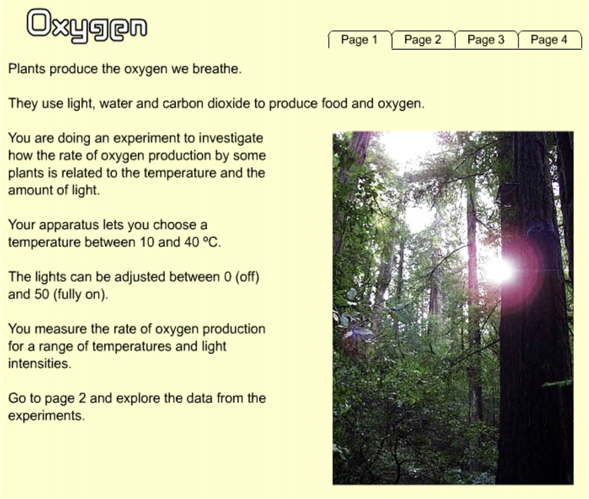
Toss Up (SP): Students are required to interpret histogram data concerning a coin tossing experiment, firstly in descriptive terms then in terms of the likelihood that particular histograms are ‘made up’.

Handguns (M7CH): Students are presented with an excerpt concerning use of handguns in one US city and asked about their perception of risk in another US city.

Killer Cars (M8QU): Students are asked to question a media excerpt claiming a relationship between the rise of car use and corresponding rise in heart deaths over the previous twenty years.

Figure 2. The seven paper-based tasks

Screen dumps of the computer-based task Oxygen are shown below as Figures 3.1 – 3.4, followed by a description of the scoring rubric in Table 2. The paper-based task Handguns is shown below as Figure 4, and the associated scoring rubric is shown in Table 3. A complete set of tasks used in the study is given in Appendix 1. Students were encouraged to ‘do their best’ on each task, and not to rush through the complete set of tasks.



Plants produce the oxygen we breathe.

They use light, water and carbon dioxide to produce food and oxygen.

You are doing an experiment to investigate how the rate of oxygen production by some plants is related to the temperature and the amount of light.

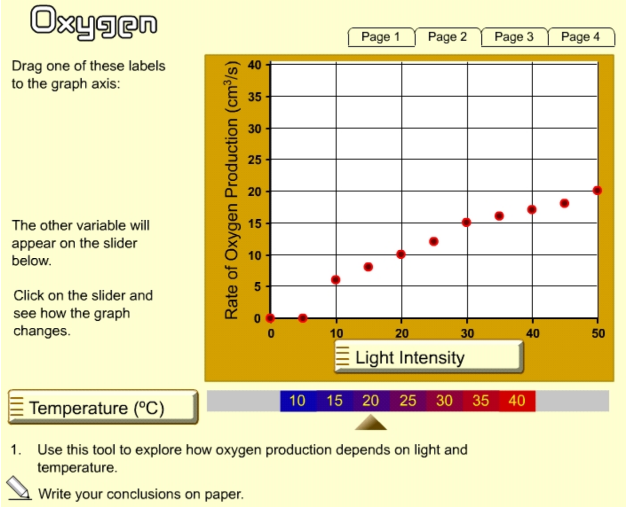
Your apparatus lets you choose a temperature between 10 and 40 ºC.

The lights can be adjusted between 0 (off) and 50 (fully on).

You measure the rate of oxygen production for a range of temperatures and light intensities.

Go to page 2 and explore the data from the experiments.

Figure 3.1. Page 1 of Oxygen Task



Drag one of these labels to the graph axis:

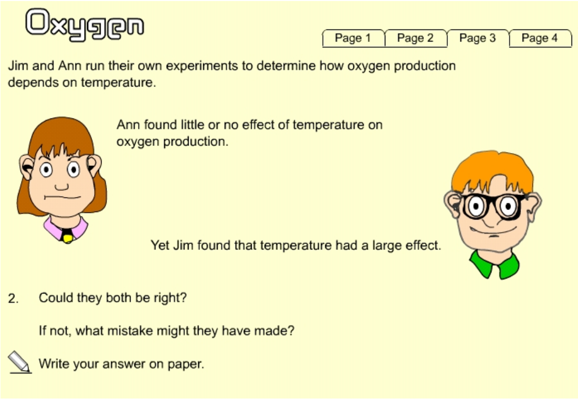
The other variable will appear on the slider below.

Click on the slider and see how the graph changes.

1. Use this tool to explore how oxygen production depends on light and temperature.

Write your conclusions on paper.

Figure 3.2. Page 2 of Oxygen Task



Jim and Ann run their own experiments to determine how oxygen production depends on temperature.

Ann found little or no effect of temperature on oxygen production.

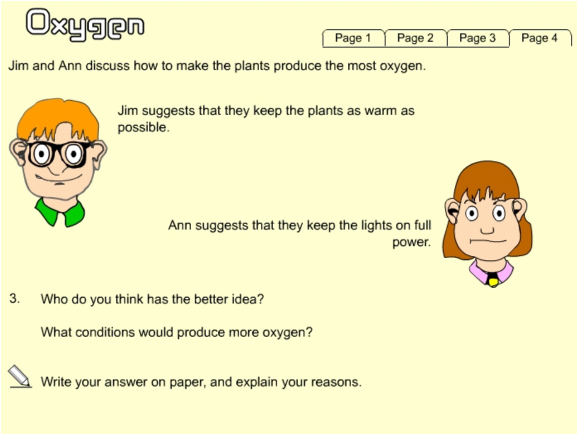
Yet Jim found that temperature had a large effect.

2. Could they both be right?

If not, what mistake might they have made?

Write your answer on paper.

Figure 3.3. Page 3 of Oxygen Task



Jim and Ann discuss how to make the plants produce the most oxygen.

Jim suggests that they keep the plants as warm as possible.

Ann suggests that they keep the lights on full power.

3. Who do you think has the better idea?

What conditions would produce more oxygen?

Write your answer on paper, and explain your reasons.

Figure 3.4. Page 4 of Oxygen Task

Handguns. Would you make any criticisms of the claims in this article? If you were a high school teacher, would this report make you refuse a job offer somewhere else in the United States, say Colorado or Arizona? Why or why not?

> ABOUT six in 10 United States high school students say they could get a handgun if they wanted one, a third of them within an hour, a survey shows. The poll of 2508 junior and senior high school students in Chicago also found 15 per cent had actually carried a handgun within the past 30 days, with 4 per cent taking one to school

Figure 4. Handguns paper-based task

Table 2. Scoring rubric for the Oxygen computer-based task

Question	Possible Responses	Points	Section points
Q1. Oxygen production:	is not affected by light intensities below 5	1	
	increases with light intensity above 5.	1	
	Part mark: If pupil states that 'as light increases, more oxygen is produced' then give (1).		
	• increases as temperature rises up to 30°C	1	
	• decreases as temperature rises over 30°C	1	4
	Part mark: If pupil states that best temperature is 30°C then give (1)		
Q2. Could both be right?	Correct (Ignore reason for mark)	1	
	Reason:		
	They hadn't controlled the light intensity.	1	2
	Part mark: If pupil states that Ann was looking at low light intensity and Jim was looking at high light intensity.	(1)	
Q3. Who has better idea?	Ann 's idea is better than Jim's. (Ignore reason for mark)	1	
	Reason: e.g. Jim's idea might overheat the plants.	1	
	A better idea still would be to have Temp = 30°C with light intensity =50.	1	4
	Total Points	10	10

Table 3. Scoring rubric for the *Handguns* paper-based task

Code 4	(a) Only Chicago has been asked
Code 3	No, 2508 is small sample; (b) Maybe not in Arizona
Code 2	No, not everyone; Reliability of measurement: No, could be lying; Whole of USA would be the same
Code 1	Shouldn't have guns
Code 0	No Response

Data Analysis

In the case of the Watson and Callingham tasks, the 'partial credit' scoring system was taken from the original Watson and Callingham studies. Essentially, tasks have a number of components (we call them 'items'). For each item, student responses are categorized in order of sophistication and accuracy, and students are assigned a number label that corresponds to their response (Full details of the scoring rubrics are available from the authors). Some items will have several categories, others as few as 2. In the case of the World Class Arena tasks, the number of categories was adapted from the scoring systems developed for WCA. On WCA, a student's test score is an aggregation of their marks obtained by the application of a scoring rubric to performance on each item in each task.

Here, for WCA items, each scoring rubric was analysed carefully; in cases where marks were allocated for progressively better solutions, these were treated as partial credit scores; where marks were given one at a time for unrelated components of performance, these were retained (not summed) as individual indicators of performance (so were each allocated a label of 0 or 1, depending on the student's success). It follows that items do not have the same number of categories as each other; codes for different items ranged from 0-1 to 0-5. Students who made no response to an item were allocated a code of zero if it was in the middle of their work, but if they did not reach items at the end of the test, no code was recorded, and the analysis treated this as though the item were not on their test. The surveys were all coded by one of the researchers who has extensive experience grading student work for high-stakes assessment. These data were then analysed via partial credit Rasch scaling. This produces an odds-ratio scale where every item can be related to every other item in terms of its difficulty.

RESULTS

Usually an item-map produced by Rasch scaling will display the item difficulty against the student ability on each side of a figure. However, for the current analyses, the student ability display has been removed from the figure below to allow easy comparison between paper-based and computer-based tasks.

The scale represented in the middle of Figure 5 is the Rasch scale of difficulty. Items towards the top of Figure 5 with positive scale values are the more difficult items within the test. The figure has been edited to allow each item a column of its own. The levels of response to each item are represented by a suffix. E.g. OXY1 represents the first part of the task 'Oxygen' and OXY1_3 represents a code of '3' on OXY1 (increasing number codes refer to better student performances; in many cases, number codes can be interpreted as points awarded in conventional scoring systems). In cases where two item codes are at the same level, they are separated by a '/'. Some items appear with 2 coding levels e.g. B3.1 and B3.2. This refers to two separate aspects within the same task part. In this case, B3 (Bingo – Part 3) required the students to place numbers on a bingo card. B3.1 is associated with the selection of numbers and B3.2, with their positioning.

	Paper-based			Most Difficult		5			Computer-based		
						.					
						.	OXY1_4				
						.					
						.					
						4					
						.					
						.					
						.					
						.					
						.					
						3	OXY2_2	B3.2_3			RF3_3
						.	OXY3_4	B4._1/B3.1_3			
						.		B2.2_3			
	M7CH_4					.		B4.1_1			RF4_2_2
HSE3_2						.	OXY1_3	B1.2.2_1	BW2.2_2		
		SP11_2				.	OXY3_3				
						2			BW2.1_2		
						.					
	M7CH_3		M8QU_2			.					
						.		B3.1_2/B3.2_2			
						.	OXY2_1/OXY3_2	B2.2_2	BW2.3_2		
HSE1_2						.					RF2.2_3
					MPDifG_2	1	OXY1_2			WF2.4_2	RF3_2
		SP10				.		B1.2.1_2	BW2.2_1/BW3.3_1	WF2.3_2	
HSE2_2	M7CH_2					.					
HSE3_1						.					
						.		B2.2_1	BW3.2_1	WF2.2_1	RF2.2_2
		SP9/SP7				.					RF4_1
						0		B3.1_1			
			M8QU_1			.	OXY3_1		BW2.1_1	WF2.4_1	
		SP11_1				.		B3.2_1	BW3.1_1	WF2.3_1	
		SP8				.		B1.2.1_1	BW2.3_1		RF2.2_1
HSE2_1/HSE1_1						.					
						.					RF3_1
						-1	OXY1_1				
	M7CH_1					.					
		SP6				.			BW1_1		
						.				WF1_1	RF2.1_1
					MPDifG_1	.					
						.					
						-2				WF2.1_1	
						.					
						.					
						.					
						.					RF1_1
				TRV2/TRV5		.					
						-3					
						.					
						.					
				TRV3		.					
						.					
						.					
				TRV1_1		-4					
						.					
						.					
						.					
						.					
						.					
				Least Difficult		-5					

Figure 5. Rasch Item Map of Paper-based and Computer-based Tasks

Computer-based items have been abbreviated as follows - OXY: Oxygen Task; B: Bingo Task; BW:Big Wheel Task; WF:Water Fleas Task; RF: Rare Fish Task. Paper-based items have been coded as follows (following the convention used in all the Watson and Callingham publications) - HSE:House Prices; M7CH: Handguns; SP:Toss Up; M8QU: Killer Cars; TRV: School. The newly written item is coded MPDifG: Mobile Phone.

Can Students from a Broad Ability Range Reason with Multivariate Data?

The computer-based tests were developed for high attaining students. One of our ambitions for this work was to explore the extent to which students from a broad ability spectrum can reason with multivariate data. Figure 5 shows that students from a broad ability range can make good progress on tasks that present multivariate data – these tasks are hardly more difficult than cognitively simpler tasks, presented on paper (discussed below). This is an important finding, and suggests that multivariate tasks are likely to be useful for curriculum and assessment purposes in mainstream education,

Comparing the Relative Difficulty of Computer and Paper-based Tasks

Examination of Figure 5 shows that the computer-based items cover much of the same spread of difficulties as the paper-based items. Given the nature of Rasch scales, and the opportunistic choice of multivariate tasks, it is appropriate to examine the relative difficulties of paper-based and computer-based items qualitatively. The few paper-based items at the bottom of the figure represent items requiring simple counting operations within the School task (e.g. students were given a pictograph and were asked 'how many pupils travelled to school by car'). Items at this level of difficulty had not been designed into the WCA, because their original function was to identify high attaining students. The computer-based tasks require reasoning with multivariate data, unlike Watson and Callingham's paper-based tasks. The computer-based items appear slightly more difficult than the paper-based items. Overall, however, there is no dramatic difference in difficulty between the paper-based task and the computer-based tasks. We conclude that the computer-based tasks, which require students to engage in reasoning with multivariate data, are no more difficult than the cognitively simpler paper-based tasks.

Is There a Single Scale of 'Reasoning with Data'?

The fit statistic in Rasch analysis is a measure of the extent to which any item or item part fits the model used for the analysis. The fit statistic can be used to identify or highlight items which may not fit the model or behave in the expected way. It must be stressed that just because an item does not meet the fit criteria (usually .77 < fit statistic <1.3) the item should not be rejected automatically. The fit statistic should be used as a guide and in concert with inspection of the item to see how the misfitting item should be treated. Two measures of fit are usually used, namely Infit and Outfit. Infit is an information weighted statistic and so is less susceptible to outliers. Outfit is an unweighted statistic, and so is sensitive to outliers. Ideally an item would lie between the fit boundaries for both fit statistics. However, the value of both fit statistics yields information about the item, which can then be used to make decisions concerning rejection or modification of the item or scoring scheme. Where Figure 5 showed the difficulty measure associated with each item within a task part, the graphs below show a mean measure of difficulty and fit associated with each task part within the test.

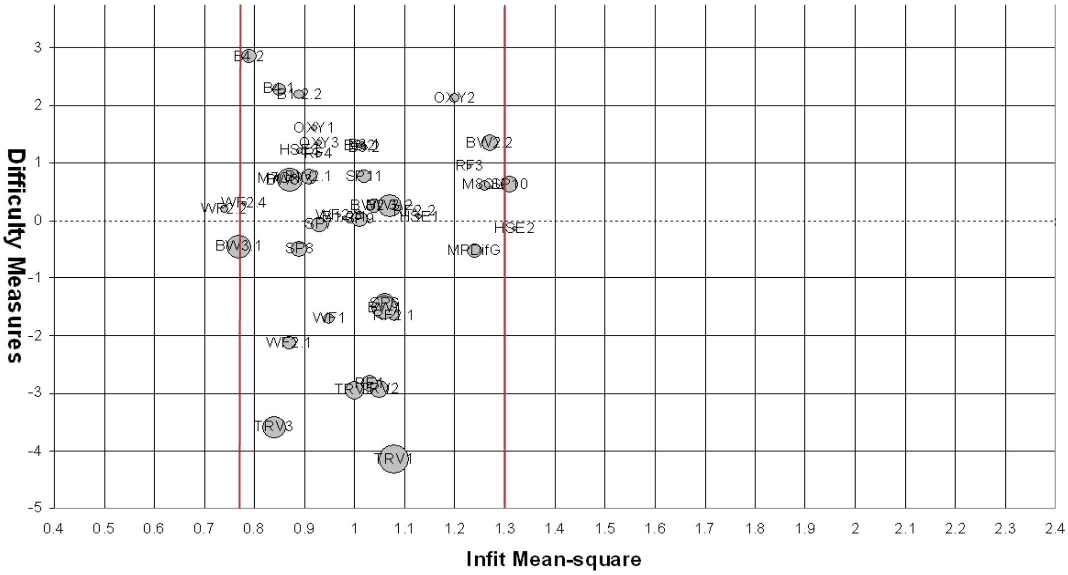


Figure 6. Infit statistic from Rasch scaling of test items

Figure 6 shows the infit statistic for all the item parts to be adequately within the conventional bounds (i.e. .77 < fit statistic <1.3) and could reasonably be used to support the assertion that the test is measuring a single scale. Looking at the Outfit (unweighted) statistic in Figure 7, the most striking feature is the high outfit measure associated with RF1 (Rare Fish Part1). This combination of good Infit and high Outfit usually implies that there was a certain degree of carelessness or guessing associated with the item, i.e. either students were getting it wrong when all other items indicate that it should have been achievable, or that students were getting it right when all other items indicated that they should be getting it wrong. At the design stage of the WCA items, great care was taken to ensure that marks were not awarded for items which could be guessed at. Inspection of the first part of the Rare Fish task shows that students were required to read fish populations for 3 separate years for a bar chart. The mark was awarded only if all three readings were correct. This is a fairly routine and simple task where carelessness on any one part could lead to achieving no marks.

A few of the items also show a low Outfit statistic with a good Infit statistic. This pattern is usually associated with outliers created by the assumptions being made about the unanswered items, for instance, awarding a zero to an unanswered item within a body of correct answers, where most people would have answered all items correctly. Overall, the Rasch model provides a good fit for the data. The Rasch analysis supports the idea of a single scale of ‘Reasoning From Data’ running through all the items and item parts.

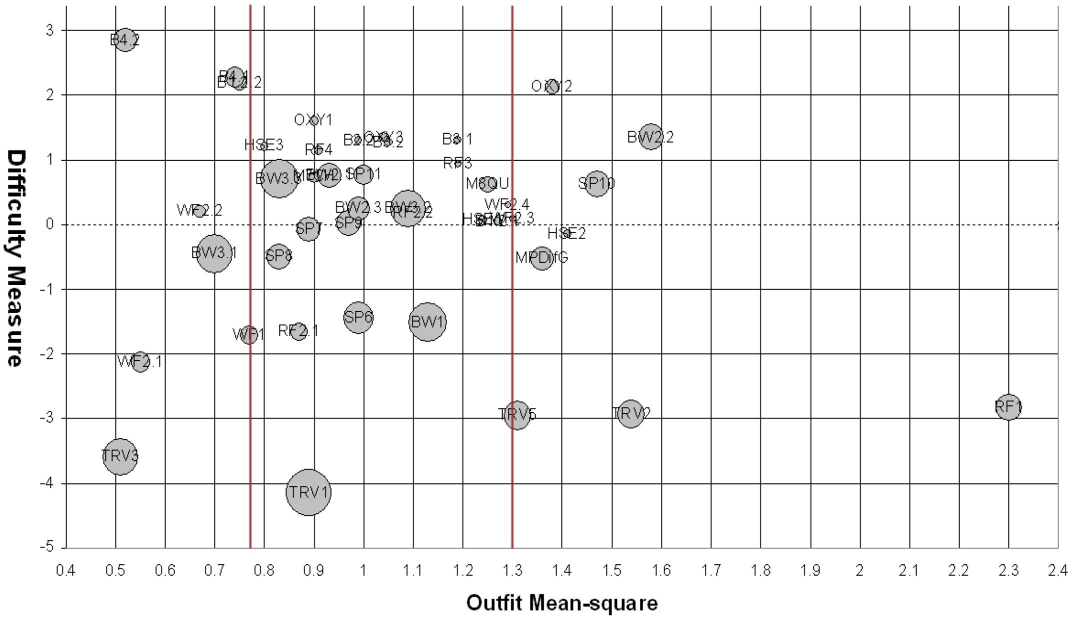


Figure 7. Outfit statistic from Rasch scaling of test items

When grading student work, we noticed that with multivariate problems, when more than one factor is seen to make a difference, almost all students can describe accurately the effects of one or other factor, but some will stop at that. Better responses move through making a brief reference to the second factor, to describing both factors fully, and the best can also describe interactions between the two factors. For example, in the Oxygen task, an example of a good response was: ‘There needs to be a reasonable light level before any oxygen is produced, and beyond that level the higher the light intensity, the more oxygen is produced at any given temperature, but the rate of oxygen production increases with temperature up to around 300 and decreases thereafter’.

It is interesting to note that items in which there were substantial differences in both explanatory variables were easier for students to deal with than very similar items where a substantial difference was present in only one of the explanatory variables. We conjecture that this result can be removed by appropriate teaching; students should discuss plausible sources of variation (here, sex of students on mobile phone ownership) which are not, in fact, associated with differences in the dependent variable.

The fact that the tasks presented here can be fitted to a single scale does not necessarily mean that such a scale has any psychological reality. It may be that each task measures distinct attributes which are highly correlated. To support the idea of a scale onto which all of these tasks fit, we need to inspect each task and fit every item to a position within a conceptual scale, be it one using the Solo taxonomy (Biggs and Collis, 1982), or the specific variant created by Watson and Callingham (2003) for understanding the development of statistical reasoning and literacy. The data presented here are rather sparse for drawing any precise conclusions about the nature of the hierarchy of 'Reasoning with Data'. Nevertheless the scale of computer-based items based items does lend itself to division according to 'clusters' of items, albeit with a certain degree of arbitrariness. The idea of the scale stands not on the presence of the clusters, but rather on a conceptual analysis of the items within each cluster. For the current analysis we shall only inspect a few of the items on the boundaries of the divisions as illustrative examples (Table 5).

Table 5. Descriptors in a Hierarchy of Statistical Reasoning

Level 6. Graph interpretation, comprehensive interpretation and understanding of the relationships between three variables (OXY1_4)

Level 5. Graph interpretation, integrating two components, real-world application (BW2.1_2)

Level 4. Evaluating a statement with reference to data involving three factors (WF2.3_2); reasoning based on more than one factor (B3.1_2/B3.2_2)

Level 3. Graph reading, eight observations (RF2.2_1); graph reading, eight observations & categorisation – incomplete or minor flaws (RF2.2_1)

Level 2.Graph interpretation, single observation with two factors (WF2_1); graph interpretation, single observation (RF3_1)

Level 1. Straightforward graph reading, no interpretation required (RF1_1)

Conceptually, the boundaries seem to delineate levels of understanding of the data. It must be stated that the boundaries were drawn by inspecting the distribution of item difficulties for clusters separated by 'notches', and that this was done before any analysis of the items. To further support this idea of the conceptual boundaries, readers are invited to examine the scoring scheme from 'Oxygen', and to locate different student performances in the framework.

The above performance descriptors are consistent with the Watson and Callingham (2003) framework, and with that of Biggs and Collis (1982). More work needs to be done to characterise the nature of the performance on every task, to provide some appropriate labels for the levels within our framework, to distil descriptors of the reasoning skills involved, and to explore links with the Watson and Callingham hierarchy. However some tentative observations can be made from our initial analysis.

DISCUSSION

Our computer tasks required students to work with multivariate data, where the paper-based tasks did not. We contend this makes the computer tasks inherently more complex cognitively, but the analysis showed them to be hardly more difficult than paper tasks. We conclude that students aged 12-14 years in the normal ability range can reason with multivariate data if they have the appropriate tools and support for visualization.

All the tasks – both paper-based and computer-based - fit well onto a single Rasch scale, which would be the case if reasoning with multivariate data is an integral part of statistical literacy. Given the small and potentially biased student sample, and the small set of tasks considered, we take this as a working hypothesis, to be explored in detail in further studies. We are encouraged by the finding (not reported here) that the model provided a good fit for students, as well as for tasks.

The finding that young students can solve problems involving multivariate data has a number of implications for assessment, curriculum, and the public presentation of data. Tasks which require reasoning with multivariate data can be used for assessment purposes across a wide range of student ages and abilities. In our view, this should be done as an integral part of high-stakes testing. This is appropriate for two distinct reasons. Reasoning with multivariate data is an important component of statistical literacy, and so should be assessed formally, to ensure appropriate coverage of the domain. Second, high-stakes examinations have a profound effect on the curriculum, and the inclusion of multivariate problems in formal assessments will have the direct effect on the curriculum of forcing the inclusion of reasoning with more complex data.

One can identify some possible barriers to such developments. One might be a resistance from teachers or students to the inclusion of materials that seem more difficult than those currently faced. We think this is unlikely; in our discussions with teacher groups, and in working with students, there is an enthusiasm for more realistic tasks, and very positive engagement with the problems we set. A second barrier might be the availability of the technical infrastructure to support national, computer-based testing. In some countries at least, there is a commitment to widespread adoption of computer-based assessment (Department for Education and Skills, 2005).

The assessment of multivariate reasoning does not require on-line testing. E-portfolios are well suited to recording the results of locally administered tests, and this model could be adopted in countries without national e-assessment capacity. More extensive use of e-portfolios could open up further possibilities. In particular, the assessment of statistical literacy need not be done as part of mathematics. Portfolio assessment would allow evidence of reasoning with multivariate data to be gathered from a wide range of curriculum areas.

There are also a number of implications for curriculum planning. The finding that young students can work effectively with multivariate data opens up some rich opportunities for work in a variety of curriculum subjects. In our study, tasks that used a wide variety of contexts (biology, physics, citizenship) could be located on a common scale. This suggests that a coherent approach to cross curriculum planning for statistical literacy could have cross-curricular and extra-curricular benefits. It will take some time before we can gather evidence as to whether students better understand critical issues facing them and their world if they are better at reasoning from multivariate evidence, but we conjecture that this will be the case (Ridgway, McCusker and Nicholson, 2006)

Finally, this study has implications for interface design in general. We have clear evidence that young students can reason from multivariate data, and there is an opportunity for presenting evidence relevant to social policy (e.g. crime, health, road traffic accidents, global economics) in ways which are accessible to the majority of citizens, if appropriately designed interfaces are used.

A number of countries have initiatives from central statistics agencies to make government data accessible in a useable form, and to encourage the use of their data by schools (e.g. Canada, Portugal, and the UK). For example, Grunewald and Mittag (2006) refer to data from Zigure (2005) that pupils were responsible for about 25% of the hits on the UK Office for National Statistics website. In the European Union, Eurostat (http://epp.eurostat.ec.europa.eu/portal/) has been developing interactive displays (e.g. Mittag & Marty, 2005) to complement their on-line publications. These are a great advance on static text (especially over extensive tabular data), but are restricted to simple two variable displays. although it seems likely that they would benefit greatly from multidimensional displays. The OECD also have improved the readability of their Factbook (compare 2000 and 2006), and are actively looking for ways to present data better via ICT, see Ridgway, Nicholson and McCusker (2007d). These developments have the potential to both increase the accessibility of the data, and to help develop the statistical literacy of citizens – le savoir liberateur – as Condorcet advocated in 1792.

REFERENCES

Aoyama, K. (2007). Investigating a hierarchy of students' interpretation of graphs. *International Electronic Journal of Mathematics Education*, this issue.

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.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the quality of learning: The SOLO Taxonomy*. New York Academic Press.

Bond, T., & Fox, C. (2001) *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah NJ: Erlbaum.

Callingham, R., & Watson, J. (2005). Measuring statistical literacy. *Journal of Applied Measurement*, 6(1), 1-29.

Chance, B.L. (2002). Components of statistical thinking and implications for instruction and assessment. *Journal of Statistics Education*, 10(3). Online http://www.amstat.org/publications/jse.

Condorcet, J. (1994). *Foundations of social choice and political theory*. Aldershot and Brookfield, VT: Elgar (original work published in 1792).

Cremin, L. (1988). *American education: The metropolitan experience 1876-1980*. New York: Harper and Row.

Department for Education and Skills (2005). Harnessing technology transforming learning and children's services. Online: http://www.becta.org.uk/corporate/publications/.

Dewey, J. (1999). Democracy and education: an introduction to the philosophy of education. New York : Free Press.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatt, J., & Kruger, L (1989). *The Empire of Chance*. New York: Cambridge University Press.

Grunewald, W., & Mittag, H. J. (2006). The use of advanced visualisation tools for communicating European data on earnings to the citizen. In A. Rossman, & B. Chance (Eds.), *Proceedings of the Seventh International Conference on Teaching Statistics, Salvador, Brazil*. International Statistical Institute and International Association for Statistical Education. On line: http://www.stat.auckland.ac.nz/~iase.

Mittag, H. J., & Marty, U. (2005). *Gross earnings in Europe 2002*. CDROM. Luxembourg: Eurostat.

Monteiro, C., & Ainley, J. (2007). Investigating the interpretion of media graphs among student teachers. *International Electronic Journal of Mathematics Education*, this issue.

Orrill, R. (2001). Preface. In L. Steen (Ed.), *Mathematics and democracy; the case for qualitative literacy* (pp. xiii-xx) Princeton: The National Council on Education and the Disciplines.

Ridgway, J. (1997). Why measure development? In L. Smith, J. Dockrell, & P. Tomlinson (Eds.). *Piaget, Vygotsky, and beyond* (pp. 201-210). Routledge: London.

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

Ridgway, J., Nicholson, J. R., & McCusker, S. (2007d). Engaging citizens in reasoning with evidence. Paper presented at the *56th Session of the International Statistical Institute*, Lisbon, Portugal.

Ridgway, J., Nicholson, J. R., & McCusker, S. (2007a). Teaching statistics – despite its applications. *Teaching Statistics*, 29(2), 44-48.

Ridgway, J., Nicholson, J. R., & McCusker, S. (2007b). Using multivariate data as a focus for multiple curriculum perspectives at secondary level. Paper presented at the *56th Session of the International Statistical Institute*, Lisbon, Portugal.

Ridgway, J., Nicholson, J. R., & McCusker, S. (2007c). Embedding statistical assessment within cross-curricular materials. Paper presented to the *IASE Satellite Conference on Assessing Student Learning in Statistics*, Guimarães, Portugal.

Ridgway, J., McCusker, S., & Nicholson, J. (2006). Reasoning with evidence – Development of a scale. Paper presented at the *32nd Annual Conference of the International Association for Educational Assessment - Assessment in an Era of Rapid Change*, Singapore, May 2006. Online: http://www.iaea2006.seab.gov.sg/conference/programme.html.

Silva, P. (2006). Statistical education for doing statistics professionally: some challenges and the road ahead. Keynote presentation at the *Seventh International Conference on Teaching Statistics*. Salvador (Bahia), Brazil.

Schield, M. (2006). Statistical literacy survey analysis: Reading graphs and tables of rates and percentages. In A. Rossman, & B. Chance (Eds.), Proceedings of the *Seventh International Conference on Teaching Statistics, Salvador, Brazil*: International Statistical Institute and International Association for Statistical Education. Online: http://www.stat.auckland.ac.nz/~iase.

Watson, J. M. & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, 2(2), 3–46.

Zigure, A. (2005). Focusing on non-professional users. Proceedings of the *91st Conference of the Directors-General of the National Statistics Offices* (pp. 93-105). Luxembourg: Eurostat.

APPENDIX 1. COUNTRIES WEBSITES FOR USE BY SCHOOLS

Canada: http://www.statcan.ca/start.html

Portugal: http://alea-estp.ine.pt/ingles/index.html

England: http://www.stats4schools.gov.uk/default.asp

APPENDIX 2. REASONING FROM EVIDENCE TASKS

1. COMPUTER TASKS

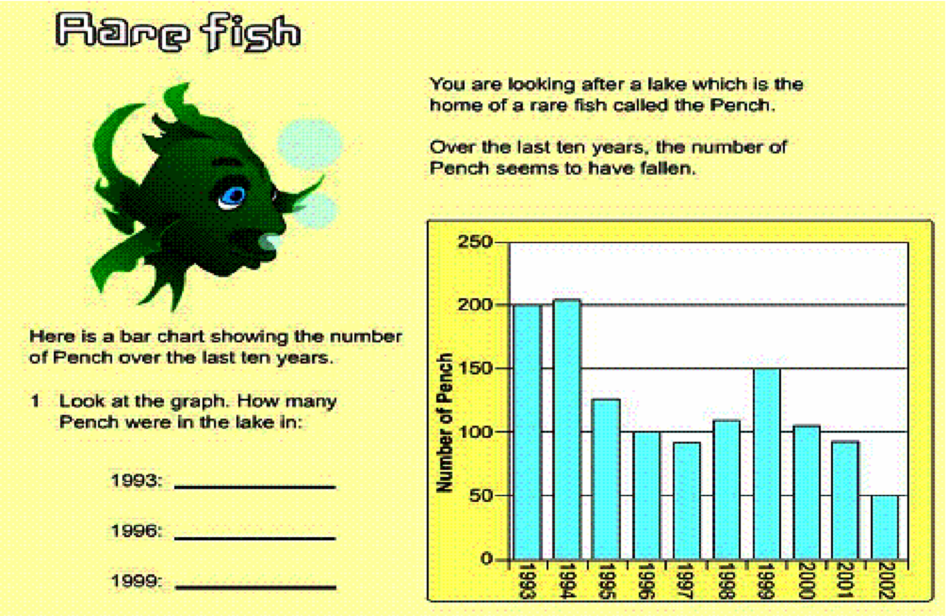
Task 1. Waterfleas.

Look at Page 1 until you understand what is happening.

1. Say how Craig's bar charts show that he is wrong.
2. Say who you think has the best idea. Explain how Craig's bar charts show who is right or wrong.

Task 2. Rare Fish

1. Complete this part.



Rare fish

You are looking after a lake which is the home of a rare fish called the Pench.

Over the last ten years, the number of Pench seems to have fallen.

Here is a bar chart showing the number of Pench over the last ten years.

1 Look at the graph. How many Pench were in the lake in:

1993: ___________

1996: ___________

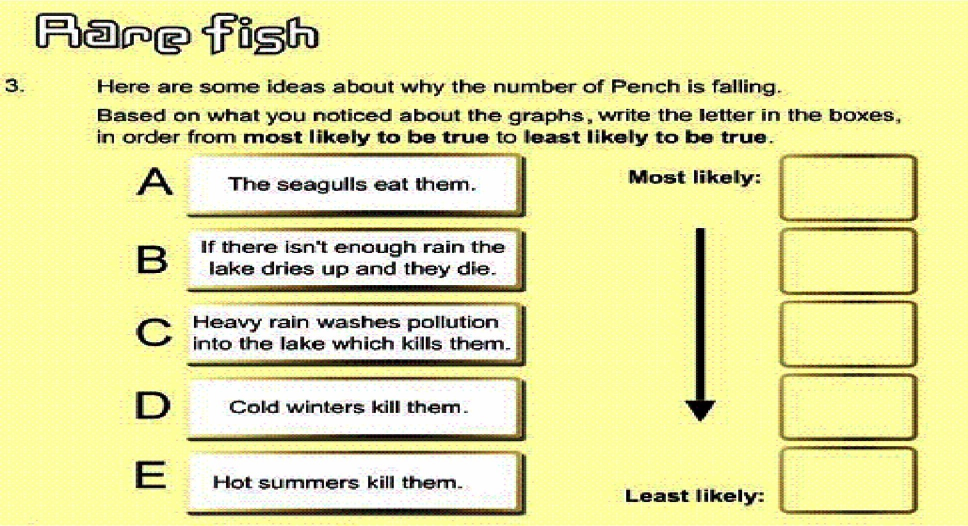
1999: ___________

2. Write down what you notice about the graphs. For each graph, say:

In which year it was highest.

In which year it was lowest.

For each of those years, was the number of pench high, low or in between?



Rare fish

3. Here are some ideas about why the number of Pench is falling.
Based on what you noticed about the graphs, write the letter in the boxes, in order from **most likely to be true** to **least likely to be true**.

A The seagulls eat them.

B If there isn't enough rain the lake dries up and they die.

C Heavy rain washes pollution into the lake which kills them.

D Cold winters kill them.

E Hot summers kill them.

Most likely: □

□

□

□

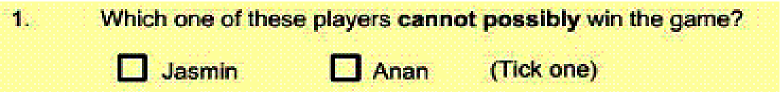
Least likely: □

How did you choose the **most** likely reason? Say which years' data helped you decide this.

How did you choose the **least** likely reason? Say which years' data helped you decide this.

Task 6. Bingo

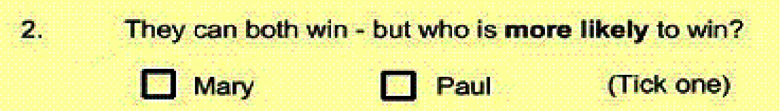
1. Press "Go" and watch the game.



1. Which one of these players **cannot possibly** win the game?

□ Jasmin □ Anan (Tick one)

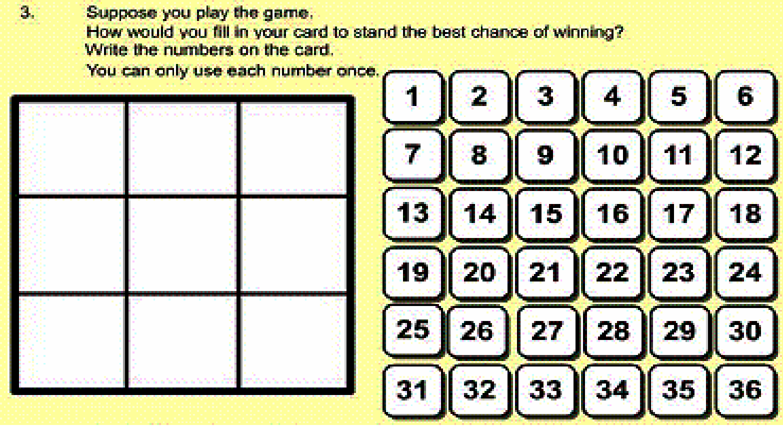
How do you know this?



2. They can both win - but who is **more likely** to win?

□ Mary □ Paul (Tick one)

How do you know this?



3. Suppose you play the game.
How would you fill in your card to stand the best chance of winning?
Write the numbers on the card.
You can only use each number once.

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36

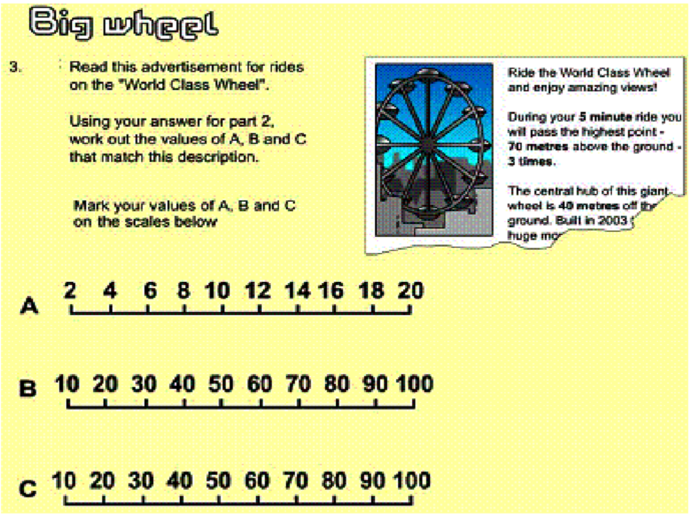
3. Explain carefully why you chose those numbers. How you decided where to put them on the card?

Task 7. Oxygen

Plants produce the oxygen we breathe. They use light, water and carbon dioxide to produce food and oxygen. You are doing and experiment to investigate how the rate of oxygen production by some plants is related to the temperature and the amount of light. Your apparatus lets you choose a temperature between 10 and 40oC. The lights can be adjusted between 0 (off) and 50 (fully on). You measure the rate of oxygen production for a range of temperatures and light intensities.

1. On the computer, use the graph tool to explore how oxygen production depends on light and temperature. Write your conclusions here
2. Jim and Ann run their own experiments to determine how oxygen production depends on temperature. Ann found little or no effect of temperature on oxygen production. Yet Jim found that temperature had a large effect. Could Both Ann and Jim be right? If not, what mistake might they have made?
3. Jim and Ann discuss how to make the plants produce the most oxygen. Jim suggests that they keep the plants as warm as possible. Ann suggests that they keep the lights on full power. Who do you think has the better idea? What conditions would produce more oxygen? Explain your reasons.

Task 12. Big Wheel



Big wheel

3. Read this advertisement for rides on the "World Class Wheel".

Using your answer for part 2, work out the values of A, B and C that match this description.

Mark your values of A, B and C on the scales below

Ride the World Class Wheel and enjoy amazing views!

During your **5 minute** ride you will pass the highest point - **70 metres** above the ground - **3 times**.

The central hub of this giant wheel is **40 metres** off the ground. Built in 2003 huge mo

A 2 4 6 8 10 12 14 16 18 20
B 10 20 30 40 50 60 70 80 90 100
C 10 20 30 40 50 60 70 80 90 100

Describe carefully what A, B and C tell you about the wheel

2. PAPER TASKS

Task 3. Smoking

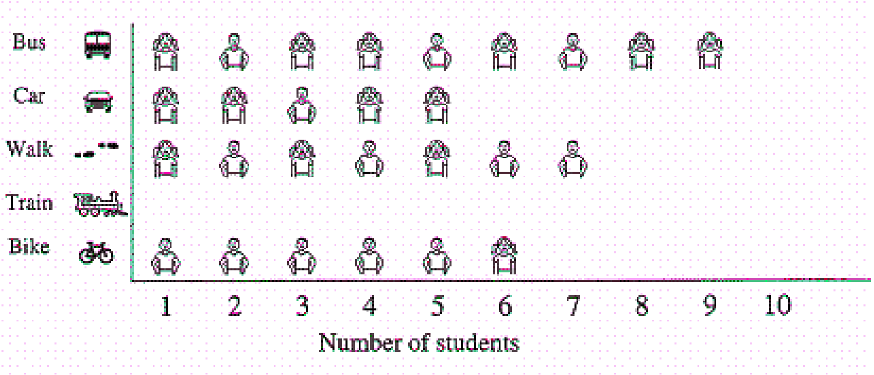
The following information is from a survey about smoking and lung disease among 250 people.

	Lung disease	No lung disease	Total
Smoking	90	60	150
No smoking	60	40	100
Total	150	100	250

Using this information, do you think that for this sample of people, lung disease depended on smoking? Explain your answer.

Task 4. School

How children came to school one day



1. How many children walk to school?
2. How many more children come by bus than by car?
3. Would the graph look the same everyday? Why or why not?
4. A new student came to school by car. Is the new student a boy or a girl? How do you know?
5. What does the row with the Train tell about how the children get to school?
6. Tom is not at school today. How do you think he will get to school tomorrow? Why?

Task 5. House Prices

Hobart defies homes trend

AGAINST a national trend, Hobart's median house price rose to $88,200 in the March quarter - but, Australia-wide, the average wage-earner finally can afford to buy the average home after almost two years of mortgage pain.

1. This article appeared in a newspaper. What does "average" mean in this article?
2. What does "median" mean in this article?
3. Why would the median have been used?

Task 8. Handguns

ABOUT six in 10 United States high school students say they could get a handgun if they wanted one, a third of them within an hour, a survey shows. The poll of 2508 junior and senior high school students in Chicago also found 15 per cent had actually carried a handgun within the past 30 days, with 4 per cent taking one to school.

Would you make any criticisms of the claims in this article? If you were a high school teacher, would this report make you refuse a job offer somewhere else in the United States, say Colorado or Arizona? Why or why not?

Task 9. Killer Cars

Family car is killing us, says Tasmanian researcher

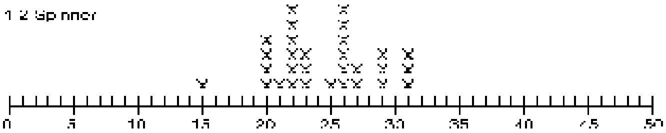
Twenty years of research has convinced Mr Robinson that motoring is a health hazard. Mr Robinson has graphs which show quite dramatically an almost perfect relationship between the increase in heart deaths and the increase in use of motor vehicles. Similar relationships are shown to exist between lung cancer, leukaemia, stroke and diabetes.

1. Draw and label a sketch of what one of Mr. Robinson's graphs might look like.

2. What questions would you ask about his research?

Task 10. Toss Up

A class did 50 spins of a fair coin many times and the results for the number of times it landed on heads are recorded below.



1. What is the lowest value?

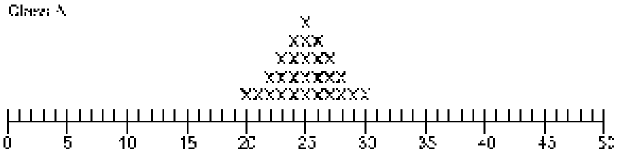
2. What is the highest value?

3. What is the range?

4. What is the mode?

5. How would you describe the shape of the graph?

6. Imagine that three other classes produced graphs for a coin. In some cases, the results were just made up without actually doing the experiment.

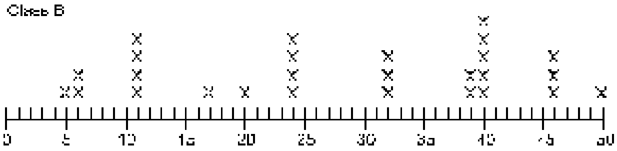


Do you think class A's results are made up or really from the experiment?

__Made up __Real from experiment

Explain why you think this.

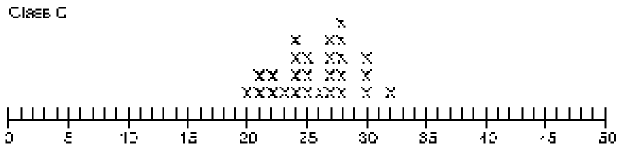
7. Do you think class B's results are made up or really from the experiment?



__Made up __Real from experiment

Explain why you think this.

8. Do you think class C's results are made up or really from the experiment?

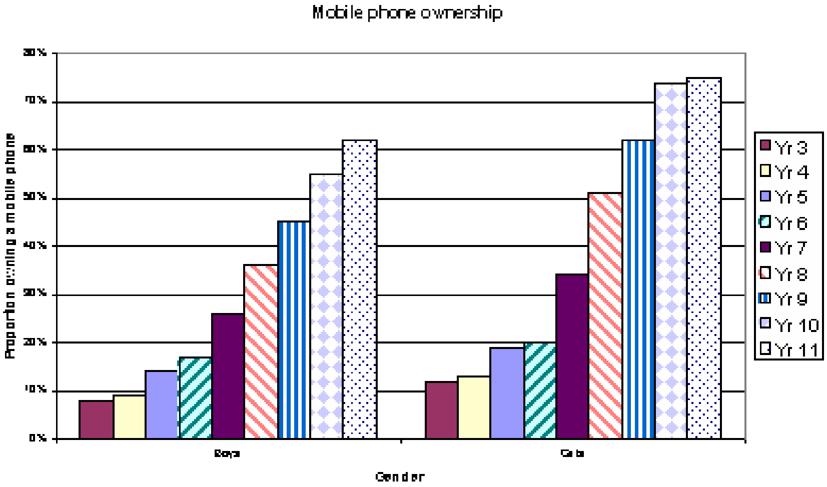


__Made up __Real from experiment

Explain why you think this.

Task 11. Mobile

A large scale survey gathered information from school children in Wales. The graph below shows the proportions of boys and girls owning a mobile phone across different year groups.



Based on this data only: What does the graph tell you about mobile phone ownership by school children in Wales? [please give reasons for your answers]:

Author : **Jim Ridgway, James Nicholson, and Sean McCusker**

E-mail : Jim.Ridgway@durham.ac.uk, j.r.nicholson@durham.ac.uk, Sean.McCusker@durham.ac.uk

Address : School of Education, University of Durham