REASONING WITH EVIDENCE – NEW OPPORTUNITIES IN ASSESSMENT

Jim Ridgway, James Nicholson, and Sean McCusker
University of Durham, United Kingdom
Jim.Ridgway@durham.ac.uk

Computers facilitate reasoning with complex data. 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. All the tasks fitted well onto a single Rasch scale; computer tasks were cognitively more complex, but ranked only slightly more difficult than paper tasks on the Rasch scale. Implications for assessment, the curriculum, and public presentations of data are discussed.

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 we are to use evidence to inform and understand public debate, to function in complex, fast-moving commercial environments, and to make important personal decisions, we need high levels of statistical literacy.

Statistical literacy encompasses a number of ideas. Wallman (1993) emphasised the ability to value, understand and evaluate statistical evidence that influences our daily lives; Gal (2002) emphasised the ability to interpret, evaluate and communicate statistical evidence. There is increasing sophistication in the ways that data are presented and analysed in the media when addressing complex problems facing society. Haggett (2000), for example, illustrates new ways of mapping disease spread, and the importance of recognizing significant regional variability in the behaviour of epidemics. In the USA, the Centres for Disease Control and Prevention (www.cdc.gov) are using mapped disease and immunisation data as part of their public information initiative. Comprehending and communicating with multivariate numerical data is increasingly important, and appropriate skills need to be developed that can be applied in a range of meaningful contexts. The need for citizens to reason with multivariate data goes way beyond earlier conceptions of statistical literacy, and is the focus of this paper.

The UK National Curriculum presents very few opportunities for students to engage with multivariate data. Our recent analysis of all statistics components in high-stakes mathematics examinations for students aged 17 and 18 years showed that students never encountered more than 2 variables at a time, and when they did work with two variables, relationships were always linear. About 75% of available marks were awarded for demonstrating statistical technique. We concluded that students are being taught a limited range of techniques that narrow their conceptions of reasoning from evidence, and are offered no preparation for reasoning with multivariate data (Ridgway, Nicholson, and McCusker, in press).

New technology offers opportunities for change. The World Class Arena (WCA) project (http://www.worldclassarena.org/) 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 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 and McCusker, 2003).

The WCA project provided some evidence of what high-attaining students can do, and raises some important questions for educational practice in general. In particular:

- To what extent can students from a broad ability range reason with multivariate data presented via Information and Communication Technology (ICT)?
- How does reasoning from multivariate data develop?
- How does reasoning from multivariate data relate to other components of statistical literacy?

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 believe 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. We believe that the curriculum (say in geography, citizenship, science and psychology) could be made more realistic and more relevant to students; students could leave school with a far greater understanding of ways to work with multivariate data. For this to happen, we need a good understanding of what develops and how, in order to plan curriculum progression and coordination across subjects, and to have reasonable expectations of students. Moreover, we need to demonstrate that assessment is effective and manageable within current curriculum constraints.

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 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. They describe a hierarchy of statistical literacy evident in the data, which they propose 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 and 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 with very few tasks. These would occur if the notches 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, but 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

One hundred three students from an academically selective school in Northern Ireland and 92 students 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, across a broad range of difficulty levels, a new paper based task that required reasoning with multivariate data, and computer based tasks from the WCA work. The computer based tasks were chosen to be representative of different types of tasks, to see what levels of difficulty they appeared to represent. Surveys 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 three of the five computer based tasks is given below (sample tasks can be found at http://www.worldclassarena.org/), followed by a brief description of three of the seven paper based tasks.

- *Computer tasks*:

Waterfleas: presents graphical data on the activity of waterfleas over time in water of different temperature, containing different amounts of pollutant. Students have to judge the correctness or otherwise of claims about the data, and to justify their answers.

Rare Fish: presents graphical data on the population of a rare fish over time, with evidence on changes in temperature, rainfall and the numbers of seagulls. Students draw conclusions about the likely causes of the decline of the rare fish population.

Oxygen: provides graphical data on the amount of oxygen produced by plants under different conditions of light and temperature. Students evaluate statements made, resolve a paradox of experimental design, choose values for light and temperature that optimize oxygen production, and justify their decisions.

- *Paper tasks*:

School: presents a pictograph of the number of students who travel to school using different forms of transport. Students are asked to answer questions that involve ideas of variability, and uncertainty, and to justify their answers.

*Handguns***:** presents a description of a survey and the conclusions drawn. Students are required to critique claims made, and to extrapolate from the data, and to justify their conclusions.

Mobile: presents evidence on mobile phone ownership by boys and girls of different ages, and asks students to draw conclusions.

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. Some items will have several categories, others as few as 2. In the case of the WCA 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 were asked to work through the items in the order in the booklet, using the computer when required, but not to worry about trying to get all the tasks finished and to work as quickly as they felt comfortable with. 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 (Nicholson) who has extensive experience grading student work for high-stakes assessment.

These data were then analysed via partial credit Rasch scaling (Figure 1).

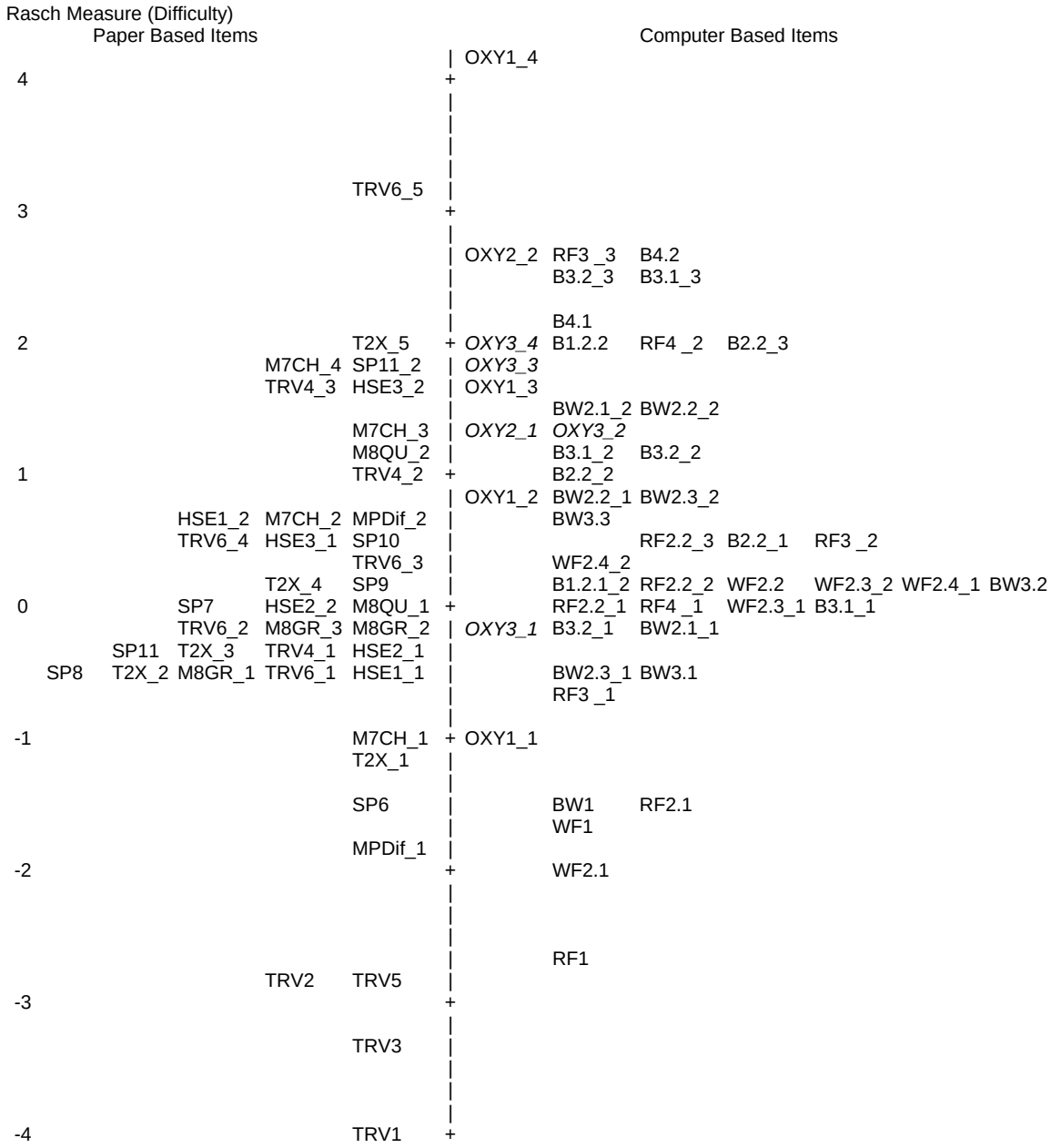

Figure 1: Multivariate Reasoning Scaled with Statistical Literacy

This procedure puts the categories of responses to each item into their difficulty order; here, the most difficult items appear at the top of Figure 1.

Figure 1 shows the location of responses to the paper based and computer based items against the same vertical Rasch scale. Each label represents a particular response to a particular item. We have edited the computer output to show all the responses to 'Oxygen' in a single column as far as possible (OXY3_2 is to be found at the same difficulty level as OXY2_1 in the second column), and have shown all the levels of response to a single item (OXY1) in bold. OXY1_4 refers to a perfect response to the first item in the Oxygen task, OXY1_3 refers to a good but partial response, and so on. In some of the tasks, there are two levels of numbering, so B3.1 and B3.2 are two separate aspects of the third part of the Bingo task, and each of these aspects was coded at three levels. It can be seen that the computer based tasks that require multivariate reasoning have a similar spread of difficulties as the paper based tasks. They appear slightly more difficult on average than the paper based tasks but are certainly not dramatically more difficult. Although only 5 WCA tasks were used in this study, we are planning to look at a number of other tasks in further studies to see if this initial finding is more generally applicable.

One can be more confident about the existence of a single scale if every item response and every student fits the model well. Figure 1 does not show any information about these measures of fit, but the software to do Rasch analysis does provide them. In this study, 5 out of 50 item responses (2 computer responses and 3 paper responses) and 13 out of 195 students failed to fit the scale adequately. Overall, these data represent a good fit to a single Rasch scale.

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 30^0 and decreases thereafter.'

It is interesting to note that items where differences were substantial in both of the 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.

DISCUSSION

A lot more work needs to be done with these new tasks 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.

The data show clearly that all the tasks fit well onto a single Rasch scale. We conclude, on the basis of this evidence, that reasoning with multivariate data is an integral part of statistical literacy. Given the small and potentially biased student sample, we take this as a working hypothesis, to be explored in detail in further studies. We are encouraged by the finding that the model provided a good fit for students, as well as for tasks.

Computer tasks required students to work with multivariate data, so inherently were cognitively more complex, but the analysis showed them to be hardly more difficult than paper tasks. We conclude that students can reason with multivariate data if they have the appropriate tools and support for visualization. This has a number of implications for assessment. 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 experienced curriculum, and what is not assessed formally will struggle to maintain its position.

It is possible to identify some potential 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. Two sorts of responses can be made to this challenge. First, in the UK, there is a commitment to widespread adoption of computer based assessment (Department for Education and Skills (D*f*ES), 2005). Second, the infrastructure for national computer-based testing will soon be in place. Every student's ability to use ICT will be assessed on-screen at age 13 years by 2008 (Qualifications and Curriculum Authority, 2005).

The assessment of multivariate reasoning does not require on-line testing. In the UK, every student will have an e-portfolio (D*f*ES, 2005). 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 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. Students appear to be able to engage effectively with cognitively more complex tasks using computers across a broad spectrum of difficulty of items. 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.

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.

REFERENCES

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

Department for Education and Skills. (2005). Harnessing technology: Transforming learning and children's services, http://www.successforall.gov.uk/downloads/harnessingtechnologye-strategy-151-225.pdf

Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, 70, 1-51.

Haggett, P. (2000). *The Geographical Structure of Epidemics*. Oxford: Clarendon Press.

Nicholson, J. R., Ridgway, J. and McCusker, S. (2006). Reasoning with data – time for a rethink? *Teaching Statistics*, 28(1), 2-9.

Qualifications and Curriculum Authority. (2005). QCA Testing, www.rm.com/qca/default.asp.

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

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

Watson, J. M. and 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.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the Statistical Association*, 88(421), 1-8.