

Assessment in statistics education: Obstacle or opportunity?

Jane M. Watson

University of Tasmania, Australia



RL

Regular Lecture

Abstract

Assessment is one of those words that conjures up different meanings and feelings for different people. In choosing between using the words “assessment” and “measurement” in the title, assessment was chosen because, despite the anxiety it may create for some, it seems to encompass a broader range of ideas. Measurement seems to imply turning “something” into numbers and that is not always appropriate. Intuitively measurement is usually considered as a subset of assessment and the aim of this paper is to view the subject from the widest angle possible. The perspective of the paper is not that of a specialist educational measurement or assessment person. Most people in education, however, are forced, or even choose, to be in situations where they have to carry out assessments. This paper hence aims to present some of the variety that can exist in the field of assessment in statistics education. After briefly considering some of the potential obstacles of assessment, some of the wonderful opportunities that present themselves are explored.

Obstacles

Traditional views of statistics

Nearly everyone who has been enrolled in a statistics course has faced assessment in the form of a test. In the “olden” days this used to be a test where students had to remember formulae, insert the numbers in appropriate places, and calculate (without a calculator) the correct value to two decimal places. Perhaps this then had to be interpreted in the light of a table of significant values. In more recent times, progressive assessment may have allowed the provision of a formula sheet so that memorization was unnecessary. Recognition still was necessary, however, to pick the right formula before carrying out the rest of the process as before, maybe now with a calculator, basic or scientific. Another typical assessment was to draw a graph of a certain type from a data set, perhaps a histogram, where the trick was to guess the same interval widths as those in the mind of the examiner.

Luckily today the existence of graphics calculators and other technology makes it possible to assess students on much more interesting tasks, like what can be *learned* from statistics or graphs. It also makes possible the use of real data that would have been impossible in the days of hand calculation. Even further along, it is possible to consider actual contexts that have relevance to students’ lives, such as drug tests or environmental experiments.

Views about teachers

As assessment becomes more creative, there is then the debate about who assesses, and how capable they are of judging increasingly sophisticated responses. Teachers were trusted to mark procedural and multiple choice questions, but can they be trusted to mark open-ended high-level tasks reliably? Further, can they write assessment tasks at levels appropriate for a range of abilities? The folk lore has generally held that

teachers are unaware of what to assess, of how to write good “creative” items, and of how to develop rubrics to score responses in a hierarchical fashion. Although there may be some truth in this lore, one may suspect it is spread by measurement specialists who wish to protect their empires and their jobs!

The belief behind this paper is that if teachers are an obstacle, then they must be educated in how to carry out the task. If assessment is an integral part of the learning cycle, then difficulties with appreciating higher order assessment are likely to be associated with difficulties in teaching and assisting students to achieve these higher level outcomes. It may be that some professional development is necessary to raise teachers’ awareness, understanding, and skills in this area.

Does assessment drive content?

In the days when system-wide testing was common in many parts of the world, including the Australian state of Tasmania, there was a maxim, “if you want teachers to teach a topic, put it on the end-of-year exam.” It worked in Tasmania in that from the middle of the year on, teachers in years 11 and 12 based most of their assignment questions on old exam questions. In fact, the main annual money-making enterprise of the local Mathematical Association was selling booklets of solutions to the previous year’s examinations. Over the years quite a good bank of questions could be built up by teachers.

One has to be very careful condemning such a way of implementing the curriculum if in fact the test items reflect adequately the goals of the curriculum. As the curriculum has changed in recent times and assessment in many places has become school-based, however, this “solution” is no longer viable. The obstacle may even become more difficult if goals in areas such as statistical literacy are nebulously stated. The question becomes how to ensure that new assessment tasks measure the content that statistics educators want to be covered.

These are a few of the obstacles facing educators in devising new assessments in statistics. Besides this, tasks should be motivating, capable of being answered at various levels, useable in various classroom settings (besides paper-and-pencil tests), and perhaps even entertaining!

These potential obstacles are not very far removed from the challenges of assessment for statistics educators, as initially put forward in a book edited by Iddo Gal and Joan Garfield (1997). They saw the three main challenges as: to assess what the curricular goals actually represent, not some technical subskill; to assess conceptual understanding rather than the number correct; and to provide innovative models beyond paper-and-pencil tests (Gal and Garfield, 1997). Using their framework, suggestions are made in this paper of some opportunities for assessment that can enhance learning and increase the chances of reaching the goals of statistics education.

The opportunity to assess curricular goals

In considering curricular goals it is necessary to realize that these vary from country to country depending on national statements or internally depending on local requirements. In some places the curriculum for chance and data is very skills based, for example working out probabilities of simple events based on sample spaces, calculating means and medians, or drawing certain kinds of graphs. These goals are assessable by conventional production-of-single-outcome tasks when stated in this way, and it is important to move to more interesting goals.



These interesting curricular goals relate to higher order thinking, for example to appreciation that variation is the foundation of all experiences in chance and data (Moore, 1990), to making inferences based on consideration of all possible information, to critical thinking about questionable claims, and to seeing issues when they are embedded in realistic contexts. In advocating that these goals be assessed it is not necessary to reject paper-and-pencil instruments, but to devise innovative tasks that motivate students to be involved in finding solutions. Consider a few examples.

Variation

If it is desired to assess students' appreciation of variation in chance or data settings, it is necessary to give students the opportunity to show variation as they conceive it in an actual situation. Mike Shaughnessy started the ball rolling on this one with what is called the "lollie" problem in Australia or the "candy" problem in the United States.

Imagine a container with 100 candies in it, of three colours: 50 red, 30 green, and 20 yellow. First imagine drawing 10 candies without looking and predict the number of reds. Then imagine doing this many times, with replacement after each draw, and predicting successive numbers of reds. What would you predict for six draws? What would your frequency graph of 40 draws look like? The first question is about chance and proportion but the others concern "variation about expectation" and can tell much about students' beliefs. Students with a deterministic view of theoretical probability, for example, are likely to respond "5, 5, 5, 5, 5, 5" for six draws but will they do the same for 40 draws? There are many issues here and students can also be asked to select the most reasonable outcome from multiple choice options or to provide a range of outcomes. Various kinds of tasks, for both written surveys and interviews, have been trialled based on this scenario and there is a growing literature analysing the research outcomes (Shaughnessy, Watson, Moritz, and Reading, 1999; Reading and Shaughnessy, 2000; Torok and Watson, 2000; Kelly and Watson, 2002).

Statistical literacy

As curricular expectations broaden from just providing subject-matter experts at the end of secondary schooling, say in mathematics or foreign languages, it is important to express goals for everyone. Hence statistical literacy, or "statistics for all" (to borrow a phrase from mathematics), should be a goal for all students when they leave school. This objective has been described by Gal and Garfield (1997) in the first of their two goals for students of statistics at school:

"Comprehend and deal with uncertainty, variability, and statistical information in the world around them, and participate effectively in an information-laden society." (p. 2)

In an attempt to characterize or measure the development of this type of understanding, Watson and Callingham (2003) used a set of 80 questions varying in difficulty, content, context, and familiarity to school students. With data from about 4000 students, Partial Credit Rasch analysis (Masters, 1982; Rasch, 1980) was used and a complex hierarchy of tasks identified. Moving up through the hierarchy, expectations for tasks were clustered into six levels reflecting increasing appreciation of elements of the school curriculum, of context, of mathematical ideas such as proportion, and of critical thinking. These are briefly summarized in Table 1 with a few examples of typical achievement at the levels.

RL

Regular Lecture



RL

Regular Lecture

Level	Description
1. Idiosyncratic	Tautologies, one-to-one counting, read cells.
2. Informal	Intuitive non-statistical beliefs (3 is lucky), one-step calculations.
3. Inconsistent	Limited appreciation of content and context without justification; qualitative ideas.
4. Consistent Non-critical	Straight-forward engagement with context; means, simple probabilities and graphs.
5. Critical	Questioning engagement; appreciation of variation; qualitative interpretation of chance.
6. Critical Mathematical	Questioning critical engagement with context, proportional reasoning, subtle language.

Table 1. Levels of Statistical Literacy (from Watson and Callingham, 2003)

This survey method of assessment is vastly different from interviews where students carry out experimentation with containers full of candy. Here assessment does become “measurement” and students (and items) do appear on a scale relative to one another. Care must be taken in this case to carry out measurement in such a way that outcomes are interpretable. Watson and Callingham (2003) believe that Table 1 provides information useful for planning interventions to assist students to progress through the hierarchy; further research by others will hopefully confirm this belief. This model for measuring statistical literacy is considered briefly again later in the paper.

The opportunity to assess conceptual understanding

In discussing conceptual understanding, examples like variation, discussed in the last section, or sampling, or inference, that are not based on procedural numerical calculations that can be memorized, are useful starting points. The concepts are complex enough that one can envisage them being built up in stages, perhaps by putting successively more elements together. Although various types of frameworks exist for describing complex understanding, ones that reflect intermediate steps using more relevant ingredients are attractive. Outcomes often fit with the steps observed as children mature and can give clues to instructional opportunities.

A model that is useful in analysing student progress toward conceptual understanding is one from cognitive psychology devised by Biggs and Collis (1982). Called SOLO, it is based on the Structure of Observed Learning Outcomes, with an emphasis on what a student actually produces rather than what the researcher thinks the student might have meant. Such a scheme is hence based on classifying a response, not a student, acknowledging that at a different time the response may be different. This puts pressure on the teacher to devise tasks that are valid and reliable in providing opportunities for responses at all possible levels. A task that is too easy, so that students are able to provide a reasonable or acceptable answer at a lower than optimal level, does not allow or challenge the student to show what can be accomplished.

For most practical purposes the model can be applied to a concept as a four- or five-step progression toward a goal set in a task. The steps as adapted from Biggs and Collis (1982) are the following.

- (i) *Prestructural (P)* responses do not engage any facet of the task reflecting the concept in question.
- (ii) *Unistructural (U)* responses employ single elements relevant to the task and if a contradiction of terms occurs, it is not recognized.
- (iii) *Multistructural (M)* responses employ multiple elements in a sequential fashion and if conflict occurs it is likely to be recognized but not resolved.
- (iv) *Relational (R)* responses integrate multiple elements into a whole and resolve conflict should it arise with respect to a task set within the scope of these levels of response.
- (v) *Extended Abstract (EA)* responses go beyond the expectations of the task and bring in unexpected more sophisticated insights.

Although tasks can be devised to require thinking at any of the four top levels of the hierarchy, often tasks are written for the Relational level. In fact it is possible to devise a task one considers complete, without any reference to the SOLO structure, and find that when responses are ordered by complexity, they fit the model.

Consider for example a task set by Anthony Kelly and his colleagues (Kelly, Sloane, and Whittaker, 1997) to summarize a data set of heights of 20 women in any way considered appropriate. They were interested in how their students would consider the distribution of data values using an available statistical package and this influenced the four ways they categorized the responses. A close look, however, reveals structural similarities to the SOLO model in the increasingly appropriate categories of response given by college students:

- (1) mindless relevance on statistical packages,
- (2) generating a mean value without plotting the data,
- (3) plotting the data and generating a mean value uninformed by the plot, and
- (4) plotting the data and linking the choice of statistic to distributional assumptions.

With respect to the goal of the task, the first category of response is prestructural, as the students do not indicate they would make any decisions but just “believe” what the software says, whatever that is. There is no indication that any further interpretation would take place. For the second category students rely on a single element for their solution, the mean. For the third category (M), students carry out two procedures, plotting the data and calculating the mean, but then make no effort to combine the elements to make meaning from the data. In the final category (R) responses consider the plot of the data, realizing the presence of an outlier and choosing an appropriate statistic based on the distribution. One might be tempted to claim that good tasks allowing high level conceptual outcomes are very likely to be analysed in this hierarchical fashion.

A categorization, such as described above, suits two purposes. On one hand, if used as an assessment task, then giving increasing points to each category reflects the sophistication of the understanding of the usefulness of considering data distributions. On the other hand, if the desire is to assist students in reaching higher categories, it

appears clear from a student's placement in the hierarchy, what is needed in the way of instruction.

Variation revisited

Returning to the example on Variation discussed in the previous section, the analysis of interview data on the candy problem gives another example of the potentially hierarchical structure in assessment (Kelly and Watson, 2002). Students in pre-Grade 1 and Grades 3, 5, 7, 9, were interviewed on a protocol related to the questions stated earlier and they provided examples of a wide range of performance. There was interest in students' appreciation of the need to consider both the proportion of red candies in the container and the variation about this proportion that might occur in repeated draws of 10 candies (with replacement after each draw). This was judged by the reasoning given to support the numerical values students suggested and the graphs they drew to represent what would happen over 40 repeated draws.

At the Prestructural Level students provided intuitive explanations, such as "6 red, because it is my favourite number" and drew pictures of children drawing candy from bags. At the Unistructural Level, they recognized the importance of there being "more red" in the container, a single feature, but were likely to be inconsistent across different forms of the task, for example choosing values for repeated trials that were inconsistent with the ranges they themselves suggested. Graphs at this level were likely to be of a time-series type with values between 0 and 10 but representing uniform rather than centred variation. At the next level, students considered more features, including the ideas of "more" or "half" red and an appreciation of 5 as the centre of the values they were likely to obtain. Given help in the form of a blank graph with axes, many of these students could draw idiosyncratic representations that demonstrated clustering about the centre. At the highest level, students related together the information about the proportion of reds in the container and typical variation for 40 trials and could produce a centred frequency representation that was reasonable (although often still with more variation than appropriate). Typical graphs for these four levels of response are shown in Figure 1.

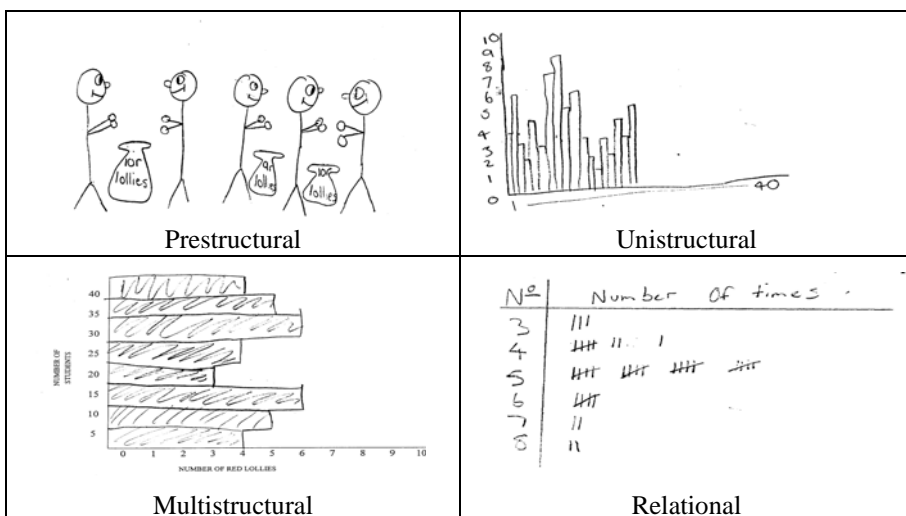


Figure 1. Typical graphs for the four levels of response for the candy problem (from Kelly & Watson, 2002)



C M E
1 0
2 0 0 4

RL

Regular Lecture



I C M E
1 0
2 0 0 4

RL

Regular Lecture

Beginning inference

It would not be appropriate to assume, however, that all assessment in statistics can or should follow a hierarchical structure as suggested by Biggs and Collis. Sometimes quite straightforward tasks allow the discovery of interference from elsewhere in the mathematics curriculum, and categories of response are difficult to distinguish structurally but provide valuable information for teachers. Consider a task that has been used in various forms in interviews and surveys (e.g., Watson and Kelly, 2003). In Figure 2 there are six questions, some included as introductory and to ensure that younger students can engage in the task. In terms of considering variation and drawing tentative inferences from basic graphs, the last four questions are of more interest. For the question about the graph looking the same everyday, a large majority of students (70% in Grades 3 to 9) use language like “might” and “could” to indicate that change is likely from day to day. Similar language is *not* used (by 84% of students) to explain about the meaning of the row with the train that is empty. It is hence of interest to explore the arguments used in the other two questions, which ask for inferences to be drawn in the context of the graph: (1) A new student comes to school by car. Is the new student a boy or a girl? How do you know? (2) Tom is not at school today. How do you think he will get to school tomorrow? Why?

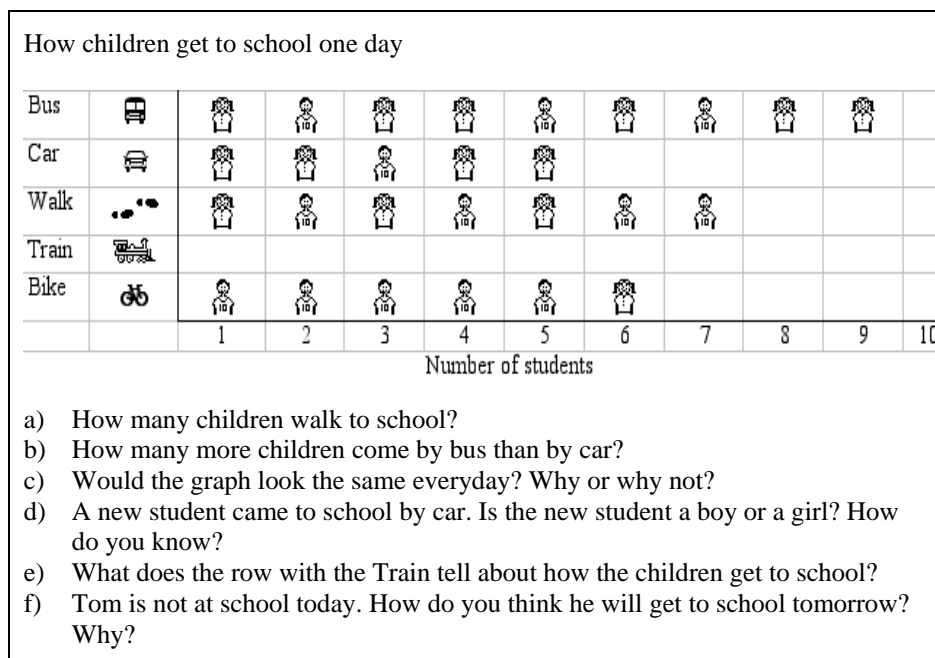


Figure 2. Travel graph and survey questions

Statistical thinkers know that they can make suggestions to answer these two questions based on the information in the graph but that these must be speculative because the graph provides likely but not certain information. Hence one might say, “The new student is probably a girl because more girls than boys come to school by car.” Or “Tom probably comes by bus because more students come by bus.” Or even “Tom might come by bike because more boys come by bike.” In contrast to these responses,

what do students in the middle school grades say to these questions and what kinds of understanding are displayed?

Watson and Kelly (2003) identified six categories of response that they considered generally hierarchical; other researchers, however, may disagree about the relative merits of the responses. Table 2 summarises the categories.



RL

Regular Lecture

Summary	New student boy or girl?	How does Tom get to school?
[0] Inappropriate [Don't seem to understand question]	There were more kids. Don't know.	He will feel better after a day off. Not sure.
[1] No interaction with the graph	Boy, I just guessed. Not sure, not enough information.	Car, so he doesn't get a cold. The same way he does every day.
[2] Patterns or Anything can happen	Boy, because there is a pattern (GGBGG...) Girl, she is at the end. Could be either.	Car, because there is a pattern (GGBGG...) Anything, it's chance!
[3] Balancing	Boy, cause it's the only boy that goes by car. Boy, it could make 14 of both in the class.	Train, because there is no one on the train today.
[4] Statistical reasons but no uncertainty	Girl because the majority who come by car are girls.	Bike, the majority of boys ride to school. Bus, more people catch the bus.
[5] Statistical reasons acknowledging uncertainty	You don't know but it is just more likely to be a girl 'cause more come by car.	Probably by bus because 1/3 of the children caught it today.

Table 2. Responses to Pictograph Question (from Watson & Kelly, 2003)

What is intriguing in Table 2 are the numbers of students who respond in the middle two categories ([2] and [3]), especially for the new student question. Teachers are not at all surprised and suggest that there is interference from the pattern work that is done these days in the effort to prepare students for work with algebra. Yes, patterns are important all across the mathematics curriculum, but in statistics the interest focuses on different sorts of patterns than those related to algebra. It is also of note that very few middle school students acknowledge uncertainty and potential variation when giving responses based on the information in the graph. Perhaps some teachers may be over-emphasizing the deterministic power of information obtained from graphs.



RL

Regular Lecture

The opportunity to develop innovative models for assessment

Claiming to be innovative in assessment is fraught with tension because it may appear that someone is always prepared to express strong opinions from one perspective or another. Practicality is a problem for some in looking at Peter Holme's (1997) innovative suggestions for assessing project work. How do you choose topics? How do assessment criteria change? How long does assessment take? On the other hand multiple choice questions, which are practical and economical to mark, are often criticised for not addressing critical ideas or for providing too much scaffolding in suggesting the correct response within a series of alternatives. Chris Wild and his colleagues' (Wild, Triggs, and Pfannkuch, 1997) attempt to address these issues but also confess that the preparation of excellent multiple choice questions is a time-consuming process. Assessment for students working in groups (Chick and Watson, 2001) or using technology (Lajoie, 1997) also raises issues for educators.

One assessment context that appears not yet to have been used extensively, is that of the video extract. Until recently it has not been easy to organize equipment and video players to show video extracts; but with digital video and the availability of computers, this becomes a possibility. Sometimes a well-known television snippet has the potential to make an excellent stem for an assessment task. One such extract is an advertisement for a Ford automobile shown a few years ago in Australia. It features the statement, "The average young family has 2.3 children" and a small boy with a large ".3" printed on his t-shirt. The advertisement lauds the merits of a small car that has "room for all of the .3s of Australia."

Why is this a good starter for an assessment task? First, the advertisement is very humorous: the ".3" boy is a sort of mathematical-statistical pun. This advertisement, however, is presented as fact in the media and it is important for people to be able to question the information provided and have the knowledge to know that the three children appearing in the extract cannot be legitimately described as "2.3 children" regardless of the space they take up in the car. This is a basic requirement of statistical literacy as noted earlier by Gal and Garfield (1997).

What kind of a task, then could be based on this advertisement. Perhaps one that asked, "What does it mean for a family to have 2.3 children?" and "How can you interpret the meaning of this advertisement?" Many might think that the tasks are very easy but throughout the middle school years when decimals are first introduced and the definition of the arithmetic mean is not well-entrenched, the responses to the first question are perhaps not what would be expected.

- Well it might be because they might just want that many children or something like that.

Some students have an appreciation of the part-whole nature of decimals but not an understanding of how the arithmetic mean operates within that structure. This may result in quite different interpretations of 2.3 children.

- Well, someone might have 2 children, a mum might have 2 children or something, and she might be pregnant.
- Well, some have 2 children and others have 3. More likely to have 2 though
- That out of a certain amount of Australian families, the most common amount of children is 2.3.

- When looking at Australian families they usually have between 2 and 3 children, because you can't get .3 of a person, so they would have to have between 2 and 3 children.
- Most families would have around 2.3 children, and it's not like there's a child who's so small that it's a .3, but it's all divided ... like you add up the number of children and then you divide it by the number of families and it came out at 2.3.

In interpreting the meaning of the advertisement, however, some students take it quite seriously and develop elaborate descriptions that fit the story line.

- Well, that they have got two full grown kids, and one's not full grown yet.
- It says that most Australian families have two older children and say one infant or child under the age of 5 or whatever.
- Well because the average is of like the older children, which they could say is fully grown or my age or whatever. And the .3 is a child that is growing up to be an older child. So that, like, say the kid is 3 now, once it turns to be 10, it will get to be 1, so they will have 3 children sort of thing.

An intuition about the mean appears to be present in the following answer.

- That people, evened out, have 2.3 children. You can't have .3 of a child but that is just how it worked out. I know because my aunty has 4 children and I'm an only child. That's five for two families. That's 2 and a half people.

An extension to the use of video extracts is the showing of extracts of some of these responses to challenge other students' understanding, depending on their initial responses. This has been done in exploring students' understanding of other statistical concepts, for example beginning inference (Watson, 2002), and appears to be motivating to students.

Not only are there innovative models for how assessment occurs but also of what criteria are important for students to satisfy and in what contexts they are expected to do so. The structural model introduced in the previous section makes no judgements about what statisticians think is important for people to know. This has to be included in the task set for assessment. As noted at the beginning, the reflection of school curriculum objectives is a starting point but many statistics educators would like to go further than the expectations there. By the time students leave school they should be critical thinkers and ask questions in contexts such as the media they encounter everyday. In fact Gal (2003) has suggested that people should also be expected to be critical consumers of government and other public reports. He further emphasises the importance of motivation to be critical thinkers (Gal, 2002), but again devising effective assessment for this aspect is a challenge (Likert scales are unlikely to be sufficient).

One way to consider the development of critical thinking is through a hierarchy that acknowledges the components that appear to be related and contribute to critical awareness. The three tiers suggested by Watson (1997) are:



RL

Regular Lecture

- (i) understanding of the statistical terminology to be used,
- (ii) understanding of the terminology when it appears in various contexts, including social, scientific and technical contexts appearing in media or other reports,
- (iii) ability (and motivation) to question claims that are made without proper statistical justification (and even to explore and assess those made with adequate justification).

The hierarchical nature of this model may suggest similarities to the structural model noted earlier but recent research has identified U-M-R cycles within and across these tiers, depending on the tasks used (Watson and Moritz, 2000). Based on four tasks, the information in Table 3 shows how the levels referred to in the previous section can be observed within each of the tiers. The Tier 1 task is the definition of Sample. The Tier 2 task relates to the information used to purchase a car. The two Tier 3 tasks are based on a voluntary radio poll on legalizing marijuana and a newspaper report of a non-representative sample of access to guns in the United States based only on Chicago. In the first two tiers three or four SOLO levels are identified as students develop appreciation of terminology and context, whereas five levels appear when critical thinking is involved.

Level	Tier (Associated Questions)		
	1: Defining Sample Terminology	2: Applying Sampling in Context	3: Questioning Sample Claims
P	Something too hot	-	Shouldn't have guns Should decriminalise marijuana
U	A part <u>OR</u> A test	(T) Experience of three friends	Some could be lying
M	Small part, not all <u>OR</u> A blood test	(=) Could be unlucky either way	Not all schools like that Need to ask everyone
R	Small part representing total	(H) Report based on 800 cases	Large sample good Sample not large enough
EA	-	-	Chicago does not represent the US <u>AND/OR</u> Voluntary poll, only JJJ listeners

Table 3. Examples of Response Levels within Tiers of Statistical Literacy (from Watson & Moritz, 2000)

Applying the hierarchical model to a single simpler task, consider the pictograph task introduced in the previous section (Figure 2). For the goal of critical questioning in relation to graph interpretation, the development of student thinking can be seen across the tiers from the responses in Table 2.

- Pre-Tier 1: Students do not interact with the graph at all, perhaps making suggestions based on their personal experience [0], or if they look at the graph they do not know what to do with the information in it [1].
- Tier 1: Students are able to read the information from the graph (they know what a graph represents) but their interpretations are based on information that is not relevant to the context of the question. They look at patterns in the graph that may be interesting but are not relevant to the questions asked [2] or they try to identify special cases or make suggestions considered “fair,” again not relevant to the statistical point of view [3].
- Tier 2: Students are able to read the graph in the intended context and use it to make appropriate interpretations for the data [4]. These statements, however, are deterministic in nature.
- Tier 3: Students go beyond the basic interpretation of the information in the graph to include an element of uncertainty in their predictions, acknowledging that variation is possible [5]. This uncertainty is the foundation of the questioning attitude associated with critical statistical literacy.

On a much broader scale it is also possible to suggest links between the three-tiered hierarchy and the six levels of statistical literacy understanding suggested early in this paper based on Rasch analysis. The use of Tier 1 skills and terminology appears throughout the tasks at all levels, with the most sophisticated appearing at the highest levels. Engaging in the contexts of statistical literacy, the concern of Tiers 2 and 3, appears from Level 3 upward, whereas thinking critically as required in Tier 3 appears in Levels 5 and 6. This is not surprising given the hierarchical nature of the model and the Rasch measurement technique. It is interesting that the development of each tier of the hierarchy continues to be observed in successive levels of the statistical literacy construct, indicating for example that the development of skills and terminology continue alongside the development of appreciation of context and then critical thinking. Statistical literacy is indeed, a complex, interwoven phenomenon.

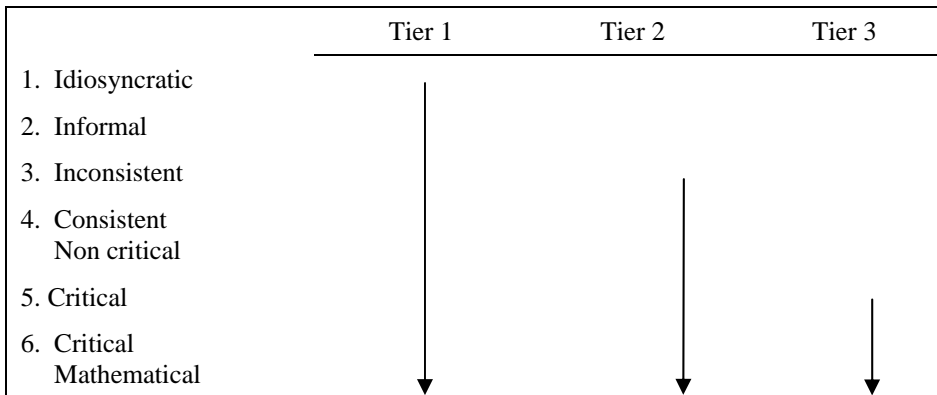


Figure 3. Tiers and levels of statistical literacy from Rasch analysis

Returning briefly to the tasks related to the Ford advertisement, students who accept the advertisement at face value are Pre-Tier 1 in the statistical literacy hierarchy. Students who know about average in the sense of most, middle, or mean are in Tier 1, whereas those who begin to grapple with the meaning of the advertisement in relation to their understanding of average are in Tier 2. Finally those who can explain the humour of the advertisement in relation to the meaning of the arithmetic mean and part-whole numbers are operating in Tier 3.

Conclusion

There are many possible angles for considering assessment in statistics education and those chosen for this paper reflect particular interests and research studies. One of the concerns for statistics educators is not to turn some of these opportunities into obstacles in the way they are implemented. The issue of measurement is very likely to enter the assessment scene at some point but it must be carried out in a manner that numbers, if used, are interpretable and useful in a descriptive way to add meaning to the process. This was the intention of the analysis employing Rasch methods described here. Far from being a boring or technical topic, assessment opens windows to understanding and can assist in planning future learning experiences.

References

- Biggs, J., & Collis, K. (1982). *Evaluating the quality of learning: The SOLO taxonomy*. New York: Academic Press.
- Chick, H.L., & Watson, J.M. (2001). Data representation and interpretation by primary school students working in groups. *Mathematics Education Research Journal*, **13**, 91-111.
- Gal, I. (2002). Adults' statistical literacy: Meanings, components, responsibilities. *International Statistical Review*, **70**, 1-51.
- Gal, I. (2003). Functional demands of statistical literacy: Ability to read press releases from statistical agencies. In *Bulletin of the International Statistical Institute 54th Session Proceedings Berlin* (Volume LX, Book 2, Invited Papers, Topic 49, pp. 46-49). Berlin: ISI.

- Gal, I., & Garfield, J. (1997). Curricular goals and assessment challenges in statistics education. In I. Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 1-13). Amsterdam: IOS Press and The International Statistical Institute.
- Holmes, P. (1997). Assessing project work by external examiners. In I. Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 153-164). Amsterdam: IOS Press and The International Statistical Institute.
- Kelly, A. E., Sloane, F., & Whittaker, A. (1997). Simple approaches to assessing underlying understanding of statistical concepts. In I. Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 85-90). Amsterdam: IOS Press and The International Statistical Institute.
- Kelly, B. A., & Watson, J. M. (2002). Variation in a chance sampling setting: The lollies task. In B. Barton, K.C. Irwin, M. Pfannkuch, & M.O.J. Thomas (Eds.), *Mathematics education in the South Pacific: Proceedings of the Twenty-fifth Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 366-373). Sydney: MERGA.
- Lajoie, S. P. (1997). Technologies for assessing and extending statistical learning. In Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 179-190). Amsterdam: IOS Press and The International Statistical Institute.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, **47**, 149-174.
- Moore, D. (1990). Uncertainty. In L. Steen (Ed.), *On the shoulders of giants: new approaches to numeracy* (pp. 95-137). Washington, D.C.: National Academy Press.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (expanded edition). Chicago: The University of Chicago Press. (Original work published 1960).
- Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th annual conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima, Japan: Hiroshima University.
- Shaughnessy, J. M., Watson, J., Moritz, J., & Reading C. (1999, April). *School mathematics students' acknowledgment of statistical variation*. In C. Maher (Chair), There's more to life than centers. Pre-session Research Symposium conducted at the 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.
- Torok, R., & Watson, J. M. (2000). Development of the concept of statistical variation: An exploratory study. *Mathematics Education Research Journal*, **12**, 147-169.
- Watson, J. M. (1997). Assessing statistical literacy using the media. In I. Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 107-121). Amsterdam: IOS Press and The International Statistical Institute.
- Watson, J. M. (2002). Inferential reasoning and the influence of cognitive conflict. *Educational Studies in Mathematics*, **51**, 225-256.
- Watson, J. M., & Callingham, R.A. (2003). Statistical literacy: A complex hierarchical construct. *Statistics Education Research Journal*, **2**(2), 3-46.

- Watson, J. M., & Kelly, B. A. (2003). Inference from a pictograph: Statistical literacy in action. In L. Bragg, C.Campbell, G. Herbert, & J. Mousley (Eds.), *Mathematics education research: Innovation, networking, opportunity: Proceedings of the Twenty-sixth Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 720-727). Sydney: MERGA.
- Watson, J. M., & Moritz, J. B. (2000). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior*, **19**, 109-136.
- Wild, C., Triggs, C., & Pfannkuch, M. (1997). Assessment on a budget: Using traditional methods imaginatively. In I. Gal & J.B. Garfield (Eds.), *The assessment challenge in statistics education* (pp. 205-220). Amsterdam: IOS Press and The International Statistical Institute.



RL

Regular Lecture