Reasoning With Evidence – Development of a Scale

Professor Jim Ridgway, Dr. Sean McCusker & James Nicholson
University of Durham, UK

Email jim.ridgway@dur.ac.uk; sean.mccusker@dur.ac.uk, j.r.nicholson@dur.ac.uk

Introduction

Much of the data presented by politicians and the media is multivariate in its nature. However, in the UK at least, the general public has little training to deal with such information. It is reasonable to explore the school curriculum to determine the nature and extent of students' preparation for dealing with multivariate data. In the UK, high-stakes examinations have a profound effect on the curriculum; it follows that one can best understand the curriculum by examining the content of high-stakes tests. We analysed all of the specimen assessment materials produced for the year 2004 by all the examination bodies in the UK for statistics courses taken as part of mathematics, by students aged 17 and 18 years. Not one question required students to work with 3 or more variables. Furthermore, no examples were found where the relationship between two variables was anything other than linear. (Ridgway, Nicholson & McCusker, 2006, in press). The result of this disjuncture is that we are not equipping citizens to be able to take part in social debates or make important decisions concerning life choices or well-being.

It may be the case that reasoning with multivariate data is actually very difficult; certainly, students struggle to master relatively simple statistical concepts (Batanero, Godino, Vallecillos, Green, and Holmes, 1994). Here, we present evidence that, when supported with appropriate technology, students from even a young age are well able to work with multivariate data, and to draw informed conclusions from complex data. This has profound implications for the curriculum, and for assessment. It suggests that education may well be able to improve the quality of political debate, and the quality of the life choices made by individuals.
Our research poses as many questions as it answers, and there is an urgent need to understand in detail the nature of reasoning from data, and to map its stages of development.

Research Design

A study was carried out with 102 students from a selective school in Northern Ireland and 92 students from a non-selective comprehensive school in the North-East of England. Students were aged between 12 and 15 years. A test of Reasoning with Data was constructed using paper-based and computer-based tasks across a broad range of difficulties. Paper based items comprised of a number of tasks from the Watson & Callingham (2003) studies of statistical literacy together with a new paper based task involving reasoning with multivariate data. The computer-based tasks were selected from a range of items from the World Class Arena (http://www.worldclassarena.org, Richardson, Baird, Ridgway, Ripley, Shorrocks-Taylor, & Swan, 2002).). which specifically focussed on reasoning from data. These items were designed originally to assess the problem solving skills of very able students. Test were administered by the students' usual teachers in the presence of one of the authors and lasted approximately 70 minutes.

Briefly, the five computer tasks were:

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

Oxygen: 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 explain their reasoning throughout.

Bingo: 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 then asked to create their own 'best' card.

Big Wheel: Students are required to interpret the data form 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 .

The six paper-based tasks were:

School: 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: 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: Students are required to interpret multivariate data concerning mobile phone ownership.

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

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

Killer Cars: 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

Watson and Callingham's paper-based items were marked according to the partial credit model used in their previous analyses. In this model, responses are categorized according to their level of sophistication and are assigned a numerical label. Responses to the WCA items were categorised according to a scheme adapted from the scoring rubrics associated with the tasks. In some cases this was a direct interpretation of the rubric, in others the rubric was generalised to accommodate a range of answers of similar structure. In cases where marks were awarded for increasingly sophisticated answers, analysis was carried out on the aggregated score using a partial credit model, in other cases where indicators were assigned for discrete performances, results were left in raw form (1 or 0 depending on achievement).

Students were encouraged to complete as many tasks as they were comfortable with, but emphasis was placed on 'doing one's best' on each item as opposed to trying to do as many items as possible. Blank responses were treated as zero if they lay within a body of other attempts and treated as blank if it appeared that the item was not attempted because it was not reached. Tests were marked by one of the authors (Nicholson) who has extensive experience of marking student work at this level for high stakes tests.

The resulting data were analysed using partial credit Rasch scaling. This technique analyses responses and places them in order of difficulty. The analysis also produces a measure of 'fit' which is an indicator of the extent to which any item fits within the single scale described by the whole test.

Results

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

Reading the Table

The scale represented in the middle of Table 1 is the Rasch scale of difficulty. Items towards the top of Table 1 with positive scale values are the more difficult items within the test. The Table 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 a 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					

Table 1 – 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 ask

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

Comparing the Relative Difficulty of Computer and Paper-based Tasks

Examination of Table 1 shows that the computer based items cover much of the same spread of difficulties as the paper-based items. The few paper based items at the bottom of the difficulty table 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 in the WCA, because their original function was to identify high attaining students. Whilst the Computer-based items appear slightly more difficult than the paper-based items overall, there is no dramatic difference in difficulty between the paper-based task and the Computer-based tasks, which require multivariate reasoning.

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 stressed that just because an item does not meet the fit criteria (usually .77 < fit statistic <1.3) it must be disposed of. 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 reject or modification of the item or scoring scheme. Where Table 1 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 1 shows the infit statistic for all the item parts to be adequately within the conventional bounds 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 2, 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 but it can be seen how 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.

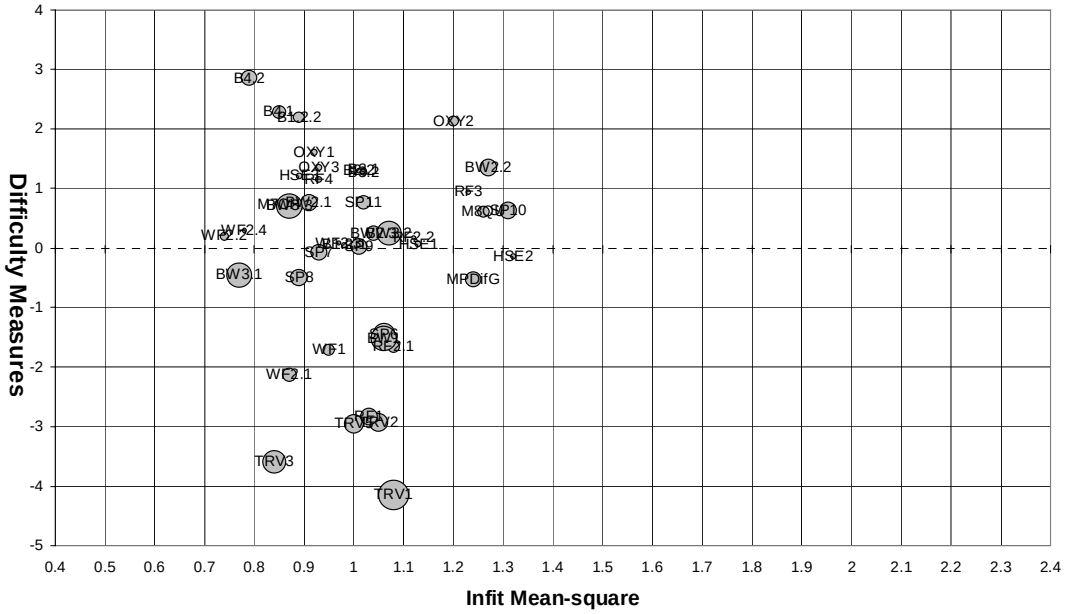

Figure 1 – Infit statistic from Rasch scaling of test items

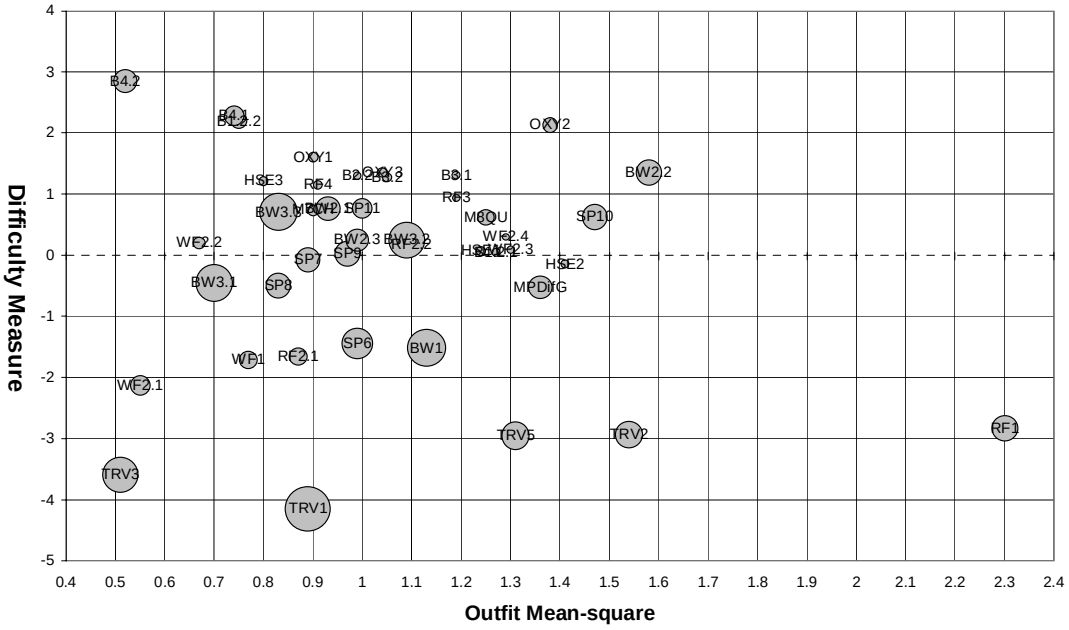

Figure 2 - Figure 1 – Outfit statistic from Rasch scaling of test items

Towards a Hierarchy of Reasoning from Evidence

Here, we begin with a detailed description of one computer-based task, then provide a tentative description of the hierarchy of competencies that underlies performance on the whole task set.

Whilst all the Computer-based tasks contain a good spread of items over the difficulty range, Oxygen is particularly impressive. A detailed analysis of the Oxygen task will follow. Oxygen has the best psychometric properties of all the tasks within the test, moreover it is generally representative of all the computer based-items. Figures 3 to 6 show screen dumps from the task. Page1 (Figure 3) merely outlines some background information and the context of the task; the student is prompted to read the information and move on to page 2 to begin the task. On page 2 (Figure 4) the student is presented with an interface which is able to present multivariate data in an accessible form. The student can choose which variable to place on the x-axis and which variable to place on the slider. The layout shown in Figure 4 allows the student to observe how the rate of production of Oxygen varies with light intensity. The student can slide the temperature values and see how the graph changes. The student is then asked an open question about the data. The mark scheme for the responses is shown in Table 2. In part 1 (OXY1) marks are awarded for single descriptors; however the marks increase as more information is constructed, and to achieve full marks (4) they are required to have fully grasped the idea of modifier effects (these marks are treated as codes in Rasch scaling). This first question on its own covers a wide difficulty range (shown in bold in Table 1) and the large separation between each level suggests that students who offer a more detailed description of the patterns in the data may well be functioning at a cognitively more complex level. Part 2 (OXY2), on page 3 (Figure 5) requires the students to evaluate the statements made by 'Ann' and 'Jim' and reconcile the apparent difference in results. One mark is awarded for identifying the source of the error. However, to gain both available marks for the item, the student must be able to recreate the circumstances under which 'Jim' and 'Ann' were working. Part 3 (OXY3) on page 4 (Figure 6) requires the students to evaluate two suggestions, then propose their own idea for increasing oxygen production.

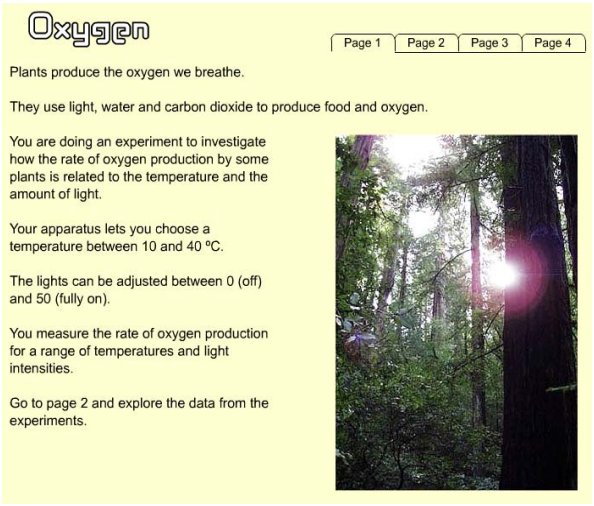

Figure 3 – Page 1 of Oxygen Task

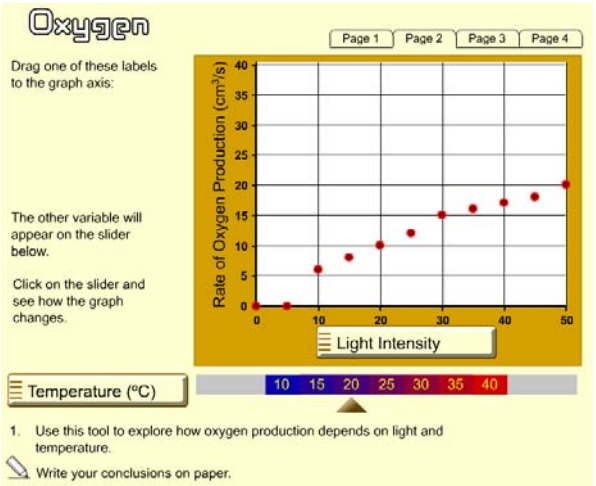

Figure 4 – Page 2 of Oxygen Task

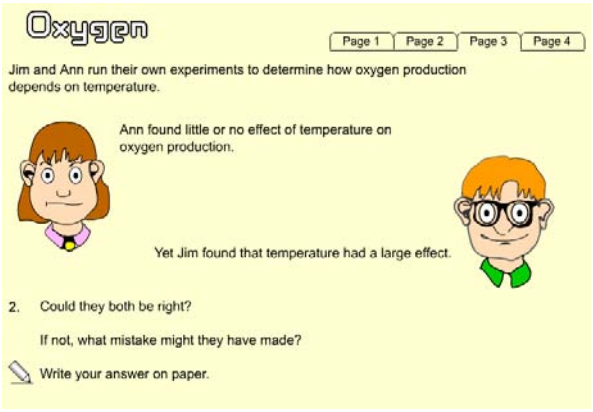

Figure 5 – Page 3 of Oxygen Task

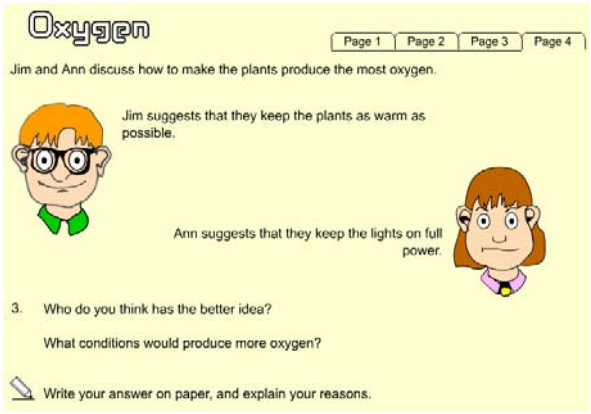

Figure 6 - Page 4 of Oxygen Task

	OXYGEN: Age 13	Points	Section points
	The core elements of performance required by this task are: • Explore relations in 3 variables in a scientific context. • Identify optimal values Based on these, credit for specific aspects of performance should be assigned as follows:		
Q1	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 is produced' then give (1). • 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 then give (1)	 1 1 1 1	4
Q2	They could both be right. (Ignore reason for mark) *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 (1)	2
Q3	Ann 's idea is better than Jim's. (Ignore reason for mark) 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

Table 2 – Oxygen Task - Scoring Rubric

The fact that the tasks presented here can be fitted to a single scale does necessarily mean that such a scale has any psychological reality or that it is being measured. 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 3).

RF1_1	*Straightforward graph reading. No interpretation required*	*Level 1*
WF2_1	*Graph interpretation, single observation with 2 factors*	*Level 2*
RF3_1	*Graph interpretation, single observation*	
RF2.2_1	*Graph reading, 8 observations*	*Level 3*
RF2.2_2	*Graph reading, 8 observations & categorisation – incomplete or minor flaws*	
WF2.3_2	*Evaluate statement with reference to data. 3 factors*	*Level 4*
B3.1_2/B3.2_2	*Reasoning based on more than one factor*	
BW2.1_2	*Graph interpretation, 2-component, real-world application.*	*Level 5*
OXY1_4	*Graph interpretation, comprehensive interpretation & understanding*	

Table 3: Descriptors in a Hierarchy of Statistical Reasoning

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, and to provide some appropriate labels for the levels within our framework,

Conclusion

Most of the situations we encounter in everyday life are multivariate, but we prepare our young people very poorly for this. We have developed interfaces which allow students to engage fully with multivariate data. This engagement, which is

cognitively more complex than that encountered by even Advanced Level Statistics students, has been found to be within the capabilities of children as young as 12 years. There is evidence which supports the existence of a hierarchy of data handling skills, from working with single values, one step computation, and elementary reasoning, through to fluency using a variety of representations, fluency with number, and in synthesising evidence and communicating results clearly. We need to capitalise on this evidence in the development of curriculum and assessment materials. We also need to understand better the structure and development of the skills required to reason from evidence. This understanding will support better curriculum planning, and also cross-curricular co-ordination to enable data intensive topics to be presented in relevant contexts such as Geography, PSHE or Citizenship.

References

Batanero, C., Godino, J.D., Vallecillos, A., Green, D., and 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

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

Ridgway, J., Nicholson, J.R., and McCusker, S. (2006, in press) Teaching Statistics – Despite its Applications. *Teaching Statistics*.

Watson, J. M. & Callingham, R. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal,* 2(2), 3–46. http://www.stat.auckland.ac.nz/~iase/serj/SERJ2(2)_Watson_Callingham.pdf