ASSESSING STATISTICAL LITERACY: A QUESTION OF INTERPRETATION?

Rosemary Callingham
University of New England, Australia
Rosemary.Callingham@une.edu.au

Current school curriculum documents stress the need for assessment to support learning. Teachers use assessment information to infer students' development and plan appropriate intervention. In order to do this, a framework is needed within which the assessment can be developed and interpreted, and a suitable task is required to obtain the necessary information about students' performances. The responses of 586 students to performance assessment tasks developed for the purpose of assessing a numeracy construct, rather than statistical understanding, were analysed against a previously identified hierarchy of Statistical Literacy. The findings suggest that the tasks provided reliable and interpretable evidence of performance in Statistical Literacy, using a classroom-based process rather than a traditional test.

INTRODUCTION

The rise of interest in socially-based curriculum frameworks and a move towards applications-based approaches to teaching and learning have promoted interest in students doing more than simply carry out mathematical procedures. There are calls for students to think critically about social situations in which data are used, sometimes referred to as applying *statistical literacy.* Statistical Literacy has been defined as

> …the ability to understand and critically evaluate statistical results that permeate our daily lives – coupled with the ability to appreciate the contributions that statistical thinking can make in public and private, professional and personal decisions.
> (Wallman, 1993, p. 1)

Such a definition requires that students must develop not only the mathematical skills required to understand statistical information, but also an appreciation of the social context in which the information is set. Indeed Gal (2002) goes further suggesting that for full participation in our increasingly 'data-drenched' society (Steen, 1999), students must

> (a) … interpret and critically evaluate statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant (b) … to discuss or communicate their reactions to such statistical information, such as their understanding of the meaning of the information, their opinions about the implications of this information, or their concerns regarding the acceptability of given conclusions. (Gal, 2002, pp. 2-3)

These ideas have led to calls for an increased emphasis on statistics and probability in the mathematics curriculum (e.g., National Council on Education and the Disciplines (NCED), 2001). Holmes (1980) described categories of statistical understanding: data collection, data tabulation and representation, data reduction, probability, and interpretation and inference. Each one of these categories involves understanding a range of statistical concepts such as sampling, measures of central tendency and range, and experimental and theoretical probability, which can be assessed in traditional ways. The emphasis on contextual understanding and critical thinking, however, presents a challenge for assessment. Clearly, from the descriptions of Statistical Literacy provided by Wallman and Gal, a number of elements entwine to create a complex construct. Critical thinking implies that higher order processes are required (Krulik and Rudnick, 1999) to make sense of both the mathematical underpinnings and the social context at the same time.

There are two challenges that assessors must meet. First, a framework must be identified that will provide information about the development of cognitive skills, including higher order thinking, in statistical contexts. Such a framework is likely to be hierarchical or developmental in nature, in which later understanding and skills are based on prior knowledge and experience, leading to the development of complex cognitive processes. Many school curricula, however, are framed in terms of subject matter, which may or may not match student development of understanding, and outcomes are specified on this basis (e.g., Australian Education Council, 1994). Second, a suitable task is required to make this framework operational in the classroom. Each of these components must provide information that will allow reliable and valid inferences to be made about students' understanding regardless of the context of the assessment task.

Watson and Callingham (2003) and Callingham and Watson (2005) approached the issue of identifying a framework for assessing Statistical Literacy through the use of Rasch modelling (Rasch, 1960). Using archived data from surveys conducted over a number of years that addressed the aspects of statistical thinking suggested by Holmes (1980), Watson and Callingham demonstrated that a unidimensional scale of Statistical Literacy could be constructed that provided interpretable information about students' achievement. They described a six level hierarchy characterised by increasingly complex cognitive processes in which statistical processes and contextual understanding were both involved. This hierarchy is summarised in Table 1. A framework, therefore, existed that could provide information about Statistical Literacy development. This framework, however, had been identified using items administered under traditional test conditions, and the issue of providing alternative types of assessment remained.

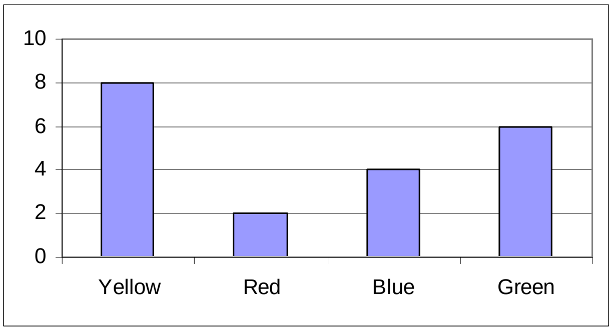
Table 1: Statistical Literacy construct (Watson and Callingham, 2003)

Level	Brief characterisation of levels
6 Critical-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.
5 Critical	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.
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.
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.
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.
1 Idiosyncratic	Idiosyncratic engagement with context, tautological use of terminology, and basic mathematical skills associated with one-to-one counting and reading cell values in tables.

Two existing tasks could provide a model for assessment design to be used in normal classroom contexts. The two tasks were part of a suite designed to make operational a generic Continuum of Competence (Callingham and Griffin, 2000) in diverse numeracy contexts. One task, called *In a Spin* (*IAS*), was set in the context of a game involving spinners, and addressed aspects of probability and statistical variation. Students made a spinner, using data provided in the form of a graph, and solved unusual problems involving two spinners with unequal outcomes. The second task, *Keep Australia Beautiful* (*KAB*), used the context of a school litter collection. Students were provided with a set of 12 cards that showed data collected from a school litter survey. They used the information on these cards to find connections, make inferences and draw conclusions about the data. The two tasks were undertaken by students in normal classroom settings, rather than under 'test' conditions. Thus, these two tasks addressed many of the aspects of Statistical Literacy described by Wallman (1993) and Gal (2002) but also provided an approach to assessment that was consistent with curriculum documents in that they were part of normal teaching activities.

Each task comprised a set of activities that was organised into several sub-tasks. Each activity had a scoring rubric which was designed to identify increasing quality of response to that activity. These varied from 0-1 to 0-4, depending on the complexity of the activity. Each one of these sub-tasks could be used on its own as a classroom activity, but together the several sub-tasks provided a scaffolded task that allowed students to demonstrate higher order thinking. Figure 1 shows the middle sub-task from *In a Spin*, together with the scoring rubrics. The three questions (6a, 6b and 6c) form a coherent task but are scored independently, so that a student could make an incorrect spinner but nevertheless demonstrate understanding of the effect of sample size and the need for no physical bias in the spinner. The scoring rubrics were based initially on anticipated responses, and modified in the light of what students actually did.

These are some results from spinning a 4 colour spinner 20 times.



	Activity	Rubric	Code
6a.	Make a spinner that would most likely give you these results.	Spinner does not fulfil criteria. Spinner shows four colours but not in correct proportions. Spinner shows 4 colours in approximate ratio Red:Blue:Green:Yellow = 1:2:3:4	0 1 2
6b.	Test your spinner and record your results in any way that you wish.	No recording or only summary information e.g., Red 2, Blue 4 etc. Recording incomplete and not systematic. Recording systematic and efficient e.g., tally marks, table.	0 1 2
6c.	Explain why your results are the same or different from the ones shown here.	No explanation or explanation restates findings e.g., Yellow came up most, then green, etc. Explanation based on a story or personal preference e.g., Red is luckier or I spun the spinner too hard. Explanation based on chance but without a sense of needing a lot of spins (sample size) e.g., It's just luck which colour comes up or on spinner bias e.g., The pencil might not have been exactly in the middle. Explanation based on sample size *and* assumption of no bias e.g., To get similar results I would need to spin the spinner lots more than 20 times and make sure that the spinner was made properly.	0 1 2 3

Figure 1: *In a Spin* sub-task and rubrics.

The tasks, *Keep Australia Beautiful* and *In a Spin*, had been used previously as part of a group of five tasks to link numeracy skills and understanding to an underlying generic continuum of cognitive competence. What needed to be established was whether these two tasks could be used together to provide information about Statistical Literacy.

METHOD

Rasch (1960) measurement provided a robust approach to identifying an underlying variable or construct. There are two reasons for using Rasch approaches. First, the model is paramount and misfit suggests that the items, questions or tasks are not working together consistently to define an interpretable construct. Hence, evaluation of the fit of the data from these two tasks to a Rasch model would provide information about the coherence of these two tasks. Second, the variable produced by the modelling process could be *back translated* (Griffin and Forwood, 1990; Callingham and Watson, 2005) on to the defined Statistical Literacy construct. Back translation is the process of comparing the continuous variable produced through Rasch modelling, with an underlying identified cognitive continuum or developmental sequence. Where the tasks, questions or items used are designed to make operational a specific sequence, this becomes a confirmatory technique that provides information about the validity of inferences that can be drawn about students' cognitive development. In this instance, however, the approach was exploratory to establish whether the *Keep Australia Beautiful* and *In a Spin* tasks could be used to measure the Statistical Literacy construct.

Using archived data, a sample of 586 Grade 10 students was identified who had completed both tasks. This sample (48.5% male, 50.8% female, 0.7% unidentified), were in 13 different schools in the state of Tasmania, Australia. The data were analyzed with the Quest computer program (Adams and Khoo, 1996) using the partial credit model (Masters, 1982).

RESULTS

Fit to the Model

Fit to the model, of both items and persons, was evaluated using the Infit Mean Square (IMSQ) statistic and the standardized infit (Infit t). The acceptable values lie between 0.77 and 1.3 (Keeves and Alagumalai, 1999) with an ideal value of 1.00. For both items ($IMSQ_I$ = 1.00, s.d. =0.1; Infit t = -0.7) and persons ($IMSQ_P$ = 0.98, s.d. =0.42; Infit t = -0.14) the overall fit was acceptable. Consideration of the fit of each item also showed that no item fell outside the limits of acceptable fit. Reliability indices were high (Item separation reliability = .97; Person separation reliability = .86; Cronbach alpha = .83) indicating that the internal consistency of the tasks was good. These findings indicated that all items worked together to measure a single underlying construct, and the persons who attempted the tasks performed in expected ways.

Interpretation of the Construct

Although the fit to the model suggested that the two tasks were measuring a single construct consistently, it remained to be seen whether the construct could be substantively interpreted in terms of the Statistical Literacy continuum. In order to establish this, an audit of the content knowledge and skills needed to address items appearing at different places along the construct was undertaken – the back translation process.

The Quest computer program produces a *variable map* showing the distribution of both items and persons along the measured construct (Adams and Khoo, 1996). This map is shown in Figure 2, with the two tasks separated. The items' names refer to the task, the number of the task activity and the partial credit rating, hence IAS6b.2 is *In a Spin* activity numbered 6b on the sheet given the partial credit rating of 2. The figures on the extreme left hand side are logits, the units of Rasch measurement, and persons are represented by X, with each X representing four persons.

Along the construct, there are gaps between items. For example, there is a gap at approximately two logits between IAS1.2 and KAB4B.4 and the two lower level items of IAS9.2 and KAB6B.3. These gaps indicate a jump in difficulty of the items, with greater cognitive demands required to achieve items at the higher level. These gaps provide an indication of suitable partitioning of the construct to give a set of descriptors of the likely behaviors of students at different points along the construct. It should be emphasized that the boundaries between the groupings of items are not hard edges, but rather provide a set of levels that give a convenient way of describing changes as students progress to higher levels of performance. The aim is to furnish teachers with a tool that can be used to assess students' Statistical Literacy without swamping them with detail.

The skills and understanding required by each group of items identified by looking for the jumps in difficulty was considered and compared with the levels of the Statistical Literacy construct previously identified (Watson and Callingham, 2003). The nature of the thinking required by students was mapped onto the Statistical Literacy behaviors described at each level. The final item groupings related to the levels of the Statistical Literacy continuum are shown by the horizontal lines on the map.

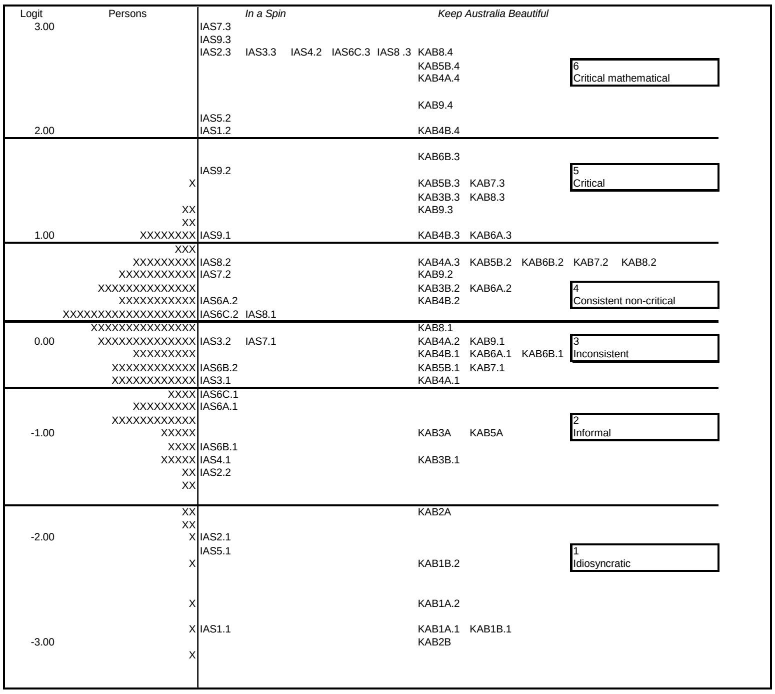


Figure 2: Variable map of IAS and KAB tasks

To illustrate this process, consider items IAS6C and KAB8. IAS6C is shown in Figure 1. KAB8 asked students to suggest some new questions that they could answer with the data given. Both items address different aspects of Statistical Literacy. IAS6C.1 appears at the Informal level. The rubric indicated that the explanation of difference was based on personal preference, which is coherent with the description of the Informal level. Similar analysis was undertaken for all items. Table 2 shows the mapping of the rubrics for the two indicator items onto the levels of the Statistical Literacy construct.

Table 2: IAS6C and KAB8 rubrics mapped onto Statistical Literacy levels

Item	Logit	Rubric	Statistical Literacy level
IAS6C.1	-0.38	Explanation based on a story or personal preference (e.g., Red is luckier or I spun the spinner too hard)	Informal
KAB8.1	0.03	Questions asked reproduce the questions in the task using different places or kinds of litter (e.g., How many hamburgers are sold in the week?)	Inconsistent
IAS6C.2	0.26	Explanation based on chance but without a sense of needing a lot of spins (sample size) (e.g., It's just luck which colour comes up) or on spinner bias (e.g., The pencil might not have been exactly in the middle.)	Consistent Non-critical
KAB8.2	0.68	Questions asked use the data but are closed (e.g., What is the average number of pieces of litter collected each day?)	
KAB8.3	1.31	Questions asked are open-ended and relate to data but do not go beyond those already asked (e.g., What is the most popular item sold by the canteen and why?)	Critical
KAB8.4	2.75	Questions asked are open-ended and use inferences about new and changed conditions (e.g., What would happen if the school did a litter survey each term?)	Critical Mathematical
IAS6C.3	2.77	Explanation based on sample size and assumption of no bias (e.g., To get similar results I would need to spin the spinner lots more than 20 times and make sure that the spinner was made properly.)	

The rubrics shown here appear to demonstrate the characteristics of the Statistical Literacy levels described in Table 1. The highest response rubrics, IAS6C.3 and KAB8.4, both demanded that students express sophisticated understanding, subtle use of language and complex cognitive skills, as is expected by the Critical Mathematical level of Statistical Literacy.

DISCUSSION

This reanalysis of data collected from the use of performance assessment tasks developed to realise a generic continuum of competence provided evidence of the applicability of the Statistical Literacy construct to alternative forms of assessment. There are two implications from this finding. First, the construct itself is not dependent on a specific set of items to measure it, implying that the notion of Statistical Literacy is an independent construct that can be inferred from a range of student behaviors. Second, forms of assessment that are different from tests or surveys are appropriate to use to assess students' development of Statistical Literacy.

This finding suggests that teachers who wish to identify their students' achievement of Statistical Literacy can make a choice about assessment methods that are appropriate to their context and the curriculum frameworks that they must address. A developmental framework is provided by the Statistical Literacy hierarchy and this can be measured using different approaches including those that are traditional and forms of performance assessment that are closer to students' classroom experiences. Identifying different forms of assessment that reliably measure the same construct is a useful step forward.

REFERENCES

Adams, R. J. and Khoo, S. T. (1996). *Quest: Interactive Item Analysis*. Melbourne: ACER.

Australian Education Council (AEC). (1994). *Mathematics—A Curriculum Profile for Australian Schools*. Melbourne: Curriculum Corporation.

Callingham, R. and Griffin, P. (2000). Towards a framework for numeracy assessment. In J. Bana and A. Chapman (Eds.), *Mathematics Education Beyond 2000. Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia*, (pp. 134-141). Fremantle: MERGA.

Callingham, R. and Watson, J. M. (2005). Measuring statistical literacy. *Journal of Applied Measurement*, 6(1), 29, 19-47.

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

Griffin, P. and Forwood, A. (1990). *Adult literacy and numeracy scales*. Canberra: Department of Employment, Education and Training.

Holmes, P. (1980). *Teaching Statistics 11-16*. Slough, UK: Schools Council and Foulsham Educational.

Keeves, J. P. and Alagumalai, S. (1999). New approaches to measurement. In G. N. Masters and J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment*, (pp. 23–42). Oxford: Pergamon.

Krulik, S. and Rudnick, J. A. (1999). Innovative tasks to improve critical and creative thinking skills. In L.V. Stiff and F.R. Curcio (Eds.), *Developing Mathematical Reasoning in Grades K-12. 1999 Yearbook.* Reston, VA: National Council of Teachers of Mathematics.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47, 149-174.

National Council on Education and the Disciplines (NCED). (2001). *Mathematics and Democracy: The Case for Quantitative Literacy.* Princeton: Woodrow Wilson Foundation.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research (Expanded edition, 1980. Chicago: University of Chicago Press).

Steen, L. A. (1999). Numeracy: The new literacy for a data-drenched society. *Educational Leadership,* 57(2), 8-13.

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

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