STATISTICAL LITERACY: A COMPLEX HIERARCHICAL CONSTRUCT

JANE WATSON
University of Tasmania
Jane.Watson@utas.edu.au

ROSEMARY CALLINGHAM
University of New England
rcalling@pobox.une.edu.au

SUMMARY

The aim of this study was, first, to provide evidence to support the notion of statistical literacy as a hierarchical construct and, second, to identify levels of this hierarchy across the construct. The study used archived data collected from two large-scale research projects that studied aspects of statistical understanding of over 3000 school students in grades 3 to 9, based on 80 questionnaire items. Rasch analysis was used to explore an hypothesised underlying construct associated with statistical literacy. The analysis supported the hypothesis of a unidimensional construct and suggested six levels of understanding: Idiosyncratic, Informal, Inconsistent, Consistent non-critical, Critical, and Critical mathematical. These levels could be used by teachers and curriculum developers to incorporate appropriate aspects of statistical literacy into the existing curriculum.

Keywords: Statistical literacy; school students; Rasch analysis; conceptual hierarchy

1. INTRODUCTION

Historically there are two antecedents to this study. One is the growing interest, from the middle of the last century, in numeracy as a co-agent with literacy becoming an essential education foundation for all school students. The other is the specific inclusion of data handling and chance in the mathematics curricula of many countries toward the end of the century.

1.1. EMERGENCE OF NUMERACY

Although the term literacy has been accepted for a long time as describing the ability to read and write, the acceptance of a similar term has been more elusive in describing number or mathematical skills. The term numeracy was introduced in the Crowther Report in the United Kingdom in 1959:

> *A word to represent the mirror image of literacy… On the one hand an understanding of the scientific approach to the study of phenomena — observation, hypothesis, experiment, verification. On the other hand… the need in the modern world to think quantitatively, to realize how far our problems are problems of degree even when they appear to be problems of kind. Statistical ignorance and statistical fallacies are quite as widespread and quite as dangerous as the logical fallacies that come under the heading of illiteracy* (quoted in Cockcroft, 1982, para. 36).

Statistics Education Research Journal 2(2), 3-46, http://fehps.une.edu.au/serj

For those interested in statistical literacy it may be regretted that this definition was not immediately and universally accepted. The dilution of the term numeracy to include only basic skills with numbers, has led some people into the multiple literacies milieu, using the phrase "quantitative literacy" to describe the broad range of understanding required when students leave school. This is particularly true in the United States where Quantitative Literacy is in the title of recent significant reports on the need to improve mathematical and statistical understanding of the general population (Steen, 1997, 2001). In Australia, however, the term numeracy continues to have a broad meaning. For example, the following description of numeracy was endorsed in 1997 at a conference organized by the Australian Association of Mathematics Teachers Inc (AAMT).

> *To be numerate is to use mathematics effectively to meet the general demands of life at home, in paid work, and for participation in community and civic life. In school education, numeracy is a fundamental component of learning, performance, discourse and critique across all areas of the curriculum. It involves the disposition to use, in context, a combination of:*
>
> - *underpinning mathematical concepts and skills from across the discipline (numerical, spatial, graphical, statistical and algebraic);*
> - *mathematical thinking and strategies;*
> - *general thinking skills; and*
> - *grounded appreciation of context* (AAMT, 1997, p. 15).

1.2. STATISTICS IN THE SCHOOL CURRICULUM

The second antecedent to interest in statistical literacy was the introduction of revised mathematics curricula in the early 1990s around the world (e.g., National Council of Teachers of Mathematics [NCTM], 1989; Ministry of Education, 1992). In Australian states these curricula were based to a greater or less extent on *A National Statement on Mathematics for Australian Schools* (Australian Education Council [AEC], 1991), which included chance and data as one of five content areas of the mathematics curriculum. Written with the advice of statisticians and educators, the chance and data curriculum could not benefit from previous research into school students' understanding of probability and statistics concepts as, with the exception of some work in the area of probability (e.g., Fischbein, 1975; Green, 1982, 1983b), there had been virtually none.

Most national curriculum documents (e.g., AEC, 1991; Ministry of Education, 1992; NCTM, 1989) reflected the five components that comprise a statistical investigation based on a question of interest, as suggested by Holmes (1980): data collection, data tabulation and representation, data reduction, probability, and interpretation and inference. In Australia Holmes' categories were aggregated into three major subheadings of the mathematics curriculum: chance, data handling, and statistical inference. In the latest standards of the NCTM (2000) in the United States these ideas are included in the Data Analysis and Probability Strand:

> *Instructional programs from pre-kindergarten through grade 12 should enable all students to:*
>
> 1. *Formulate questions that can be addressed with data and collect, organize, and display relevant data to answer them;*
> 2. *Select and use appropriate statistical methods to analyze data;*
> 3. *Develop and evaluate inferences and predictions that are based on data;*
> 4. *Understand and apply basic concepts of probability* (NCTM, 2000, p. 48).

1.3. STUDENT UNDERSTANDING

The implications of the new curricula for the professional development of teachers and for the development of materials and activities for different grade levels were many. What did students know about the new topics? How did understanding develop over time? What alternative conceptions existed to complicate learning? These and similar questions provided the motivation for research on students' initial and developing understanding of the topics in the curriculum over the next decade.

This research has added much to our understanding of students' developing ideas on particular topics, such as average (Cai, 1995, 1998; Mokros & Russell, 1995), sampling (Watson & Moritz, 2000a), chance measurement (Metz, 1998), inference (Gal & Wagner, 1992), association (Batanero, Estepa, Godino, & Green, 1996; Moritz, 2000) and graphing (Friel, Curcio, & Bright, 2001; Mevarech & Kramarsky, 1997), as well as for the areas of probability and data handling more generally (Jones, Langrall, Thornton, & Mogill, 1997; Jones, Thornton, Langrall, Mooney, Perry, & Putt, 2000). These studies were based on interviews, moderate-sized surveys, or reviews of the literature. As well, large-scale surveys have added to the store of information on understanding of these topics (Watson & Moritz, 1998, 1999a, 2000b, 2002; Zawojewski & Shaughnessy, 2000).

In addition to research specifically related to topics listed in the curriculum, there were calls from Green (1993) and Shaughnessy (1997) for research into students' understanding of variation as the underlying factor that creates the need for the chance and data curriculum in the first place (Moore, 1990; Wild & Pfannkuch, 1999). Work in this area (e.g., Shaughnessy, Watson, Moritz, & Reading, 1999; Reading and Shaughnessy, 2000, in press; Watson & Kelly, 2002) has brought about the development of further instruments specifically to measure how understanding of variation is displayed in association with understanding of the other topics in the chance and data curriculum (e.g., Watson, Kelly, Callingham, & Shaughnessy, 2003).

1.4. FURTHER CURRICULUM CHANGE

Within a decade of the initial changes to the mathematics curriculum in Australia (AEC, 1991, 1994), new, more general moves were taking place in the school curriculum. Under descriptors such as *New Basics* (Education Queensland, 2000) and *Essential Learnings* (Department of Education Tasmania, 2002) a more integrated approach to all areas of the curriculum has focussed attention on critical skills linking literacy, numeracy, and information technology through a curriculum addressing cultural, aesthetic, scientific, and social issues in a holistic fashion. These moves force a serious consideration of the links between the two antecedents described in the opening paragraphs. The place of chance and data in the mathematics curriculum suggests a consequent need for a wider appreciation of numeracy and of the part to be played by statistical skills in relation to social and scientific thinking based on both literacy and numeracy. As statistical understanding is the foundation for many of the decisions made in society today (Wallman, 1993), aspects of statistical literacy – the application of statistical understanding in context – will be placed at the intersection of literacy and numeracy and will be essential to meeting the goals of the new curricula, particularly those associated with citizenship (Steen, 2001).

2. THE CONSTRUCT OF STATISTICAL LITERACY

The emergence of a description of a construct of statistical literacy has taken place within the contexts described in the previous section. It has links to the parallel development of descriptions of numeracy, quantitative literacy, critical literacy, and adult literacy, as well as the chance and data elements of the school curriculum. Although the building blocks for statistical literacy are found within the mathematics, chance, data, and literacy components of the school curriculum, for students to become statistically literate they also need to interact with a variety of contexts as they take their place as consumers of information in the adult world. Hence contexts also must be included in a final theoretical description of the construct. This provides a link from statistical literacy to the other notions of quantitative literacy, adult literacy, and critical literacy. As well, a final description is likely to acknowledge that growth is hierarchical in nature.

2.1. EMERGENCE OF THE STATISTICAL LITERACY GOAL

A somewhat apocryphal prophecy from H. G. Wells at the beginning of the twentieth century has been a starting point for many who argue for a high status for statistical literacy: "Statistical thinking

will one day be as necessary for efficient citizenship as the ability to read and write" (quoted in Castles, 1992, p. v). This claim is highly significant in terms of the move to an information society in the last third of the twentieth century as recognized in traditionally agriculture-dependent Australia by Jones (1982, p. 173): "Australia is an information society in which more people are employed in collecting, storing, retrieving, amending, and disseminating data than producing food, fibres and minerals and manufacturing products." Steen (1997, p. xv) reflected a similar view for a North American audience: "As information becomes ever more quantitative and as society relies increasingly on computers and the data they produce, an innumerate citizen today is as vulnerable as the illiterate peasant of Gutenberg's time."

These concerns were echoed in the desires of national governments, for example in Canada, the United States, and Australia, to survey the adult literacy skills of their citizens (Statistics Canada and the OECD, 1996; Dossey, 1997; McLennan, 1997). Although using the phrase "adult literacy" these surveys had three components: prose literacy, document literacy, and quantitative literacy. The terminology implies that numeracy skills would only be required for the last literacy, however, as is seen in the items used (e.g., McLennan, 1997), numerical information is embedded in many prose tasks and in most graphical and table-based documents to be interpreted. Further, basic statistical interpretation skills are required for many of the document and quantitative literacy tasks, for example summarizing information in pie charts and bar graphs, and working with percents and averages (e.g., Dossey, 1997; McLennan, 1997). Interest in adult literacy, at least implicitly, acknowledges a debt to statistical thinking/literacy. The stage is hence set for a consideration of the nature of the contribution of statistical literacy to the educational milieu.

For her Presidential Address to the American Statistical Association a decade ago, Katherine Wallman chose the topic of statistical literacy. In her brief definitional summary she focused on the application of understanding of the type developed during the school years for people as users rather than creators of statistics.

> *'Statistical Literacy' is 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*. (1993, p.1)

The two dimensions of statistical literacy – public and private – introduced by Wallman are significant when thinking of motivating learning. Statistical literacy is not only important to our society as a whole; it is also relevant to the individual members of society as they make decisions in their personal lives based on information and risk analysis provided by others in the community. Decisions related to where to live, what type of employment to seek, whether to gamble, or what car to buy may be influenced by data provided from outside of one's individual experience.

More recently in summarizing the current state of understanding concerning adult statistical literacy, Gal (2002) suggested that the requirements are contained in the following two components:

> (a) *people's ability to interpret and critically evaluate statistical information, data-related arguments, or stochastic phenomena, which they may encounter in diverse contexts, and when relevant,*
>
> *(b) their ability 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*. (pp. 2-3)

These components do not rely on sophisticated topics from the senior secondary curriculum such as standard deviation or hypothesis testing. They are built up as part of the chance and data curriculum, leading toward an expectation of critical thinking in many contexts across the school curriculum. This is envisaged in Australia's *National Statement*, which claims that "students should learn to question the assumptions underlying data collection, analysis and interpretation and the reasonableness of inferences and conclusions" (AEC, 1991, p. 164). It is also found in three extracts from the first NCTM statement of standards (1989):

> *In particular, citizens must be able to read and interpret complex, and sometimes conflicting information* (p. 5).

An understanding of probability and the related area of statistics is essential to being an informed citizen (p. 109).

A knowledge of statistics is necessary if students are to become intelligent consumers who can make critical and informed decisions (p. 105).

The transition from the components of a statistical investigation as set out, for example in the NCTM (2000) standards, to what is needed for survival in the world outside the classroom is illustrated in the New Zealand mathematics curriculum (Ministry of Education, 1992). "Interpreting statistical results" is one of three themes for statistics education and among the learning experiences is the following: "Investigating ways in which statistical information is presented in the media and other sources, and recognizing and identifying sources of deception in misleading graphs and their accompanying statements" (p. 189).

The use of the phrase "critically evaluate" by both Wallman (1993) and Gal (2002) in their descriptions of statistical literacy suggests a link to the area of critical literacy more generally. Adapting to statistical literacy the discussion of Luke and Freebody (1997), which related to the social practice of reading, four roles of a statistical text user are evident: the role as a code breaker, e.g., understanding the basic terminology of statistics; the role as a text participant, e.g., using knowledge to make sense of data, graphs, and chance claims embedded in text; the role as a text user, e.g., using data, graphs and chance concepts in particular social contexts; and the role as a text analyst, e.g., critically reading and seeing text as ideologically framed and constructing a position in relation to data driven claims. These categories of involvement mirror the views of Frankenstein (2001) in discussing the goals of a "critical mathematical literacy" curriculum. They are also useful in considering progression through a hierarchy of understanding.

2.2. THE HIERARCHICAL NATURE OF STATISTICAL LITERACY

Taking into account the rich and complex description of statistical literacy in the preceding section, the aim of this study is to explore and document the hierarchical stages associated with the goal as described for example by Gal (2002). Based on responses to tasks associated with statistical literacy, there are two aspects of understanding that can be described: increasing structural complexity and increasing statistical appropriateness. Two models were employed in this regard for the current study. The first was based on a cognitive framework (Biggs & Collis, 1982, 1991), which provides a structural hierarchy for responses: (i) *Prestructural* responses do not address elements of the task; (ii) *Unistructural* responses employ single elements of the task and do not recognize conflict should it occur; (iii) *Multistructural* responses employ elements in a sequential fashion and recognize conflict if it occurs but are unable to resolve it; (iv) *Relational* responses create connections among elements to form an integrated whole and resolve conflict should it occur.

The second framework was based on the work of Watson (1997) and Gal (2000) in relation to the expectations for statistical literacy of students when they leave school to participate in society. Watson suggested a three-tiered framework for statistical literacy including (i) the understanding of basic statistical terminology, (ii) the understanding of terminology when it appears in social contexts, and (iii) the ability to question claims that are made in context without proper statistical justification. This framework, in conjunction with the cognitive hierarchy, was employed for example to describe student understanding of sampling (Watson & Moritz, 2000b) and, within each tier, understanding was displayed in terms of (i) more sophisticated definitions of sample, (ii) greater engagement with samples in social contexts, and (iii) the emerging ability to question inappropriate claims made based on samples in newspaper articles. Gal provided a similar framework that included the element of motivation and the ability to communicate reactions. In the current study motivation was not addressed because the format of the items did not lend itself to any substantive measure of motivation. Overall the tasks used in this study, however, directly addressed terminology, or basic statistical skills required to address issues, associated with the first tier of Watson's framework; or were grounded in school-based curriculum or social contexts, providing the opportunity for students to demonstrate their understanding, consistent with the second tier of Watson's framework; or were based on articles

from the media, which gave opportunity for critical questioning of claims, the goal of the third tier of the framework.

In this study, it was felt that the ability to consider variation as well as concepts associated with the components of the chance and data curriculum (Holmes, 1980), and ability to interact with the contexts presented would be significant aspects of the statistical literacy construct, reflecting Gal's (2002) description of statistical literacy. This includes the requisite mathematical terminology and statistical skills appropriate to the task. Previous studies (e.g., Watson et al., 2003; Watson & Moritz, 1998, 1999a, 2000b) suggested that the expectation of a hierarchical ordering of observed responses was reasonable. Gal's (2000) concern about ability to communicate reactions was felt to be handled in a hierarchical fashion by the structural model (Biggs & Collis, 1982) that was used.

3. MEASUREMENT ASPECTS OF THE RESEARCH

Assessment instruments, such as the questionnaires used in this study, are designed to measure a particular attribute or ability. All of the items used in this study, for example, address students' understanding of aspects of chance and data, albeit in different situations. The aim of this current study is to explore the mapping of all of the items onto a hypothesized single underlying variable of statistical literacy. One approach to this problem is to use Rasch modelling techniques (Rasch, 1980).

Rasch models are a set of measurement models coming under the general heading of Item Response Theory (Stocking, 1999) that have been widely used in surveys such as the Third International Mathematics and Science Survey (TIMSS) (Lokan, Ford, & Greenwood, 1997). They use the interaction between persons and items to estimate the probabilities of response of each person to each item. This process produces a set of scores that defines the position of each item and each person against the underlying variable or construct. The unit of measurement is the logit, the logarithm of the odds of success. The specific model used in this study is the Partial Credit Model (Masters, 1982), which allows for items that have a number of hierarchical scoring categories. This model has been shown to be appropriate for use with items that have been coded using hierarchical cognitive taxonomies, such as that of Biggs and Collis (1982, 1991) (e.g., Wilson, 1990, 1992).

Although Rasch methods have been widely used in the area of school mathematics generally (e.g., Lokan, Ford, & Greenwood, 1997; Wilson, 1990, 1992) and in the area of adult literacy including quantitative aspects (Kirsch, 1997), very little research has used Rasch methods specifically in relation to statistical concepts. Izard (1992) considered some of Green's (1982) data on students' responses to probability items and confirmed Green's hypothesised hierarchical structure for an English data set, although with the suggestion of a possible additional level based on data gathered in Quebec, Brazil, and Hungary. Gagatsis, Kyriakides, and Panaoura (2001) used Rasch techniques to suggest levels of understanding in relation to probabilistic concepts found in the Cypriot school curriculum. Description of items suggested a theoretical basis for their construction. Reading (2002) also used Rasch methods to confirm a profile of statistical understanding that had been developed using the cognitive development model of Biggs and Collis (1982, 1991). Her profile included aspects related to data collection, data tabulation and representation, data reduction, and interpretation and inference.

Earlier research involving some of the items to be used in the current study employed the Partial Credit Model (Masters, 1982) on 44 items measuring understanding of chance and data with a particular emphasis on variation (Watson et al., 2003). Based on responses from 746 students in grades 3, 5, 7, and 9, the fit of the items to a unidimensional model of variation was acceptable. Analysis of the variable map, a diagram of the item difficulty and student ability distributions produced by the Quest software (Adams & Khoo, 1996), suggested four levels of understanding associated with appreciation of variation in chance, variation in data, and variation in sampling. All of these studies confirmed specific hierarchies identified through qualitative analysis. The current study differs, however, in that it hypothesises a variable, statistical literacy, which comprises a wide range of statistical understanding and skills. In this sense it is an exploratory rather than a confirmatory study, which aims to postulate the existence of an hypothesized variable, rather than confirm the presence of a construct previously identified by other means.

3.1. TEST EQUATING

Rasch modelling techniques provide a way of linking, or equating, tests through the use of common items or common persons (Bond & Fox, 2001; Kolen, 1999). Generally this is done to ensure that two tests that purport to measure the same construct can be used interchangeably to measure student performance. In this study, however, the techniques were used to bring together questionnaires that were designed to measure both common and varying aspects of a theorised construct of statistical literacy. If it can be shown that when different tests are combined, the items in the various tests work together in a consistent and predictable fashion, that is fit the model, this provides evidence of a single underlying dominant variable and it can be argued that they are likely to be measuring the same construct (Bond & Fox, 2001). Placing all items together on the same scale then provides an opportunity to examine the nature and validity of the underlying theorised construct.

Where different tests are used to measure the same underlying construct, equating is used to align these different measures so that all the items, and the students who attempt them, can be consistently described with the same measures. In addition, in relation to students, their performance on different tests can be mapped onto a single scale. In this study, the different questionnaires are hypothesised to address component parts of a variable termed "statistical literacy". To examine the hypothesis that this is indeed a single, unitary variable, the responses to different items across the questionnaires needed to be linked, or equated, in order to provide a basis for the assertion that these are part of the same construct, and for the interpretation of that construct. The process is described in some detail because the later variable interpretation is dependent on this.

Tests can be equated through Rasch techniques if an anchor or link-set of items can be found or if the same people undertake the questions (Kolen, 1999). The link set of items needs to be common across two or more questionnaires but it does not need to be common to every questionnaire. Linacre (1997) refers to this as connectedness: Providing that all items, regardless of which questionnaire they appear in, are directly or indirectly connected to every other test item in the pool, then all items can be mapped onto a single underlying scale. Links between tests that have some common items enable a "common item equating design".

In this study, 24 common items were used in the 1993, 1995 and 1997 questionnaires. Of these, eight were used across all grades, and all others were used across at least two grades. These items provided a means of linking all the different questionnaire forms from these years across the range of school grades. Within the 2000 questionnaire, 20 items were used across all grades, and, apart from one, all others were presented across at least two grades. Hence the different questionnaire forms within this year were also linked across all grades. In addition, there were four items that were common across all questionnaires, 1993, 1995, 1997, and 2000. These four common items linked the 2000 questionnaire to the earlier questionnaires and provided a means of meeting Linacre's (1997) connectedness criterion. There is no recommended minimum number of items for linking using the Rasch partial credit model, although it is usually suggested that the link items have a range of difficulty (Kolen, 1999). In this instance the range of task-step difficulty was from -1.72 logits to 3.02 logits, providing a wide difficulty span along the variable. Although more common items would provide a stronger link, the link established here appears satisfactory for this initial exploration. Every item from every questionnaire form administered across the years was hence connected directly or indirectly to every other item. The process of test equating for this study is shown diagrammatically in Figure 1.

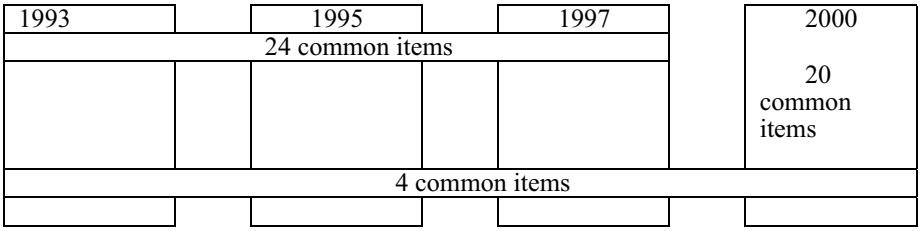


Figure 1. Linking process for equating questionnaires

3.2. VALIDITY OF THE UNDERLYING CONSTRUCT

Validity is the extent to which the inferences drawn from scores on a test, or any other form of assessment, are supported empirically and theoretically. Traditionally, evidence of three kinds of validity was gathered (Messick, 1989). Criterion validity compared the scores on a test with a different measure of the same variable within a short time frame (concurrent validity) or over time (predictive validity) (Anastasi, 1988). Content validity was considered in relation to the possible universe of content that could be included in the test, with the aim of avoiding construct under-representation (Messick, 1989). Construct validity was concerned with the underlying cognitive processes, and Messick (1989) argued, should be considered to be of overarching importance.

Many different sources of evidence may contribute to the establishment of construct validity, but the validity is strongest when the fit of the information gathered from the assessment is explicitly related to an underlying theoretical perspective (Messick, 1989). To establish construct validity two aspects must be considered. First, there should be some underlying substantive theory on which the construct is based. Second, the measuring instrument should address this explicitly – that is, it should be designed to measure a theoretical construct. In this situation, it has been claimed that the Rasch model (Rasch, 1980) is a useful approach to establishing construct validity (Fisher, 1994).

Rasch models make three assumptions. The first of these is that the underpinning construct is unidimensional. Although different domains or categories may be part of the assessment, these all form a single dimension. Second, the variable is hierarchical, or has direction. The construct is measurable with an additive unit of measure that is repeated along the variable; that is, the scoring or coding assigned to each item describes an increasing "quantity" of the construct. The third assumption is that each item is independent of all other items. This means that responses to later items do not depend on a correct response to an earlier item. The extent to which these three assumptions are met is a measure of the validity of the underlying construct (Wright & Masters, 1982).

The purpose of this study is to establish support for the existence of a unidimensional scale that provides interpretable information about hierarchical levels of a hypothesized construct of statistical literacy. In order to do this, the three assumptions must be tested. There are a number of measures provided by Rasch measurement techniques that may be used to ascertain the measurement characteristics of the variable, and test the assumptions outlined above. Two that are widely used are Infit Mean Square (IMSQ) and Item Separation Reliability (R_I). If the measurement characteristics suggest that the assumptions are met, then a substantive interpretation of the construct provides further information about its nature. These processes together provide evidence of construct validity (Wright & Masters, 1982).

The Infit Mean Square is a measure of fit to the model and has an expected value of 1. In practice, values within the range 0.77 to 1.3 provide acceptable fit to the model (Adams & Khoo, 1996; Keeves & Alagumalai, 1999). There are two ways in which items can misfit. The first is "overfit" – the items discriminate too sharply. A group of items behaving in this manner is indicative of a different construct being measured; that is, the scale is not measuring a single construct, and it thus violates the first assumption. A second reason for overfit is that there is dependence among the items, violating the third assumption. This was of particular interest in this study as many of the items shared a stimulus. These were, however, constructed as "superitems" (Collis, Romberg, & Jurdak, 1986; Cureton, 1965) in which each question asked, although referring to the same stem, did not depend on a correct response to a previous question. It was necessary, however, to check this aspect. Finally, items may "underfit" – behave in a random fashion. This occurs if responses are random, or raters apply the scoring scale inconsistently. In the situation in this study, the items had been scored by a small number of expert raters so that inconsistency would be expected to be minimal. Consistent misfit could, however, also indicate that the second assumption of an additive unit of measure is not met.

The Item Separation Reliability measures the extent to which the items are separated along the scale. Items that all cluster closely together do not provide sufficient information about the direction and meaning of the variable, the second assumption. This measure indicates how well the items are

separated in difficulty and can thus provide information about a range of development along the variable (Wright & Masters, 1982).

Once a scale has been produced, it is necessary to interpret the underlying construct. This is a process of criterion referencing (Glaser, 1963, 1981). A criterion is the point at which a student moves from one level of competence to another and is sometimes called the step or threshold. By considering the common demands made on knowledge and understanding by items that appear near each other on the scale produced by Rasch analysis, required levels of performance at particular points along the variable can be identified. The thresholds or steps at which pertinent aspects of competence evolve from one level to the next can then be defined. Since the variable describes a continuum of competence, determining the thresholds becomes a process of judgement that must be justified and understood in terms of the relevance of the criteria to the underlying variable (Eisner, 1993; Wright & Masters, 1982). This provides a conceptual interpretation of the variable that can be compared with the theorised construct. This process of evaluating the fit of the information obtained from a test, or, in this instance questionnaires, to the theoretical rationale for the interpretation of the test outcomes, or scores, establishes strong construct validity (Messick, 1989).

4. METHODOLOGY

4.1. SAMPLE

The data used were collected from questionnaires completed by 3852 students in Tasmania, conducted in 1993, 1995, 1997, or 2000. Although it could be argued that the teaching of chance and data within the state could have changed over this period, there were indications (e.g., Watson & Moritz, 1998) that there was no change at least for some items between 1993 and 1997.

As this was a preliminary study to explore the nature of the underlying variable, it was the individual understanding shown on items that was of interest, not a comparison among students. As such the research design was cross-sectional rather than longitudinal, and this initial study did not attempt to map the rate of students' progress along the variable from year to year. Rather it was intended to establish the nature and validity of the underlying construct of statistical literacy. Although some students were involved in longitudinal data collection, only data from initial questionnaires were used in this study. The sample distribution across grades is given in Table 1.

Table 1. Sample Size for Each Grade

Grade	Frequency	Percent
3	1039	27.0
5	421	10.9
6	881	22.9
7	239	6.2
8	207	5.3
9	1065	27.6

The uneven distribution reflects the target grades of the previous questionnaires. Overall, however, the sample covers a wide range of ages and school curriculum experience. The number of students in each grade answering each item is given in Appendix B.

4.2. ITEMS/TASKS

The items used in this study were devised to measure various components of the chance and data curriculum as it was introduced to Australian schools in the 1990s. They reflect specific straightforward aspects of content, such as sampling, average, chance, and graphs. They also reflect the contexts within which the subject matter is applied, for example, school-curriculum based social

contexts, such as conducting a survey in school, and less familiar media-based contexts specifically related to statistical literacy. Issues of acknowledging variation, drawing inferences, and questioning claims were interwoven with the content and context for some of the items. The items hence covered the range of potential contributing elements to the construct of statistical literacy as outlined earlier.

Altogether 80 items were used, 48 of which were used in the previously noted study of Watson et al. (2003). There were 44 items reported in Watson et al. (2003), four of which appeared in two forms with different numbers of imagined trials used for different grades (see e.g., SP2A, SP2B in Appendices B and C). Thirty-six of the items were used in data collections that took place in 1993, 1995, and 1997. The items are detailed in the various studies that used them to explore particular topics in the chance and data curriculum. They included curriculum content-based items, such as items addressing basic probabilities, basic table reading, variation in a spinner scenario, understanding of stacked dot plots, and interpretation of a pictograph. Other items were closely related to the school curriculum but placed in some type of social setting, for example, items on conditional and conjunction probabilities, on conducting a survey of children in a school, on risk from taking a medicine, on average number of children in a family, on the median (or average) of a set of science measurements, on selecting a new car, on actors' performances, and on the story conveyed by a stacked dot plot.

These items were adapted from items used in previous research (Fischbein & Gazit, 1984; Garfield, 2003; Green 1982, 1983a; Jacobs, 1999; Konold & Higgins, 2002; Konold & Garfield, 1992; Nisbett, Krantz, Jepson, & Kunda, 1983; Pollatsek, Well, Konold, Hardiman, & Cobb, 1987; Tversky & Kahneman, 1983; Watson, Collis, & Moritz, 1994, 1997), or developed as part of previous local research (Torok, 2000; Watson, 1998a; Watson et al., 2003). Other items were based on contexts taken directly from newspaper extracts, such as the "average" house price, biased sampling of populations, odds and independent binomial events in sporting contexts, interpretation of a misleading picto-bar graph, interpretation of an incorrect pie graph, ranking chance language used in headlines, creating a graph to represent a cause-effect claim and questioning the claim, and interpreting conditional language about the effects of smoking (Moritz & Watson, 1997, 2000; Moritz, Watson, & Collis, 1996; Watson, 1998b, 2000; Watson & Moritz, 2003). Finally the scale included questions on the definitions of the terms "average", "sample", "random", and "variation" (Mortiz, Watson, & Pereira-Mendoza, 1996; Watson et al., 2003). All items, codes, and example responses are given in Appendix C. The codes reflect the theoretical frameworks introduced earlier (Biggs & Collis, 1982, 1991; Watson, 1997).

For clarity, in this study the term "task" is used to refer to the total item as it was presented to the students, and "task-step", to refer to the level of response denoted by the coding. Thus for an item that asked about guns in United States high schools, M7CH is the task, and the levels of response, coded as 0–4, are the task-steps, shown as M7CH.1 to M7CH.4, that appear on the variable map in Appendix A.

4.3. ANALYSIS

The process for calibrating and equating the questionnaires, undertaken with Quest software (Adams & Khoo, 1996) is summarized as follows:

1. The difficulty levels of four items that had the highest number of respondents, SMP4, DIE7, HAT8, and BOX9, were calibrated using a data set that included no missing data. The purpose of this step was to provide a stable set of difficulties as an "anchor set" against which all other difficulties could be estimated.
2. Using the difficulty levels from these items as an anchor, and the full data set of all 80 items and 3852 students, the different questionnaire forms were calibrated and equated in one operation (Adams & Khoo, 1996). The four items common across all questionnaires (DIE7, BOX9, RAN3, and M4DR) linked the earlier questionnaires to the 2000 questionnaire.
3. Difficulty levels and fit statistics of all steps on all items were obtained and written into a file that was used for subsequent analyses.

This procedure produced a variable map, showing the task-steps and persons distributed on a single scale, a fit map of the items to the model, and a classical item analysis. Common characteristics of task content and skill were identified by undertaking an audit of the demands of the task-steps that were close together on the map. It was then possible to describe trends in the development of concepts and cluster these together with descriptions across the topics covered in the questionnaires. This process is one of professional judgment and discussion similar to procedures suggested by Miles and Huberman (1994). Six hierarchical levels were identified as a convenient way of distinguishing overall steps in the progress along the variable for the underlying construct of statistical literacy. These levels are identified on the map in Appendix A.

5. RESULTS

5.1. MEASUREMENT CHARACTERISTICS OF THE VARIABLE

Table 2 provides estimates of Item and Case Separation Reliabilities and overall fit measures for the 80-item instrument. These separation reliabilities describe how adequately the tasks describe the underlying variable, and the extent to which the student cohort is spread out across the continuum (Wright & Masters, 1982). The Item Separation Reliability is high, suggesting that the tasks do indeed describe a spread of difficulty along the variable that will allow the underlying construct to be described. The overall fit to the model is also high, at the expected value of 1.00, suggesting that the tasks form a hierarchical unidimensional scale. Similar comments apply to the Case Separation Reliability and fit statistics, suggesting that the tasks used are appropriate for this cohort of students. The Cronbach Alpha is a measure of internal consistency taken from classical psychometric theory, and is also obtained from Quest analyses. This value is high at 0.85, showing that the scale meets not only Rasch measurement standards but also a classical standard of reliability.

Table 2. Reliability and Fit Indices for 80 Items

Item Separation Reliability	0.99
Item Infit Mean Square	1.00 (S.D. 0.13)
Case Separation Reliability	0.86
Case Infit Mean Square	1.00 (S.D. 0.42)
Cronbach Alpha	0.85

Appendix A contains the variable map of student ability and item difficulty. The measurement scale, from –4.0 to +4.0, on the left hand side is in logits, the logarithm of the odds of success. The crosses on the left hand side each represent 12 students and the tasks are presented on the right hand side. This map shows the relationship between the tasks and the students who undertook them. The representation is limited by space and printing requirements, and thus students are grouped together, in this case in groups of 12, according to their estimates of ability. Where there were fewer than 12 students at any particular ability estimate, no cross is shown. The map does not suggest that where there are no crosses there were no students – only that there were fewer than 12 students at any particular level of ability (in logits) to be grouped together and thus they do not appear in this representation. Each task having more than one coding level is described in terms of m.n, where m is a 4-letter-digit task identifier and n represents the coding level or task-step. Thus M2PI.2 is media task 2, Pie chart, coding level 2. Task identifiers, along with a statement of the question, information on codes, and typical responses, are found in Appendix C. Although the variable map in Appendix A presents both item and student data, the analysis presented here focuses on the right hand side of the map, which contains the item distribution along the variable. Only two tasks fell outside the acceptable limits of fit, both showing underfit. The first of these, AV12, required students to select from a list of alternatives to explain the meaning of “2.2 children per family”. Randomness in this situation is likely to be because students were guessing. The second task showing misfit, DIE2, asked students to predict the outcomes of 60 tosses of a single die and explain their answers. The most likely

reason for underfit here is inconsistent coding patterns. Detailed examination of individual item statistics confirmed the high level of item fit for all tasks except the two problematic ones, and measurement errors were very small for each task at each step of difficulty (which corresponded to a change in coding). In general, tasks fitted the model well, providing a basis for interpretation of the scale. In particular, there was no evidence of overfit, as would be expected for a multi-dimensional construct.

5.2. QUALITATIVE INTERPRETATION OF THE UNDERLYING VARIABLE

Having provided evidence suggesting that the variable was unidimensional, the next step was to interpret it in relation to the tasks that were used to measure it. Six levels were distinguished along the continuum, shown by horizontal lines in Appendix A, indicating clusters of task-steps that when analysed by content (e.g., mathematical skills, statistical concepts, context) suggested common characteristics in a hierarchical sequence of statistical literacy. In some instances, two task-steps appeared within the same level for the same task (e.g., TBL5.2 and TBL5.3 both appear in Level 3 of the continuum). Usual measurement practice would be to recode these and collapse the two categories into one. Here, however, we have chosen to maintain the original coding because the qualitative interpretation suggested that students were responding in different ways. The implications of this are addressed further in the Discussion section. In Appendix C brief descriptors of the code levels are shown and these are amplified in the following sub-sections with examples of content and context. Summary descriptions of the characteristics of statistical literacy displayed in the tasks at the six levels are presented in Table 3.

Table 3. Statistical Literacy Construct

Level	Brief characterization of step levels of tasks
6. Critical Mathematical	Task-steps at this level demand 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	Task-steps require 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	Task-steps require 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	Task-steps at this level, often in supportive formats, expect selective engagement with context, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.
2. Informal	Task-steps require 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	Task-steps at this level suggest idiosyncratic engagement with context, tautological use of terminology, and basic mathematical skills associated with one-to-one counting and reading cell values in tables.

5.3. LEVEL 1 – IDIOSYNCRATIC

Task-steps associated with concepts such as average or with the definition of terminology do not appear at Level 1, indicating lack of engagement with their associated ideas and contexts. At this level personal beliefs and experience dominate, for example, with the task-step to identify unusual features of an article about guns in high schools in the United States satisfied by "people should not have guns" (M7CH.1) and a task-step on new car selection satisfied by the alternative "rely on friends" (SM19.1). In terms of data representation in straightforward contexts, task-steps demand only reading

specific values from a simple two-way un-nested table (Q10A.2, TBL1), choosing the highest value from a row or column in a table (Q10C.2, TBL2), and determining a frequency and a difference from a pictograph (TRV1, TRV2).

Chance task-steps at this level suggest idiosyncratic beliefs, for example in drawing names from a hat of boys and girls (HAT8.1), "a girl because the teacher is a girl"; in describing odds from a newspaper article on a sporting event (M3OD.1), "it is the current score in the game"; in describing the chance of a 50-50 spinner landing on a certain half (SP1.1), "bad chance" or "1 in 10". For a task to interpret "a 15% chance of getting a rash" (CH11.1), colloquial alternatives "good chance" or "hardly any chance" appear at this level.

The only task-steps associated with inference appearing at Level 1 suggest concern for certainty rather than uncertainty (TRV5.1), refusal to predict or belief in "no change" (TRV6.1). Similarly for tasks related to variation, only those task-steps which require basic acknowledgement of change, as in a travel-to-school graph "won't look the same every day" (TRV3), or idiosyncratic predictions of chance outcomes without justification (SP2A.1, SP2B.1, DIE2.1), appear at this level.

5.4. LEVEL 2 – INFORMAL

Although the task-steps appearing at Level 2 demand engagement with more contexts, the engagement is still intuitive, non-statistical, or reflective of irrelevant aspects of the task context. Some task-steps require single ideas, for example in terminology associated with sampling (SMP3.1, SMP4.1) and average (AVG2.2, ME13.1). Other single aspects of sample are encompassed in a survey planning task in a school-based context, where the task-step requires features such as "ask 400", "ask everyone", or "ask the people I meet", without considering the need to represent the population (MVE1.1). For a task on new car selection (SM19.2), the alternative "it doesn't matter whether a person uses friends' advice or data from a consumers' report" appears at this level.

At Level 2, table reading task-steps demand comparing cells to determine the highest or most even counts (Q10D.2, Q10E.2) and finding a total greater than 100 (TBL4). Graphing task-steps require identifying the smallest data value in a stacked dot plot (SP6) but only idiosyncratic arithmetic strategies for working out prices from a picto-bar graph from the media (M9C.1, M9D.1). In relation to chance, a simple 50-50 spinner task (SP1.2) demands a correct response equivalent to a half, but no recognition of variation. "Anything can happen" is a justification accepted for task-steps associated with picking a boy's or girl's name from a hat (HAT8.2) or with comparing two boxes of marbles for the chance of choosing a single marble of a particular color (BOX9.2). In a media task of ordering chance headlines by likelihood, the step at this level only requires placement of phrases in the appropriate half of the 0-1 number line (M1CH.1).

In terms of inference at this level, task-steps are satisfied by story-telling (TRV6.2) or pattern recognition (TRV4.1) in predicting from a pictograph. A task to judge the better of two stacked dot plots for telling a story of how long families in a class have lived in a town (TWN3.1) accepts the inappropriate choice with reasoning such as "it is well set out". Task-steps associated with variation require only appropriate "surprising" results for repeated spins of a 50-50 spinner (SP4B); too much, too little, or lop-sided predictions for repeated spinner trials (SP5A.1, SP5B.1); and patterns or strict chance in predicting 60 die outcomes (DIE2.2).

5.5. LEVEL 3 – INCONSISTENT

At this level task-steps require more engagement with context than at the previous two levels but this is dependent to some extent on the format of items, which may provide added support. Although more features are demanded, the statistical ideas required are qualitative rather than quantitative and appropriate conclusions may not be accompanied by suitable justifications.

In relation to sampling, the task-steps for tasks associated with judging plans for a school survey require suggestions in context but focusing on peripheral rather than salient features, for example a method is good because it is "easy" (MVE5.1, MVE6.1), "not too many" (MVE3.1), or "large"

(MVE4.1); or a method is bad because "more people are needed" (MVE2.1), "the wrong people might get picked" (MVE2.1), or "they're a bit young" (MVE4.2). For a task asking for methods of selecting four students to lead a parade, a representation say of boys and girls (TBL5.2) or a random method (TBL5.3) is acceptable at this level. A task-step for commenting on a voluntary poll about legalizing marijuana only requires recognition that people could be lying or the sample size is too large (M4DR.1).

For data representation the task-steps at this level demand at least one summary statement when interpreting stacked dot plots (TWN1.2, TWN2.2), a basic unlabelled graph or a labeled graph with no association when association is intended (M8GR.1), or recognition of non-salient unusual features of a media bar graph (BT1A.1). Average task-steps require colloquial expressions in an open-ended format (AVG1.1, M5AV.1).

Chance task-steps at this level generally demand qualitative rather than quantitative reasoning. Although a simple 50-50 spinner task with repeated spins requires an answer equivalent to "half" of the spins (SP2A.2), a task about drawing names from a hat requires recognition of "more girls' names in the hat" (HAT8.3), and a task about equality of dice outcomes only requires a justification of "anything can happen" (DIE7.2). A task-step supported by a selection of alternatives to interpret a "15% chance of getting a rash" demands an exact numerical interpretation (CH11.2). The language associated with ordering conjunction events appropriately is needed for two tasks at this level (CF15 and CP18), whereas demands for quantifying outcomes from four coin tosses at a sporting match are less stringent, values greater than a half being acceptable (M10A.1). A task-step for defining "random" requires single or multiple elements (RAN3.1, RAN3.2).

Few task-steps associated with inference or variation appear at Level 3. Although the appropriate choice is required in determining which of two stacked dot plots is better at telling a data-based story, justification for the choice is not needed (TWN3.2). Only recognition of chance, not variation, is required in predicting repeated outcomes with a 50-50 spinner (SP2A.2) and only a single aspect is demanded in defining the term variation (VAR.1). Improvement required in quantitative skills at this level is associated with task-steps requiring recognition rather than creation of appropriate responses (e.g., AV12.2, CH11.2).

5.6. LEVEL 4 – CONSISTENT NON-CRITICAL

The task-steps appearing at Level 4 demand a consolidation of appropriate contextual but non-critical engagement by students in various contexts. In terms of the Statistical Literacy Hierarchy discussed in Section 2.2, the task-steps require an understanding of social contexts that are not associated with critical questioning or partial context-only reasoning where critical thinking is the ultimate aim.

For the definition of sample, two aspects are required, such as "you have a small piece of something" (SMP3.2). The task-step associated with suggestions for surveying a school demands representative but not random methods (MVE1.2). Task-steps for evaluating other survey methods require peripheral or partial recognition of salient features associated with appropriate "good" or "bad" judgements (MVE2.2, MVE7.2, MVE4.3). The media task based on the less familiar context of a non-representative sample of United States high schools demands only contextual recognition, such as that people could be lying or the whole United States would be the same (M7CH.2).

Graph recognition task-steps demand the highest data value (SP7), the range of the data (SP8), a qualitative description of the shape (SP10), and appropriate reasoning for selection of the scale as the better of two stacked dot plots (TWN3.3). For media-based graph tasks, however, only partial recognition or representation is required, for example in criticizing a pie chart summing to 128.8% (M2PI.1) or graphing an association of heart deaths and car usage (M8GR.2). Average task-steps at this level require describing the mean or middle appropriately (AVG2.3, M5AV.2), and finding the mean of a small data set (AVG1.2), without recognition of the effect of an outlier.

Chance task-steps present a variety of contexts and demands at Level 4. A task to select which of two boxes with the same ratio but different numbers of marbles is more likely to produce a certain outcome demands appropriate proportional reasoning (BOX9.3), whereas one to justify belief about

die outcomes requires only "same" chance reasoning (DIE7.3). Probability tasks set in a media context require a correct response for a single coin toss (M10B.2) but the same answer for four tosses (M10A.2), indicating lack of knowledge of compound events. An odds task-step accepts predicted scores (M3OD.2). Where language rather than numerical calculation is involved, task-steps demand appropriate ordering of chance newspaper headlines on a number line (M1CH.2) and appropriate interpretation of straightforward conditional statements (M6AB.2, M6D.1).

Except for the task of distinguishing between the appropriateness of two stacked dot plots (TWN3.3), the task-steps for inference at this level require limited recognition of the implications of representations, for example balancing information presented in terms of boys and girls in a pictograph (TRV4.2, TRV6.3) or reflecting a majority (TRV6.4). A media task-step for a suspicious cause-effect relationship demands only engagement with the context and questioning of data collection rather than questioning of the association (M8QU.1).

Task-steps dealing with variation in chance settings that appear at Level 4 demand a reason associated with variance in explaining differences in repeated sets of trials with a 50-50 spinner (SP3A.3, SP3B.3), and realistic variation in numerical predictions of outcomes for six sets of repeated trials (SP5B.2) and of outcomes for 60 tosses of a die (DIE2.3). The task-step for deciding the authenticity of sets of spinner trials requires both appropriate choices and reasoning (SP11.2). For media tasks, however, task-steps demand less sophisticated reasoning, with a media bar graph interpretation task requiring focus on single columns rather than comparisons across columns (BT1B.1) and the definition task requiring multiple relevant features, such as "variation means to change something" and "the weather is going to vary over the next few days" (VAR.2).

The task-steps at this level that demand consolidation of the mathematical and statistical skills include those associated with the mean, simple probabilities, and graph characteristics, all in straightforward settings. Task-steps require appreciation of setting but rarely critical questioning.

5.7. LEVEL 5 – CRITICAL

Task-steps at the top two levels of the statistical literacy construct demand similar critical thinking skills associated with the third tier goal of the Statistical Literacy Hierarchy. What distinguishes them is the level of mathematical skill required to engage in critical questioning. At Level 5 sophisticated use of proportional reasoning is not required, but in contexts, particularly familiar ones, critical thinking is otherwise expected, as in appropriate use of terminology, appreciation of variation, and qualitative interpretation of chance.

The task-steps related to defining a sample require the relating of several elements in describing a sample and its purpose (SMP3.3, SMP4.3). The task on surveying a school demands random methods or random methods combined with representation, such as "10 from each grade, 5 boys and 5 girls picked at random" (MVE1.3). For task-steps to evaluate three other suggested surveying methods – a random method, a choice of friends, and a booth for volunteers – appropriate decisions and statistical justifications are required (MVE2.3, MVE5.3, MVE6.3). For the task of selecting a car, the appropriate task-step of using the report on 800 cases is needed (SM19.3). For the task of assessing a voluntary poll on legalizing marijuana, focusing on the central issues, for example, the type of listeners to the radio or that only motivated people telephone the station, is required (M4DR.2). The task-step for assessing an article about access to guns by school students in the United States, however, only requires recognition of the non-representative nature of the sample with the support of an additional question about other regions of the United States (M7CH.3).

In terms of graphing at this level, task-steps require appropriate representations for a claim about the association of heart deaths and car usage (M8GR.3), representing the ability to handle two variables at the same time and show corresponding increases, or recognition of the error in a pie graph that sums to 128.8%, focusing on the total percent or the shapes of the segments of the graph in comparison to the percents they represent (M2PI.2). At this level, for the idea of average, there is the demand to find the median or mean of a small data set (ME13.3).

Chance task-steps at this level demand a consolidation of ordered estimates of conditional statements (CP16) and of giving appropriate "if ... then ..." statements for an embedded conditional

statement in a newspaper article on smoking and wrinkles (M6D.2). In media contexts with mathematical skills required, however, task-steps require qualitative rather than quantitative recognition (M10A.3) or use of ratio without appropriate interpretation (M3OD.3). Few tasks on inference appear at Level 5, with a task-step for selecting actors by audition who later perform less well than expected demanding the choice of an alternative reflecting regression to the mean (Q20).

Two task-steps require appreciation of variation at Level 5. For a task predicting the outcomes of spinning a 50-50 spinner repeatedly, responses must spontaneously use words like "about" or "probably" in suggesting numbers of successes or phrases like "it will be close to half" (SP2A.3, SP2B.3). For a task to describe unusual features of bar graphs in a report on boating deaths, an increase or change in the data over time or acknowledgment of variation explicitly in the visual appearance of the graphs is required (BT1B.3).

5.8. LEVEL 6 – CRITICAL MATHEMATICAL

As noted previously proportional reasoning skills are demanded by many of the task-steps that appear at Level 6, particularly in chance or media contexts. As well task-steps require sensitivity to the need for uncertainty in making predictions and appreciation of subtle aspects of the language for some tasks.

In relation to sampling, detection of the two flaws in a survey method suggesting 10 students from a computer club, for example, "there are not enough people and they are selectively picked," is required (MVE3.3). The task-step concerning a sample from Chicago in relation to the United States (M7CH.4) requires the recognition of the non-representative nature of the sample, without any support. A task to suggest two methods to select children to lead a parade demands either two different random methods or a combination of random and representative methods (TBL5.4).

In terms of graphing, two summary statements involving the context, rather than just data reading, are required to describe stacked dot plots about how long families have lived in a town (TWN1.3, TWN2.3). The mode must be recognized in relation to a stacked dot plot (SP9). Finding errors in bar graphs about boating deaths is required (BT1A.2), as are appropriate rate calculations associated with a complex picto-bar graph (M9C.2, M9D.2). Recognition of outliers is demanded when calculating a mean (AVG1.3) and suggesting the median as the appropriate measure of middle in relation to house prices in the context of a newspaper article (M5AV.3).

At the highest level of statistical literacy task-steps require quantitative reasoning for chance tasks. For straightforward task-steps such as those involving outcomes for a single die and drawing names from a hat, numerical (e.g., fractions) rather than qualitative descriptions are demanded (DIE7.4, HAT8.4). For a classic fish-tagging task, proportional reasoning to obtain the solution of "2000" is required (Q17). For a task from the media on explaining odds, proportional reasoning and the correct direction for interpreting the result appear in the response (M3OD.4). For a task based on an article on tossing coins at the start of a cricket match, independence and correct calculations are required (M10A.4). Integrated descriptions for the term random are also demanded (RAN3.3).

Task-steps related to inference at this level reveal subtleties in thinking. Task-steps requiring predictions for a pictograph on how children travel to school, for example, demand inclusion of expressions of uncertainty, such as "probably a [new] girl comes by car – more girls get a car" (TRV4.3). A task concerning a newspaper article about heart deaths and car usage requires responses that ask the salient question about a cause-effect relationship (M8QU.2).

The mathematical/statistical skills demanded by task-steps at the highest level include proportional reasoning associated with ratio and appropriate part-whole interpretations, the ability to use rates in calculating costs, understanding of independence and its implications for calculating probabilities, an overall quantitative view of chance as probability, and a memory for terms such as "mode". Further some task-steps require an ability to account for subtleties in language and context.

These extended summaries of the levels of statistical literacy based on the tasks employed are intended to portray the detail and richness of the information obtained from the questionnaires. As every task-step code is described with examples in Appendix C, it is further possible to link task

demands for every task-step displayed in Appendix A. At the other extreme, the summaries in Table 3 and the level labels in Appendix A are intended to provide brief indications of the differences among levels.

6. DISCUSSION

Following comments on the limitations of the study, the discussion will focus on five aspects of the outcomes of this research: the identification and exploration of a hierarchical construct of statistical literacy; the relationship of this construct to previous research; the complex nature of the statistical literacy construct, particularly in relation to context; implications for future research; and implications for classroom planning.

6.1. LIMITATIONS OF THE STUDY

The data used in this study were collected as parts of other research studies into student understanding of the chance and data curriculum over an eight-year period (Watson, 1994; Watson et al., 2003). Although initial indications over the first four years were that curriculum implementation produced no improvement on average performance on many items (e.g., Watson & Moritz, 1998, 1999a, 2000b, 2003), there is no corresponding analysis for items used again in the final year covered by this study. The purpose of this study, however, was to document the hierarchical nature of the statistical literacy construct using all available data from the studies, not to consider changes across cohorts, years, or individuals. Longitudinal data, for example collected from students in 1995 and 1997, were not included in this study.

Although the data used for this study reflected a wide range of age, ability, and socio-economic status, they do only represent the Australian state of Tasmania. Other cultural settings may result in students responding differently, particularly to context-based items. It is the belief of the authors, that the school students used in this study are likely to have experiences similar to other Australian students and to students in most western countries.

The relatively small number of linking items used in the Rasch analysis means that some caution needs to be exercised in interpreting the results of the analysis. Two other factors help to mitigate this concern. First, there is a large number of responses for these linking items across all years. Second, the structure of the resulting variable map in Appendix A is very similar to the structure of the corresponding variable map found by Watson et al. (2003) for the subset of items used in 2000. The relative placement of common items engenders confidence that the Rasch analysis produced as part of the current study is a reasonable suggestion of the hierarchical nature of the construct.

6.2. IDENTIFICATION AND VALIDATION OF A HIERARCHICAL CONSTRUCT OF STATISTICAL LITERACY

The findings from the application of the Rasch model suggest a unidimensional character of the variable. Fit to the model was excellent overall, and individual items also showed no overfit, which might have been expected if a multi-dimensional construct was being addressed.

The scale established from the 80 items had a high Item Separation Reliability ($R_I = 0.99$) and provided sufficient information to give a criterion-based hierarchical profile of the underlying construct, hypothesised as statistical literacy. The large item pool provided considerable detail about the variable without over-sampling particular concepts, and confirmed that mathematical skills and understanding of contexts, as well as content from the school curriculum, were all aspects of the same construct. These aspects are summarized in Table 3 and further discussed below.

This good model fit, together with the coherence of the interpretation of the underlying variable with the hypothesised construct of statistical literacy discussed earlier in this paper, suggests strong construct validity (Messick, 1989). It seems that the questionnaires that targeted varying aspects of the

chance and data curriculum can, when combined, provide useful and interpretable information about hierarchical levels of statistical literacy.

6.3. RELATIONSHIP TO PREVIOUS RESEARCH

In comparing the results of the current study with the previous analysis of Watson et al. (2003) that focused on the construct of variation, several points can be made. In the earlier analysis a subset of the current 80 items was used and the particular focus on variation led to the identification of four levels of the underlying variation variable rather than six as in the current study. The spread of items along the variable was similar, however, with the top 11 items in the variation map appearing at Level 6 in this study, and the bottom 6 items appearing at Level 1. This, together with the good fit to the model, indicated that variation is a sub-domain of statistical literacy appearing across difficulty levels. The greater number of items used in the current study, particularly reflecting more curriculum-based chance tasks and more media-based social contexts, gave greater opportunity to distinguish characteristics of increasingly sophisticated performance. This allowed for the more detailed and complex description than earlier.

There is a close relationship of the characteristics associated with levels of statistical literacy and Watson's (1997) three-tiered framework of statistical literacy. The mathematical and statistical skills noted at the different levels reflect the terminology of statistical ideas and its usage, which are suggested as essential in Tier 1 of the framework. The engagement with the context of statistical inquiry reflects Tiers 2 and 3 of the statistical hierarchy. Applying terminology in interpreting a context, which is the goal of Tier 2, appears from Level 3 in this profile, and thinking critically to question inappropriate claims and methods, the goal of Tier 3, appears from Level 5 onwards. The use of open-ended tasks that allowed for the identification of bias or errors in subtle settings gave students the opportunity to display these understandings at increasingly higher levels of the construct. In particular these tasks reflect the transition to the needs of adults in society as users of statistical information that were recognized by Wallman (1993) and Gal (2002). The written nature of the questionnaire further satisfies at least one dimension of Gal's requirement to communicate reactions to statistical information.

6.4. THE COMPLEX NATURE OF STATISTICAL LITERACY

The title of this paper reflects our view that statistical literacy is indeed a complex construct. Interpretation of the variable suggests that it encompasses all individual components of the chance and data curriculum (AEC, 1991, 1994), as well as the foundational aspect of variation (Wild & Pfannkuch, 1999). Beyond these characteristics is the realisation of the importance of engagement with context in defining the underlying construct for statistical literacy. The emergence of context was a distinguishing feature in the higher levels of the construct. The interaction of mathematical skills from the curriculum with the increasingly subtle contexts involving statistical bias or misinformation, creates situations that only students at the highest ability level can interpret successfully. In saying this we realise it is important to recognise that this is as much related to the opportunity to learn as it is to innate ability. The reasoning associated with the application of high-level mathematical skills in a subtle social context is unlikely to emerge through happenstance.

By including many tasks embedded in social settings that require interpretation, this study has identified an important factor leading to high achievement in the realms of statistical literacy. Statistical literacy is not just knowing curriculum-based formulas and definitions but integrating these with an understanding of the increasingly sophisticated and often subtle settings within which statistical questions arise. Using a metaphor suggested by Tognolini (1996) it appears that statistical literacy is a complex construct that may be thought of as a thick thread or rope comprising two interwoven and essential strands: mathematical/statistical understanding of the content and engagement with context in exploiting this understanding. In the past, assessment has focussed almost exclusively on curriculum-based mathematical skills. This study suggests that measurement of

statistical literacy is incomplete without the opportunity to engage with genuine social contexts, particularly such as those found in the media items.

6.5. IMPLICATIONS FOR FUTURE RESEARCH

Although the unistructural, multistructural, and relational aspects of responses to individual items (Biggs & Collis, 1982) could be identified in many cases to aid in developing hierarchical codings, the combination of mathematical skills and engagement with context provided the opportunity to describe six rather than three levels of the overall statistical literacy construct. In terms of the earlier work of Campbell, Watson, and Collis (1992) on students' understanding of volume measurement, the categorisation of levels depends to some extent on the strength of the microscope used to view the phenomenon. Viewing the statistical literacy construct from "afar", it is possible to speculate on the existence of two unistructural-multistructural-relational cycles, similar to those identified by Campbell et al. and for beginning inference by Watson and Moritz (1999b). At Levels 1, 2, and 3, success on items reflects the increasingly structured use of data and information in a highly organised task environment. At Levels 4, 5, and 6, open-ended tasks and less familiar settings provide contexts where success is associated with using more complex mathematical skills and engagement within increasingly complex settings. These levels progressively appear to reflect simple single classroom settings (like using dice), multiple aspects of settings (such as surveys within the school environment), and complex relational settings (such as finding bias in unfamiliar social settings presented in the media). More research, however, is required to provide convincing evidence of this hypothesis.

The appearance of different task-steps for the same task at the same level of difficulty in some instances also provides some insight into students' achievement. It suggests that higher levels of sophistication in thinking are not always related to higher ability. Rather, students appear to be drawing on different ways of conceptualizing the question, and thus, in some situations, students at the same ability level have two ways of responding to a particular question. Improved identification of different conceptual frameworks could provide useful information to teachers about appropriate interventions for students at the same level of understanding.

Another step in future research is to analyse longitudinal data on individuals to explore the hypothesis that indeed the hierarchical structure observed in this study represents a developmental sequence that could be expected to be observed over the years of schooling. Several studies based on subsets of items included in this study suggest that such a hypothesis is reasonable (Watson & Moritz, 1998, 1999a, 2000b).

Although the characterization of the underlying statistical literacy construct appears sound based on the items used and the data collected in this study, using all of the 80 items would be impractical for an instrument to establish statistical literacy standards or benchmarks in the classroom. Items providing redundant data could be eliminated from any new instrument designed to assess statistical literacy. The choice of which items to leave out is dependent on the test writers but any new questionnaire or test purporting to measure statistical literacy should have a test specification that includes items that address both mathematical skills and contextually based application of these. This is consistent both with the conceptualisation of statistical literacy and the findings of this study.

Some gaps in the content covered with respect to topics in the curriculum also emerged when the overall scale was considered. For example, there were few difficult items relating to tables or to more complex graph types, such as those with non-linear association. New items will be needed to attend to this. The next stage of the research will address the preparation of an improved test for the construction of such a scale and its trialling in schools. Associated with this will be the identification of expectations for particular grade levels within the overall hierarchy of the statistical literacy construct. The current research has provided a foundation for future work, and confidence that statistical literacy is a single hierarchical construct that can be measured as students progress through school.

6.6. IMPLICATIONS FOR CLASSROOM PLANNING

Even before a shorter instrument is developed, recommendations for curriculum planners and teachers can be made based on the observed statistical literacy construct. We feel there needs to be more use of context, particularly socially-based media examples, in teaching statistics, both within mathematics and in other curriculum areas. We would support a concerted effort to devise activities specifically to assist students to move from non-context based application of statistical skills, such as "add them up and divide" interpretations of average, to an appreciation of context, and then to an awareness of its importance in decision making, including developing the skills to identify bias and misrepresentation. Some of these activities could be based on media items and interview protocols used in research (e.g., Watson & Moritz, 1999b, 2000a). Explicit discussion of the interwoven nature of the two strands of statistical literacy may help students appreciate its importance. We feel curriculum planners need to develop materials that enhance mathematical and statistical skills at the same time as the qualitative understanding of statistical reasoning.

It should be noted that we are not suggesting that teachers should neglect developing the separate underlying concepts, such as average, chance, variation, or sampling. Nor are we suggesting that improving a student's understanding of one of these concepts will also improve understanding of another, different idea. Rather, this research indicates not only that the underlying ideas are important, but also that students need to have opportunities to address these ideas in a range of contexts, including non-school-based ones. This requires a balance of concept and skill development and application of the ideas in authentic situations, and makes increased demands on teachers and curriculum planners.

In the light of changed curriculum expectations (e.g., Education Queensland, 2000) and extended social expectations for quantitative literacy generally (e.g., Steen, 2001), we believe that teachers across the curriculum will also have increased expectations placed on them in terms of appreciating statistical literacy and how to develop it. It is likely that professional development for teachers will be needed if they are to assist their students to achieve the highest levels of statistical literacy observed here before they leave formal schooling.

ACKNOWLEDGMENTS

This research was funded by the Institutional Research Grant Scheme at the University of Tasmania in 2001. At the time of the initial research Rosemary Callingham was employed by the University of Tasmania. The authors thank Dr. John Izard for helpful comments during the initial preparation of this paper and the anonymous referees for suggestions for necessary revisions.

REFERENCES

Adams, R. J., & Khoo, S. T. (1996). *Quest: Interactive item analysis system. Version 2.1* [Computer software]. Melbourne: Australian Council for Educational Research.

Anastasi, A. (1988). *Psychological testing*. Macmillan: New York.

Australian Association of Mathematics Teachers (AAMT) (1997). *Numeracy = everyone's business. Report of the Numeracy Education Strategy Development Conference. May 1997*. Adelaide: Author.

Australian Education Council (1991). *A national statement on mathematics for Australian schools.* Carlton, Vic.: Author.

Australian Education Council (1994). *Mathematics - A curriculum profile for Australian schools.* Carlton, Vic.: Curriculum Corporation.

Batanero, C., Estepa, A., Godino, J. D., & Green, D. R. (1996). Intuitive strategies and preconceptions about association in contingency tables. *Journal for Research in Mathematics Education, 27*, 151-169.

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

Biggs, J. B., & Collis, K. F. (1991). Multimodal learning and the quality of intelligent behaviour. In H. A. H. Rowe (Ed.), *Intelligence: Reconceptualization and measurement* (pp. 57-76). Hillsdale, NJ: Lawrence Erlbaum.

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

Cai, J. (1995). Beyond the computational algorithm: Students' understanding of the arithmetic average concept. In L. Meira & D. Carraher (Eds.), *Proceedings of the 19th Psychology of Mathematics Education Conference* (Vol. 3*,* pp. 144-151). São Paulo, Brazil: PME Program Committee.

Cai, J. (1998). Exploring students' conceptual understanding of the averaging algorithm. *School Science and Mathematics, 98*, 93-98.

Campbell, K. J., Watson, J. M., & Collis, K. F. (1992). Volume measurement and intellectual development. *Journal of Structural Learning*, *11*, 279-298.

Castles, I. (1992). *Surviving statistics: A user's guide to the basics.* Canberra: Australian Bureau of Statistics.

Cockcroft, W. H. (1982). *Mathematics counts: Report of the Committee of Inquiry into the Teaching of Mathematics in Schools*. London: HMSO.

Collis, K. F., Romberg, T. A., & Jurdak, M. E. (1986). A technique for assessing mathematical problem solving ability. *Journal for Research in Mathematics Education, 17*, 206-221.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cureton, E. E. (1965). Reliability and validity: Basic assumptions and experimental designs. *Educational and Psychological Measurement, 25*, 326-346.

Department of Education Tasmania. (2002). *Essential learnings framework 1*. Hobart: Author.

Dossey, J. A. (1997). National indicators of quantitative literacy. In L. A. Steen (Ed.), *Why numbers count: Quantitative literacy for tomorrow's America* (pp. 45-59). New York : The College Board.

Education Queensland (2000). *New Basics Project technical paper*. Retrieved January 11, 2002, from http://education.qld.gov.au/corporate/newbasics/html/library.html

Eisner, E. W. (1993). Reshaping assessment in education: Some criteria in search of practice. *Journal of Curriculum Studies, 25*(3), 219-233.

Fischbein, E. (1975). *The intuitive sources of probabilistic thinking in children*. Dordrecht: D. Reidel.

Fischbein, E., & Gazit, A. (1984). Does the teaching of probability improve probabilistic intuitions? An exploratory research study. *Educational Studies in Mathematics, 15*, 1-24.

Fisher, W. P. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Vol. 2* (pp. 36-72). Norwood, NJ: Ablex.

Frankenstein, M. (2001). Reading the world with math: Goals for a critical mathematical literacy curriculum. In *Mathematics shaping Australia* (Proceedings of the 18th Biennial Conference of the Australian Association of Mathematics Teachers, Inc.). [CDROM] Canberra: AAMT.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education, 32*, 124-158.

Gagatsis, A., Kyriakides, L., & Panaoura, A. (2001). Construct validity of a developmental assessment on probabilities: A Rasch measurement model analysis. In M. van den Heuvel-Panhuizen (Ed.), *Proceedings of the 25th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 2, pp. 449-456). Utrecht, The Netherlands: Freudenthal Institute.

Gal, I. (2000). Statistical literacy: Conceptual and instructional issues. In D. Coben, J. O'Donoghue, & G. E. Fitzsimons (Eds.), *Perspectives on adults learning mathematics: Research and practice* (pp. 135-150). Dordrecht: Kluwer.

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

Gal, I., & Wagner, D. A. (1992). *Project STARC: Statistical reasoning in the classroom*. (Annual Report: Year 2, NSF Grant No. MDR90-50006). Philadelphia, PA: Literacy Research Center, University of Pennsylvania.

Garfield, J. B. (2003). Assessing statistical reasoning. *Statistics Education Research Journal, 2*(1), 23-38.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519-521.

Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*, 923-936.

Green, D. R. (1982). *Probability concepts in 11-16 year old pupils*. Loughborough, UK: Center for Advancement of Mathematical Education in Technology, University of Technology.

Green, D. (1983a). Shaking a six. *Mathematics in Schools, 12*(5), 29-32.

Green, D. R. (1983b). A survey of probability concepts in 3000 pupils aged 11-16 years. In D. R. Grey, P. Holmes, V. Barnett, & G. M. Constable (Eds.), *Proceedings of the 1st International Conference on Teaching Statistics* (Vol. 2, pp. 766-783). Sheffield, England: Teaching Statistics Trust.

Green, D. (1993). Data analysis: What research do we need? In L. Pereira-Mendoza (Ed.), *Introducing data analysis in the schools: Who should teach it and how?* (pp. 219-239). Voorburg, The Netherlands: International Statistical Institute.

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

Izard, J. (1992). Patterns of development with probability concepts: Assessment for informative purposes. In M. Stephens & J. Izard (Eds.), *Reshaping assessment practices: Assessment in the mathematical sciences under challenge. Proceedings from the First National Conference on Assessment in the Mathematical Sciences* (pp. 355-367). Melbourne, VIC: Australian Council for Educational Research.

Jacobs, V. R. (1999). How do students think about statistical sampling before instruction? *Mathematics in the Middle School, 5*(4), 240-263.

Jones, B. (1982). *Sleepers wake: Technology and the future of work.* Melbourne: Oxford University Press.

Jones, G. A., Langrall, C. W., Thornton, C. A., & Mogill, A. T. (1997). A framework for assessing young children's thinking in probability. *Educational Studies in Mathematics, 32*, 101-125.

Jones, G. A., Thornton, C. A., Langrall, C. W., Mooney, E. S., Perry, B., & Putt, I. J. (2000). A framework for characterizing children's statistical thinking. *Mathematical Thinking and Learning, 2*, 269-307.

Keeves, J. P., & 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.

Kirsch, I. W. (1997). Literacy performance on three scales: Definitions and results. In W. McLennan, *Aspects of literacy: Assessed skill levels Australia 1996* (pp. 98-124). Canberra: Australian Bureau of Statistics.

Kolen, M. J. (1999). Equating of tests. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 164-175). New York: Pergamon.

Konold, C., & Garfield, J. (1992). *Statistical reasoning assessment: Part 1. Intuitive Thinking*. Scientific Reasoning Research Institute, University of Massachusetts, Amherst, MA.

Konold, C., & Higgins, T. L. (2002). Working with data: Highlights related to research. In S. J. Russell, D. Schifter, & V. Bastable (Eds.), *Developing mathematical ideas: Collecting, representing, and analyzing data* (pp. 165-201). Parsippany, NJ: Dale Seymour Publications.

Linacre, M. (1997, August). *Judging plans and facets*. MESA Research Note 3. Retrieved January 8, 2003, from http://www.rasch.org/m3.htm

Lokan, J. Ford, P., & Greenwood, L. (1997). *Maths and science on the line: Australian middle primary students' performance in the Third International Mathematics and Science Survey (TIMSS)*. Melbourne: Australian Council for Educational Research.

Luke, A., & Freebody, P. (1997). Shaping the social practices of reading. In S. Musprati, A. Luke, & P. Freebody (Eds.), *Constructing critical literacies: Teaching and learning textual practice* (pp. 185-225). St. Leonards, NSW: Allen & Unwin.

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

McLennan, W. (1997). *Aspects of literacy: Assessed Skill Levels Australia 1996*. Canberra: Commonwealth of Australia.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed.) (pp. 13-103). New York: American Council on Education and Macmillan Publishing Company.

Metz, K. E. (1998). Emergent understanding and attribution of randomness: Comparative analysis of the reasoning of primary grade children and undergraduates. *Cognition and Instruction, 16*, 285-365.

Mevarech, Z. R., & Kramarsky, B. (1997). From verbal descriptions to graphic representations: Stability and change in students' alternative conceptions. *Educational Studies in Mathematics, 32*, 229-263.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook.* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Ministry of Education (1992). *Mathematics in the New Zealand curriculum*. Wellington: Author.

Mokros, J., & Russell, S. J. (1995). Children's concepts of average and representativeness. *Journal for Research in Mathematics Education*, *26*(1), 20-39.

Moore, D. S. (1990). Uncertainty. In L. A. Steen (Ed.), *On the shoulders of giants: New approaches to numeracy* (pp. 95-137). Washington, DC: National Academy Press.

Moritz, J. B. (2000). Graphical representations of statistical associations by upper primary students. In J. Bana & A. Chapman (Eds.), *Mathematics education beyond 2000: Proceedings of the 23rd Annual Conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 440-447). Perth, WA: MERGA.

Moritz, J. B., & Watson, J. M. (1997). Graphs: Communication lines to students? In F. Biddulph & K. Carr (Eds.), *People in mathematics education: Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 344-351). Rotorua, NZ: MERGA.

Moritz, J. B., & Watson, J. M. (2000). Reasoning and expressing probability in students' judgements of coin tossing. In J. Bana & A. Chapman (Eds.), *Mathematics education beyond 2000: Proceedings of the 23rd annual conference of the Mathematics Education Research Group of Australasia* (Vol. 2, pp. 448-455). Perth, WA: MERGA.

Moritz, J. B., Watson, J. M., & Collis, K. F. (1996). Odds: Chance measurement in three contexts. In P. C. Clarkson (Ed.), *Technology in mathematics education: Proceedings of the 19th annual conference of the Mathematics Education Research Group of Australasia* (pp. 390-397). Melbourne: MERGA.

Moritz, J. B., Watson, J.M., & Pereira-Mendoza, L. (1996, November). *The language of statistical understanding: An investigation in two countries*. Paper presented at the Joint ERA/AARE Conference, Singapore. [Online: http://www.swin.edu.au/aare/96pap/morij96.280]

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: Author.

National Council of Teachers of Mathematics (2000). *Principles and standards for school mathematics*. Reston, VA: Author.

Nisbett, R. E., Krantz, D. H., Jepson, C., & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review, 90*, 339-363.

Pollatsek, A., Well, A. D., Konold, C., Hardiman, P., & Cobb, G. (1987). Understanding conditional probabilities. *Organizational Behavior and Human Decision Processes, 40*, 255-269.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press (original work published 1960).

Reading, C. (2002). Profile for statistical understanding. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa*. Voorburg, The Netherlands: International Statistical Institute.

Reading, C., & Shaughnessy, M. (2000). Student perceptions of variation in a sampling situation. In T. Nakahara & M. Koyama (Eds.), *Proceedings of the 24th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 89-96). Hiroshima, Japan: Hiroshima University.

Reading, C., & Shaughnessy, M. (in press). Reasoning about variation. In J. Garfield & D. Ben-Zvi (Eds.), *The challenge of developing statistical literacy, reasoning, and thinking.* Dordrecht: Kluwer.

Romberg, T. A., Jurdak, M. E., Collis, K. F., & Buchanan, A. E. (1982). *Construct validity of a set of mathematical superitems*. Madison, Wisconsin: Wisconsin Center for Education Research.

Shaughnessy, J. M. (1997). Missed opportunities in research on the teaching and learning of data and chance. In F. Biddulph & K. Carr (Eds.), *People in mathematics education Proceedings of the 20th annual conference of the Mathematics Education Research Group of Australasia* (Vol. 1, pp. 6-22). Rotorua, NZ: MERGA.

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*. Presession Research Symposium, 77th Annual National Council of Teachers of Mathematics Conference, San Francisco, CA.

Statistics Canada and Organisation for Economic Cooperation and Development (OECD) (1996). *Literacy, economy, and society: First results from the International Adult Literacy Survey*. Ottawa: Author.

Steen, L. A. (Ed.) (1997). *Why numbers count: Quantitative literacy for tomorrow's America*. New York: College Entrance Examination Board.

Steen, L. A. (Ed.) (2001). *Mathematics and democracy: The case for quantitative literacy.* Washington, DC: National Council on Education and the Disciplines.

Stocking, M. L. (1999). Item response theory. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 55-63). New York: Pergamon.

Tognolini, J. (1996). Rasch modelling: Advantages and limitations. In *Session notes National Meeting on Assessment and Reporting 25-26 November 1996*. Manly: NSW Department of School Education.

Torok, R. (2000). Putting the variation into chance and data. *Australian Mathematics Teacher*, *56(2),* 25-31.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgement. *Psychological Review, 90,* 293-315.

Wallman, K. K. (1993). Enhancing statistical literacy: Enriching our society. *Journal of the American Statistical Association, 88*, No. 421, 1-8.

Watson, J. M. (1994). Instruments to assess statistical concepts in the school curriculum. In National Organizing Committee (Ed.), *Proceedings of the Fourth International Conference on Teaching Statistics. Volume 1* (Vol. 1, pp. 73-80). Rabat, Morocco: National Institute of Statistics and Applied Economics

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. (1998a). Numeracy benchmarks for years 3 and 5: What about chance and data? In C. Kanes, M. Goos, & E. Warren (Eds.), *Teaching mathematics in new times.* (Proceedings of the 21st annual conference of the Mathematics Education Research Group of Australasia, Vol. 2, pp. 669-676). Brisbane: MERGA.

Watson, J. M. (1998b). The role of statistical literacy in decisions about risk: Where to start. *For the Learning of Mathematics*, *18*(3), 25-27.

Watson, J. M. (2000). Statistics in context. *Mathematics Teacher*, *93*, 54-58.

Watson, J. M., Collis, K. F., & Moritz, J. B. (1994). Assessing statistical understanding in Grades 3, 6 and 9 using a short answer questionnaire. In G. Bell, B. Wright, N. Leeson, & G. Geake (Eds.), *Challenges in Mathematics Education: Constraints on Construction Proceedings of the 17th Annual Conference of the Mathematics Education Research Group of Australasia* (pp. 675-682). Lismore, NSW: MERGA.

Watson, J. M., Collis, K. F., & Moritz, J. B. (1997). The development of chance measurement. *Mathematics Education Research Journal*, *9*, 60-82.

Watson, J. M., & Kelly, B. A. (2002). Can grade 3 students learn about variation? In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching Statistics: Developing a statistically literate society, Cape Town, South Africa*. Voorburg, The Netherlands: International Statistical Institute.

Watson, J. M., Kelly, B. A., Callingham, R. A., & Shaughnessy, J. M. (2003). The measurement of school students' understanding of statistical variation. *International Journal of Mathematical Education in Science and Technology, 34,*1-29.

Watson, J. M., & Moritz, J. B. (1998). Longitudinal development of chance measurement. *Mathematics Education Research Journal*, *10*(2), 103-127.

Watson, J. M., & Moritz, J. B. (1999a). The development of concepts of average. *Focus on Learning Problems in Mathematics*, *21*(4), 15-39.

Watson, J. M., & Moritz, J. B. (1999b). The beginning of statistical inference: Comparing two data sets. *Educational Studies in Mathematics*, *37*, 145-168.

Watson, J. M., & Moritz, J. B. (2000a). Developing concepts of sampling. *Journal for Research in Mathematics Education*, *31*, 44-70.

Watson, J. M., & Moritz, J. B. (2000b). Development of understanding of sampling for statistical literacy. *Journal of Mathematical Behavior*, *19*, 109-136.

Watson, J. M., & Moritz, J. B. (2002). School students' reasoning about conjunction and conditional events. *International Journal of Mathematical Education in Science and Technology, 33,* 59-84.

Watson, J. M., & Moritz, J. B. (2003). The development of comprehension of chance language: Evaluation and interpretation. *School Science and Mathematics, 103*, 65-80.

Wild, C. J., & Pfannkuch, M. (1999). Statistical thinking in empirical enquiry. *International Statistical Review, 67*, 223-265.

Wilson, M. (1990). Investigation of structured problem-solving items. In G. Kulm (Ed.), *Assessing higher order thinking in mathematics* (pp. 187-203). Washington, DC: American Association for the Advancement of Science.

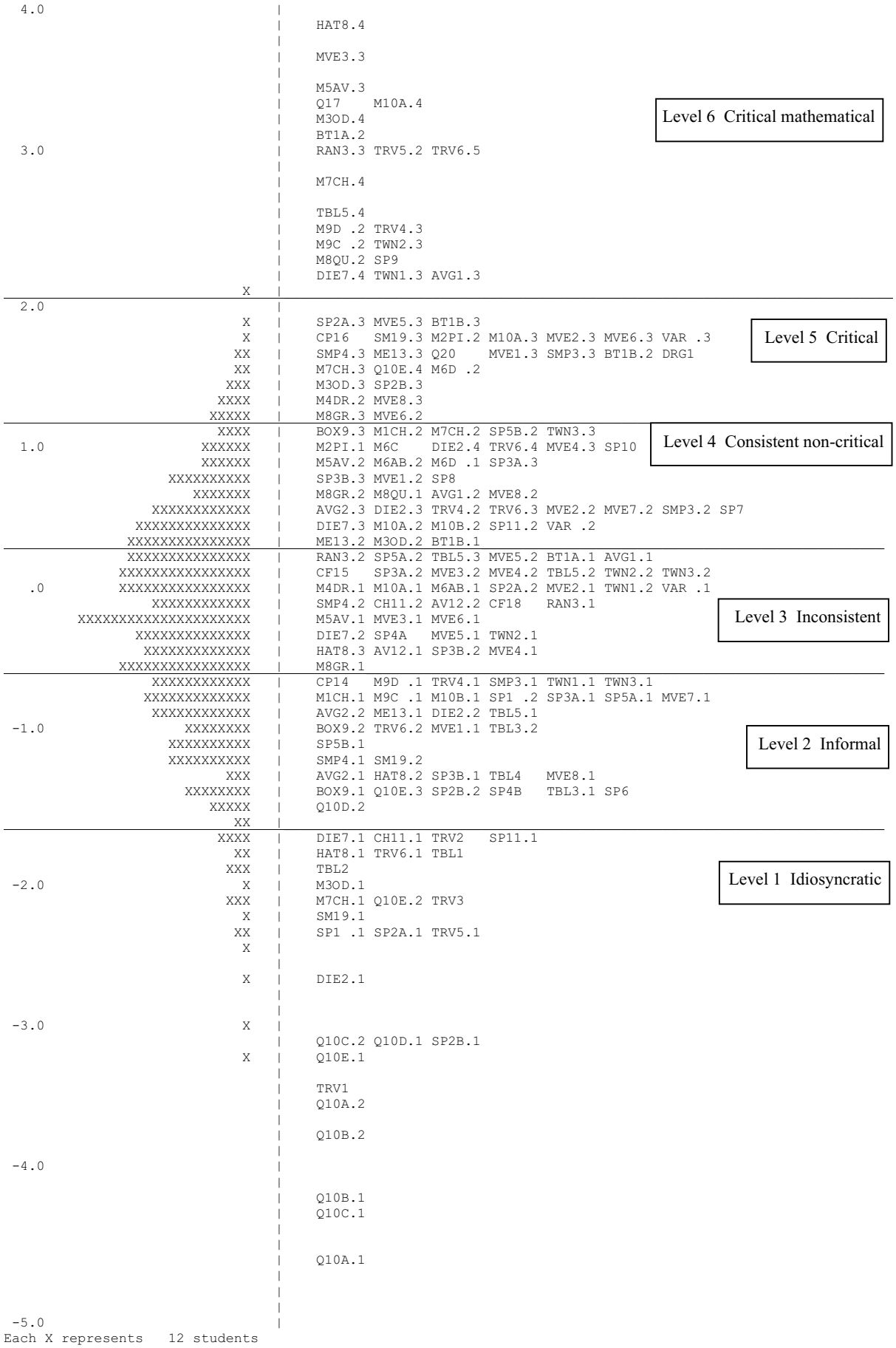
Wilson, M. (1992). Measuring levels of mathematical understanding. In T. A. Romberg (Ed.), *Mathematics assessment and evaluation: Imperatives for mathematics educators* (pp. 213-241). Albany, NY: State University of New York Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.

Zawojewski, J. S., & Shaughnessy, J. M. (2000). Data and chance. In E. A. Silver & P. A. Kenney (Eds.), *Results from the Seventh Mathematics Assessment of the National Assessment of Educational Progress* (235-268). Reston, VA: NCTM.

JANE WATSON
Faculty of Education
University of Tasmania
Private Bag 66, Hobart
TAS 7001 Australia

APPENDIX A: VARIABLE MAP FOR STATISTICAL LITERACY CONSTRUCT



APPENDIX B: NUMBER OF STUDENTS IN EACH GRADE ANSWERING EACH ITEM

	GRADE					
ITEM	**3**	**5**	**6**	**7**	**8**	**9**
AV12		218	518		167	641
AVG1				189		197
AVG2	626	218	518		167	641
BOX9	1039	421	875	239	196	1034
BT1A, BT1B				189		197
CF15, CF18, CP14, CP16		238	861	46	196	837
CH11		238	875	50	196	837
DIE2	176	183		189		197
DIE7	1039	421	875	239	196	1034
DRG1				189		197
HAT8	863	238	875	50	196	837
M10A, M10B			395		185	403
M1CH, M2PI, M3OD			695		185	746
M4DR				189	185	943
M5AV			517		165	618
M6AB, M6C, M6D					184	746
M7CH			695		185	746
M8GR, M8QU			396		185	746
M9C, M9D			521		176	647
ME13		218	518		167	641
MVE1 to MVE4, MVE7	176	183		189		197
MVE5		183		189		197
MVE6				189		197
MVE8						197
Q10A to Q10E	854	238	875	50	196	837
Q17, Q20		238	861	46	196	837
RAN3	863	238	875	239	196	1034
SM19		238	861	46	196	837
SMP3	176	183		189		197
SMP4	863	238	875	50	196	837
SP1, SP2A	176	183		189		197
SP2B				189		197
SP3A	176	183				
SP3B				189		197
SP4A	176	183				
SP4B				189		197
SP5A	176	183				
SP5B				189		197
SP6 to SP11				189		197
TBL1 to TBL5	176	183		189		197
TRV1 to TRV6	176	183		189		197
TWN1 to TWN3		183		189		197
VAR				189		197

APPENDIX C: ITEM STATEMENT AND RESPONSE CODE EXAMPLES

AV12. To get the average number of children per family in a town, a teacher counted the total number of children in the town. She then divided by 50, the total number of families. The average number of children per family was 2.2.
Tick which of these is certain to be true.
(a) Half of the families in the town have more than 2 children.
(b) More families in the town have 3 children than have 2 children.
(c) There are a total of 110 children in the town.
(d) There are 2.2 children in the town for every adult.
(e) The most common number of children in a family is 2.
(f) None of the above.

Code 3	c
Code 2	d, e, f, multiple
Code 1	a, b
Code 0	NR

AVG1. A small object was weighed on the same scales separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.

6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3

The "average" value could be calculated in several ways.

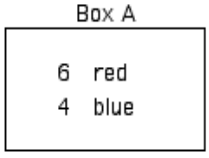
1. How would you find the average? _____
2. The average weight is _____ grams. [Show your working in the box below.]

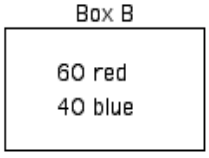
Code 3	Mode explained; Median explained and correct; Mean discarding outlier
Code 2	7.18, mean
Code 1	Any of the three measures mentioned but answer incorrect
Code 0	Incorrect or idiosyncratic method with or without unreasonable answer; NR

AVG2. If someone said you were "average", what would it mean?

Code 3	Add and divide, same as most, in the middle between good and bad
Code 2	Add, same as others, okay, normal
Code 1	Example
Code 0	Don't know, etc; No response (NR)

BOX9. Box A and Box B are filled with red and blue marbles as follows. Each box is shaken. You want to get a blue marble, but you are only allowed to pick out one marble without looking. Which box should you choose?





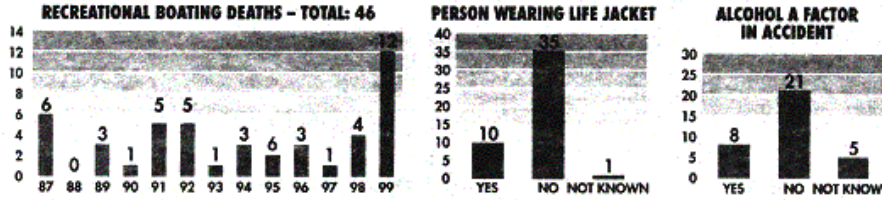
(A) Box A (with 6 red and 4 blue).
(B) Box B (with 60 red and 40 blue).
(=) It doesn't matter.

Please explain your answer.

Code 3	=, 40% chance each, B is 10 times bigger than A, same chance; 40 versus 4, 6 versus 6, similar
Code 2	=, both have more red; A, only 2 more reds, B 20 more reds; B, more blues, more marbles; A less reds, less marbles
Code 1	=, could be anything; A, B, = idiosyncratic reason
Code 0	NR, no reason

BT1. These graphs were part of a newspaper story reporting on boating deaths in Tasmania. Comment on any unusual features of the graphs. (2 spaces provided)

BOATIE'S SAFETY FAILURE



BT1A.

Code 2	Recognises mistakes: "The axes aren't named" & "The total boating deaths of 46 is not the same shown in the graph"
Code 1	Statistically specific comments about graphing elements; perhaps includes some incorrect comments
Code 0	Incorrect graph interpretation of unusual data; Inferring from graph; Advice; idiosyncratic

BT1B.

Code 3	Acknowledges variation: "In some years there were heaps and they dropped to none"
Code 2	Focuses on increase: "As the years progress the amount of years grew"
Code 1	Focuses on the highest column, a single column or only 2 columns
Code 0	No focus: "Numbers are above the shaded area"

CF15. Please estimate:
(a) The probability that you will miss a whole week of school next year.
(b) The probability that you will get a cold next year.
(c) The probability that you will get a cold causing you to miss a whole week of school next year.

Code 1	$c<\min(a,b)$; $c=\min(a,b)$
Code 0	$\min<c\leq\max$, $c>\max(a,b)$, undefined

CF18. A health survey was conducted in a sample of 100 men in Australia of all ages and occupations. Please estimate:
(a) How many of the 100 men have had one or more heart attacks.
(b) How many of the 100 men are over 55 years old.
(c) How many of the 100 men both are over 55 years old and have had one or more heart attacks.

Code 1	$c<\min(a,b)$; $c=\min(a,b)$
Code 0	$\min<c\leq\max$, $c>\max(a,b)$, undefined

CH11. A bottle of medicine has printed on it*: WARNING: For applications to skin areas there is a 15% chance of getting a rash. If you get a rash, consult your doctor.* What does this mean?
(a) Don't use the medicine on your skin - there's a good chance of getting a rash.
(b) For application to the skin, apply only 15% of the recommended dose.
(c) If you get a rash, it will probably involve only 15% of the skin.
(d) About 15 out of every 100 people who use this medicine get a rash.
(e) There is hardly any chance of getting a rash using this medicine.

Code 2	d, a and d, d and e
Code 1	e, a
Code 0	b, c, NR, multiple selections

CP14. Please estimate:
(a) Out of 100 men, how many are left-handed.
(b) Out of 100 left-handed adults, how many are men.

Code 1	b>a
Code 0	b=a/2, b=a, b<a, undefined

CP16. Please estimate:
(a) The probability that a woman is a school teacher.
(b) The probability that a school teacher is a woman.

Code 1	b>a
Code 0	b=a/2, b=a, b<a, undefined

DIE2. Imagine you threw the dice 60 times. In the table below, fill in how many times you think each number might come up. Why do you think these numbers are reasonable? [Appeared immediately after DIE7]

Number on Dice	How many times it might come up
1	
2	
3	
4	
5	
6	
TOTAL	60

Code 4	Appropriate variability in prediction and reason
Code 3	Strict probability with reason reflecting variation; Too much or too little variation with reason reflecting chance
Code 2	Strict probability with reasons reflecting classical chance or aspects of geometry; multiples of 5 with reasons reflecting chance; Did not add to 60 with reasons reflecting equality or chance
Code 1	That's what I think.
Code 0	Sums ≠ 60 and odd distributions with no reasoning

DIE7. Consider rolling one six-sided die. Is it easier to throw
(1) a one or
(6) a six or
(=) are both a one and a six equally easy to throw?
Please explain your answer.

Code 4	=, 1/6 chance every number
Code 3	=, only one of each number, same chance, cube
Code 2	=, could be anything, never know outcome
Code 1	1, 6, = idiosyncratic reason
Code 0	NR

DRG1. What was the size of the sample in this article?

Decriminalise drug use: poll
SOME 96 percent of callers to youth radio station Triple J have said marijuana use should be decriminalised in Australia. The phone-in listener poll, which closed yesterday, showed 9924 - out of the 10,000-plus callers - favoured decriminalisation, the station said. Only 389 believed possession of the drug should remain a criminal offence. Many callers stressed they did not smoke marijuana but still believed in decriminalising its use, a Triple J statement said.

Code 1	10,313; 10,000+; 10,000
Code 0	9924 out of 10 000+ callers; 96%; very small

HAT8. A mathematics class has 13 boys and 16 girls in it. Each pupil's name is written on a piece of paper. All the names are put in a hat. The teacher picks out one name without looking. Is it more likely that
(b) the name is a boy or
(g) the name is a girl or
(=) are both a girl and a boy equally likely?
Please explain your answer.

Code 4	g, 16/29 chance
Code 3	g, 13 versus 16, more girls
Code 2	=, depends on mix, same chance, could be anything
Code 1	b, g, = idiosyncratic, such as luck or teacher is a certain sex
Code 0	NR, no reason

M10A. During the recent Australian cricket tour of South Africa, the Hobart Mercury (6/4/1994, p. 52) reported that Allan Border had lost 8 out of 9 tosses in his previous 9 matches as captain. Imagine his situation at this point in time.
Suppose Border decides to choose heads from now on. For the next 4 tosses of the coin, what is the chance of the coin coming up tails (and him losing the tosses) 4 times out of 4?

Code 4	1/16
Code 3	Other number/word
Code 2	50%, 50-50, 2/4 other, word
Code 1	Value > 0.5
Code 0	NR

M10B. Suppose tails came up 4 times out of 4. For the 5th toss, should Border choose
__Heads
__Tails
__Doesn't matter
What is the probability of getting heads on this next toss?
What is the probability of getting tails on this next toss?

Code 2	=, each value 0.5
Code 1	= other values; H/T, each value 0.5
Code 0	H/T other values

M1CH. Here are eight chance words or phrases from headlines.
A. 58 per cent success at SkillShare
B. Impossible
C. It's a sure thing
D. Jack looking good for big one

E. Holden an unlikely American hero
F. No worries
G. Smith in doubt to play
H. There's a 50-50 chance

Please mark on the scale below the likelihood expressed by each of the seven phrases A to G. H is done as an example.

Likelihood

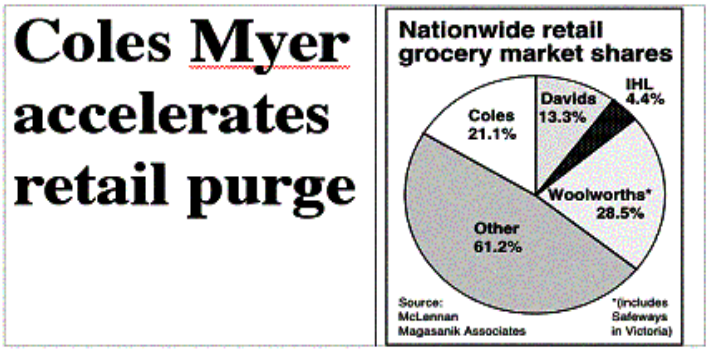
(low) (high)

|_____|_____|

0 1

Code 2	Letter in position and correct halves
Code 1	Up to 2 half errors
Code 0	>2 errors on halves

M2PI. Explain the meaning of this pie chart. Is there anything unusual about it?



Code 2	>100%; 61.2% less than half
Code 1	Large other; not explained % does match headline; company names
Code 0	Colours, no/yes, NR

M3OD. What does "7-2" mean in this headline about the North against South football match? Give as much detail as you can. From the numbers, who would be expected to win the game?

North at 7-2
But we can still win match, says coach

Code 4	South, 2/9
Code 3	North, 7/9; South, 2/7, if play, win 2/7; Chance 3.5 to 1, won 2/7
Code 2	% chance, 2 pts to 7 pts, odds; Chance, predicted score
Code 1	Current score, betting
Code 0	NR, no reason

M4DR. Is the sample reported here a reliable way of finding out public support for the decriminalisation of marijuana? Why or why not? [see DRG1]

Code 2	Only JJJ listeners, Only motivated ones phone in
Code 1	No, small sample; Yes, large sample; No, not everyone; Yes, everyone listens to radio; Reliability of measurement; No, could be lying
Code 0	Shouldn't have marijuana, NR

M5AV. What does "average" mean in this article? What does "median" mean in this article? Why would the median have been used?

Hobart defies homes trend

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

Code 3	Median not influenced by outliers as mean is; Good contrast median and mean
Code 2	Add and divide, middle value
Code 1	Normal
Code 0	Tautology/irrelevant, NR

M6. Each of the four sentences in the following article sets a condition and describes an associated outcome. In each case, state what these are.

Wrinkles ultimate smoking deterrent.

1. A study found that those who smoked a pack of cigarettes a day for less than 49 years doubled the risk of premature wrinkling.
2. For more than 50 years, the risk was 4.7 times greater than those who do not smoke.
3. He said he was not sure if the wrinkling could be reversed if people quit smoking.
4. " 'You're going to be old and ugly before your time if you smoke,' may be just the message that leads them to throw away their cigarettes for good," he said.

Condition	Outcome
Q1. _____	_____
Q2. _____	_____
Q3. _____	_____
Q4. _____	_____

M6AB.

Code 2	Both Q1 and Q2 correct: {Smoke, Cigarette, Pack} →{Wrinkle, Premature}{50} → {Wrinkle, Risk, 4.7}
Code 1	One of Q1 and Q2 correct
Code 0	Incorrect, NR

M6C.

Code 1	Q3 correct: {Smoke, Quit} → {Wrinkle, Reverse}
Code 0	Incorrect, NR

M6D.

Code 2	Q4 correct: {Message, Wrinkle, Old, Ugly} → {Quit, Not Smoke}
Code 1	Q4 correct:{Smoke} → {Wrinkle, Old, Ugly}
Code 0	Incorrect, NR

M7CH. Would you make any criticisms of the claims in this article? If you were a high school teacher, would this report make you refuse a job offer somewhere else in the United States, say Colorado or Arizona? Why or why not?

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

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

M8

Family car is killing us, says Tasmanian researcher

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

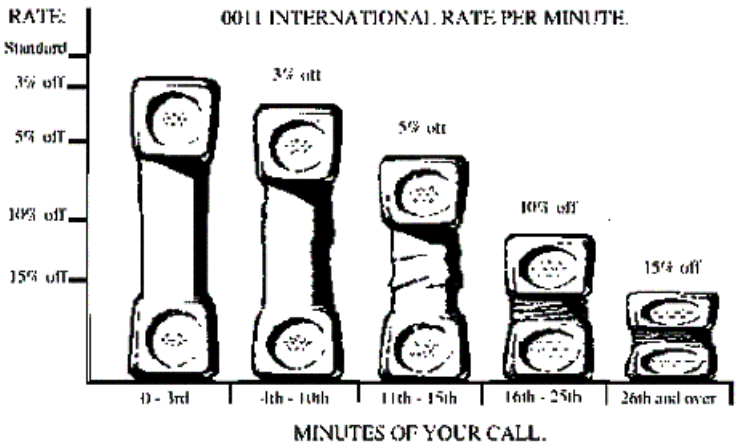
M8GR. Draw and label a sketch of what one of Mr. Robinson's graphs might look like.

Code 3	Bivariate or Series Comparison Graph
Code 2	Trend or Double comparison Graph
Code 1	Labeled or Single Comparison; Basic Graph
Code 0	No graph

M8QU. What questions would you ask about his research?

Code 2	Other causes? How linked?
Code 1	Sample size, & location; Location, size, age groups
Code 0	Can it be prevented?

M9. The longer your overseas call, the cheaper the rate.



M9a Explain the meaning of this graph. [Not coded]

M9b Is there anything unusual about it? [Not coded]

M9C. Suppose the standard rate is $1.00 for 1 minute.You have already talked for 30 minutes How much would the next 10 minutes cost?

Code 2	$8.50
Code 1	15%, $1.50, 85c; 3%, $10, $40, Other$
Code 0	NR

M9D. How much did the first 30 minutes of the phone call cost?

Code 2	$27.79 or calculation by right method; $25.50
Code 1	15%, $4.50; $30, Other$, Other %
Code 0	NR

ME13. A small object was weighed on the same scale separately by nine students in a science class. The weights (in grams) recorded by each student are shown below.
6.3 6.0 6.0 15.3 6.1 6.3 6.2 6.15 6.3
The median of this set of data is
(a) the most common value.
(b) the middle value.
(c) the most accurate value.
(d) the average value.
So, the median value is _____ grams.

Code 3	b & 6.2; d & 7.18
Code 2	a & 6.3; b & 6.1; d and 6≤#≤8
Code 1	c, other selections not above
Code 0	NR

MVE. MOVIEWORLD
A class wanted to raise money for their school trip to Movieworld on the Gold Coast. They could raise money by selling raffle tickets for a Nintendo Game system. But before they decided to have a raffle they wanted to estimate how many students in their whole school would buy a ticket. So they decided to do a survey to find out first. The school has 600 students in grades 1-6 with 100 students in each grade.

MVE1. How many students would you survey and how would you choose them? Why?

Code 3	Representative & random; Random only
Code 2	Based on one or more factors
Code 1	Just the students I meet; take them all
Code 0	Misinterpretation

MVE2. Three students in the school conducted surveys. Shannon got the names of all 600 children in the school and put them in a hat, and then pulled out 60 of them. What do you think of Shannon's survey?
__Good __ Bad __ Not Sure - Why?

Code 3	Random methods; range
Code 2	Fair chance; sample size; methodology (easy)
Code 1	Method too random, inaccurate; inadequate sample size; unfair; time consumption
Code 0	Misinterpretation; no reason or logic

MVE3. Jake asked 10 children at an after-school meeting of the computer games club. What do you think of Jake's survey?
__Good __ Bad __ Not Sure - Why?

Code 3	Detecting bias & small sample size
Code 2	Bias only, small sample size only; unfair, survey all
Code 1	Creating bias, good sample size; good method
Code 0	Misinterpretation; no reason or logic

MVE4. Adam asked all of the 100 children in Grade 1. What do you think of Adam's survey?
__Good __ Bad __ Not Sure - Why?

Code 3	Detecting bias in groups
Code 2	Sample size too large; unfair; not sure
Code 1	Large sample size good; fair
Code 0	Misinterpretation; no reason or logic

MVE5. Raffi surveyed 60 of his friends. What do you think of Raffi's survey?
__Good __ Bad __ Not Sure - Why?

Code 3	Lack of range &/or variation
Code 2	Unfair; vague friendship factor; uncertainty; adequate sample size
Code 1	Inadequate sample size; 'easy'; good to use friends
Code 0	Misinterpretation; no reason or logic

MVE6. Claire set up a booth outside of the tuck shop. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 kids to complete them. What do you think of Claire's survey?
__Good __ Bad __ Not Sure - Why?

Code 3	Non-representative
Code 2	Uncertainty; adequate sample size
Code 1	Inadequate sample size; fairness; free choice; assuming range and variation; 'easy'
Code 0	Misinterpretation; no reason or logic

MVE7. Who do you think has the best survey method? - Why?

Code 2	Shannon or Shannon plus another
Code 1	Raffi, Claire, etc., with reason
Code 0	Raffi, Claire, etc., with no reason or logic

MVE8. What percent of students in the whole school will buy a raffle ticket? - (Circle one)
a. 35% (Shannon's result) because _____
b. 90% (Jake's result) because _____
c. 50% (Adam's result) because _____
d. 75% (Raffi's result) because _____
e. 95% (Claire's result) because _____
f. I think it is best to average the 5 surveys. The average of the kids that said they would buy a raffle ticket is 69%.
g. I don't know because Raffi, Shannon, Claire, Jake and Adam all got different results.
h. I think that __percent of the kids in the whole school are willing to buy a raffle ticket because_____

Code 3	Unpersuaded by new information – Shannon; influenced choice from another earlier
Code 2	Average them with or without doubt
Code 1	Uncertainty with or without doubt; Unpersuaded by new information from inappropriate choice
Code 0	Misinterpretation; idiosyncratic; no reason or logic

Q10. A primary school had a sports day where every child could choose a sport to play. Here is what they chose:

	Netball	Soccer	Tennis	Swimming
Girls	30	5	15	10
Boys	0	20	18	20

Q10A. How many girls chose tennis?

Code 2	15
Code 1	Other
Code 0	NR

Q10B. How many boys chose netball?

Code 2	0
Code 1	Other
Code 0	NR

Q10C. How many children chose swimming?

Code 2	30
Code 1	Other
Code 0	NR

Q10D. In which sport were boys and girls most evenly divided?

Code 2	Tennis
Code 1	Other
Code 0	NR

Q10E. Were there more girls or more boys at the sports day? How do you know?

Code 4	(girls) 60 vs 58 totals correct
Code 3	(girls/boys/other) totals error add/count
Code 2	(g) other reason
Code 1	(boys/other) other reason
Code 0	NR

Q17. A farmer wants to know how many fish there are in his dam. He took out 200 fish and tagged each of them, with a coloured sign. He put the tagged fish back in the dam and let them get mixed with the others. On the second day, he took out 250 fish in a random manner, and found that 25 of them were tagged. Estimate how many fish are in the dam.

Code 1	2000
Code 0	Other response, NR

Q20. Every year, Susan selects about 5 young actors for the drama team who perform brilliantly at audition. Unfortunately, most of these kids turn out to be no better than the rest. Why do you suppose that Susan usually finds that they don't turn out to be as brilliant as she first thought?

(a) In her eagerness to find new talent, Susan may exaggerate the brilliance of the performances she sees at the audition.
(b) The actors probably just made some nice acts at the audition that were much better than usual for them.
(c) The actors probably coast on their talent alone without putting in the effort for a consistently excellent performance.
(d) The actors who did so well at the audition may find that the others are jealous, and so they slack off.
(e) The actors who did so well are likely to be students with other interests, so they don't put all their energies into acting after the audition.

Code 1	b, b + another
Code 0	a, c, d, e, other (multiple letter not including b, or including all)

RAN3. What things happen in a "random" way?

Code 3	Definition + Example; "To pick without any pattern"
Code 2	Definition – No order, choose any, unpredictable; Multiple Examples from different aspects below
Code 1	Example – Natural (Weather), Human design (Breath testing), Game/selection (Tattslotto)
Code 0	Inappropriate (ransom, fighting, everything); Chosen (weak), in order, random numbers/alphabet, NR

SM19. Mrs. Jones wants to buy a new car, either a Honda or a Toyota. She wants whichever car will break down the least. First she read in Consumer Reports that for 400 cars of each type, the Toyota had more break-downs than the Honda. Then she talked to three friends. Two were Toyota owners, who had no major break-downs. The other friend used to own a Honda, but it had lots of break-downs, so he sold it. He said he'd never buy another Honda.

Which car should Mrs. Jones buy?
(T) Mrs. Jones should buy the Toyota, because her friend had so much trouble with his Honda, while her other friends had no trouble with their Toyotas.
(H) She should buy the Honda, because the information about break-downs in Consumer Reports is based on many cases, not just one or two cases.
(=) It doesn't matter which car she buys. Whichever type she gets, she could still be unlucky and get stuck with a particular car that would need a lot of repairs.

Code 3	H
Code 2	=
Code 1	T
Code 0	NR

SMP3. What does "sample" mean? Give an example of a "sample".

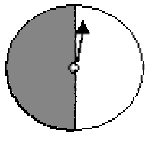
Code 3	Small part of whole to test of taste
Code 2	Small part of whole, part to test
Code 1	Test; try, piece, part
Code 0	Inappropriate; idiosyncratic; no response

SMP4. If you were given a "sample", what would you have?

Code 3	Small part of something to test
Code 2	Part of something, test of something, piece of carpet, taste of cheese
Code 1	Part, piece, test, example carpet
Code 0	NR

SP1. A class used this spinner. If you were to spin it once, what is the chance that it will land on the shaded part?

Code 2	50%, 1/2, 5/10, 1 in 2 chances, 50/50, half, same as white
Code 1	1 in 10, 80%, 20 out of 50, alright, any chance, bad chance
Code 0	I don't know



SP2A & B. Out of 10 (50) spins, how many times do you think the spinner will land on the shaded part? Why do you think this?

Code 3	Variation in one or both of response & answer
Code 2	Strict probability, implicit chance, at least 25; you can't tell (theoretically correct)
Code 1	Illogical or no reason with reasonable number
Code 0	NR

SP3A & B. If you were to spin it 10 (50) times again, would you expect to get the same number out of 10 (50) to land on the shaded part next time? Why do you think this?

Code 3	Sophisticated or simple recognition of variation
Code 2	Anything can happen, strict chance, implicit chance, contradiction
Code 1	Intuitive & primitive theories; Personal ideas & experiences
Code 0	Yes, just guessing

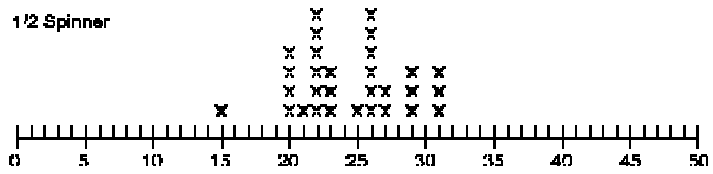
SP4A & B. How many times out of 10 (50) spins, landing on the shaded part, would surprise you?

Code 1	Grade 3 & 5 – 0,1,2,8,9,10; Grade 7 & 9 – <20, >30
Code 0	Grade 3 & 5 – 3,4,5,6,7; Grade 7 & 9 –20 to 30, Ambiguous, misinterpretation

SP5A & B. Suppose that you were to do 6 sets of 10 (50) spins. Write a list that would describe what might happen for the number of times the spinner would land on the shaded part?

Code 2	SD = 0.6-2.3 (10), SD = 1.3-5.0 (50)
Code 1	SD <0.6 >2.3 (10), SD <1.3 >5.0 (50), strict probability
Code 0	Out of range, misinterpretation

SP6. A class did 50 spins of the above spinner many times and the results for the number of times it landed on the shaded part are recorded below.
What is the lowest value?



Code 1	15
Code 0	Values with only one X above them; "1", values that have no X's above them; "0"; NR

SP7. What is the highest value?

Code 1	31
Code 0	Values with 6 X's above them; "6", "50"; no apparent logic

SP8. What is the range?

Code 1	16, 15-31
Code 0	20-27; 31; 50; 3; don't know

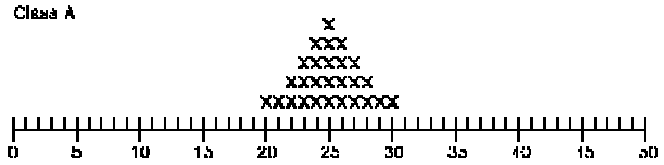
SP9. What is the mode?

Code 1	One or both of 22 and 26
Code 0	No logical reason; don't know

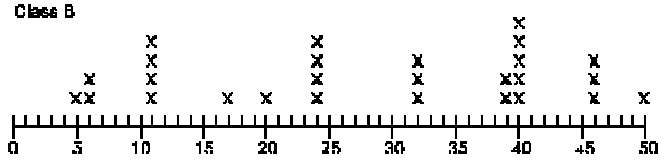
SP10. How would you describe the shape of the graph?

Code 1	Acknowledges variation; Focuses on physical objects, geometric shapes
Code 0	Focus on graph type or axes; illogical; NR

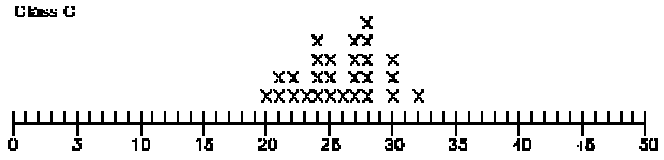
SP11. Imagine that three other classes produced graphs for the spinner. In some cases, the results were just made up without actually doing the experiment.



a) Do you think class A's results are made up or really from the experiment?
__Made up
__Real from experiment
Explain why you think this.



b) Do you think class B's results are made up or really from the experiment?
__Made up
__Real from experiment
Explain why you think this.



c) Do you think class C's results are made up or really from the experiment?
__Made up
__Real from experiment
Explain why you think this.

Code 2	"Made up: It would never be so even. Made up: Too spaced out. Real: Cause it's right." (3 correct)
Code 1	1 part incorrect with reasons; 2 parts incorrect with somewhat sensible reasoning
Code 0	Anything can happen; no reasoning for choices; NR

TBL. A primary school had a sports day where every child could choose a sport to play. Here is what they chose.

	Netball	Soccer	Tennis	Swimming	TOTAL
Boys	0	20	20	10	50
Girls	40	10	15	10	75

TBL1. How many girls chose Tennis?

Code 1	15
Code 0	Number other than 15; idiosyncratic

TBL2. What was the most popular sport for girls?

Code 1	Netball
Code 0	Partly correct (2 sports including Netball, 40); idiosyncratic, NR

TBL3. What was the most popular sport for boys?

Code 2	Soccer & Tennis
Code 1	Soccer or Tennis
Code 0	20; Sport other than Soccer or Tennis; or a number other than 20; NR

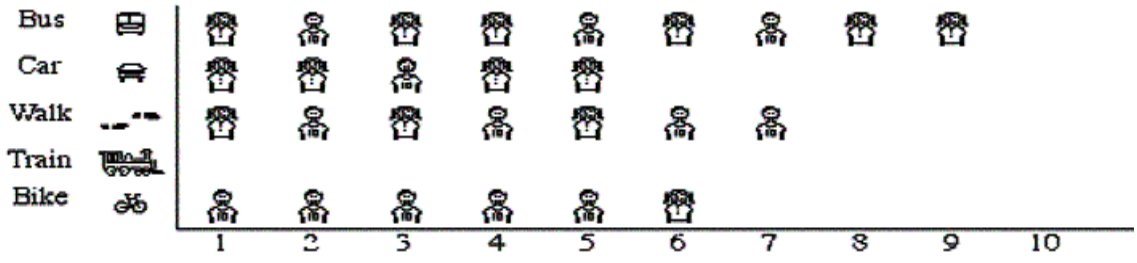
TBL4. How many children were at the sports day?

Code 1	125; Table reading only (50, 75)
Code 0	Computational fault; idiosyncratic comment; NR

TBL5. The teacher wanted to choose four children to lead the closing parade. Suggest two fair ways she could have chosen them.

Code 4	At least one method combining random selection & stratification; At least one combining random selection & stratification
Code 3	One chance method and various other possibilities
Code 2	At least one like "2 out of swimming, girl & boy"
Code 1	E.g., "The winners of each events or captains"
Code 0	Idiosyncratic methods like "They play the girls game first"

TRV. How children get to school one day
Number of students



TRV1. How many children walk to school?

Code 1	7
Code 0	Incorrect within range; odd comments

TRV2. How many more children come by bus than by car?

Code 1	4
Code 0	Bus; 9, 5, 14, a few

TRV3. Would the graph look the same everyday? Why or why not?

Code 1	Realistic or potential recognition of variation
Code 0	No variation or no reasoning

TRV4. A new student came to school by car. Is the new student a boy or a girl? How do you know?

Code 3	Explicit uncertainty - Probably a girl – More girls get a car to school; implicit - Girl – There is more chance of it being a girl
Code 2	Majority (local or global); Balance (local or global)
Code 1	Pattern in graph, could be either
Code 0	Not enough information, misinterpretation

TRV5. What does the row with the Train tell about how the children get to school?

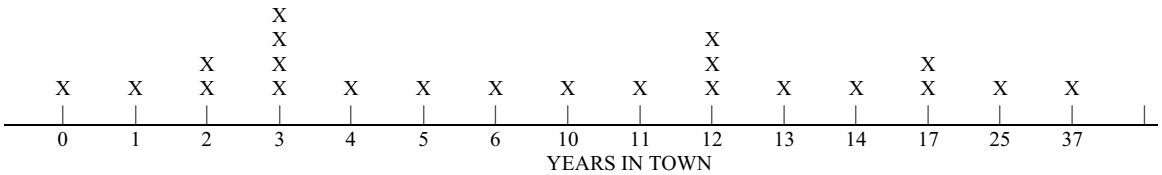
Code 2	"You can get to school by it."
Code 1	Direct interpretations, Geographical / historical assumptions, likes & dislikes
Code 0	Misinterpretations / Idiosyncratic, NR

TRV6. Tom is not at school today. How do you think he will get to school tomorrow? Why?

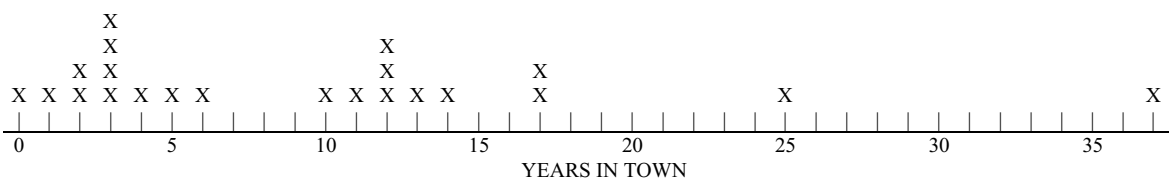
Code 5	"Probably by bus – Because 1/3 of the children caught it today." (uncertainty stated)
Code 4	Gendered or non-gendered majority; Bike, majority of boys; bus or walk are more common.
Code 3	Balancing using train or other transport; anything can happen
Code 2	Placing Tom (patterns) / Finding Tom
Code 1	No variation: Same as yesterday; Not enough information: Can't tell
Code 0	Misinterpretation, NR

TWN. A class of students recorded the number of years their families had lived in their town. Here are two graphs that students drew to tell the story.

Graph 1



Graph 2



TWN1. What can you tell by looking at Graph 1? – (2 spaces provided)

Code 3	2 summary comments; e.g., "Well, some people haven't lived there long" & "Some people have lived up to 17 years"
Code 2	1 summary plus perhaps data reading
Code 1	2 data reading comments (e.g., "There is only one in column 1" & "Column 3 has four crosses") or one appropriate and one inappropriate comment
Code 0	Graphing or idiosyncratic (e.g., "The graph is very spacey")

TWN2. What can you tell by looking at Graph 2? (2 spaces provided)

Code 3	2 summary comments
Code 2	1 summary plus perhaps data reading
Code 1	2 data reading comments or one appropriate and one inappropriate comment
Code 0	Graphing or idiosyncratic

TWN3. Which of these graphs tells the story better? - Why?

Code 3	"Graph 2 – You can see the difference between years more clearly and the graph is more spaced out"
Code 2	Indifference ; Personal preference; lack of logical reasoning
Code 1	Focused on graph spread / lay out; personal preference
Code 0	Statistically inappropriate choice with inappropriate or no reasoning, NR

VAR. What does "variation" mean? Use the word "variation" in a sentence. Give an example of something that "varies".

Code 3	"Varying is when something doesn't stay the same all the time – it varies" "That dress is a variation of the one I bought here last summer" "Clothes vary"
Code 2	More sophisticated definition with inappropriate sentence usage, or Simple but clear understanding reflected in definition
Code 1	Definition attempted, or Example given only with confused definition
Code 0	Idiosyncratic / Tautological, NR