Journal of the Korea Society of Mathematical Education Series D:
Research in Mathematical Education
Vol. 10, No. 4, December 2006, 259–288

韓國數學敎育學會誌 시리즈 D:
<數學敎育硏究>
제 10 권 제 4 호 2006 년 12 월, 259–288

Development of Numerical and Graph Interpretation Skills — Prerequisites for Statistical Literacy[1]

Watson, Jane M.
Faculty of Education, University of Tasmania, Private Bag 66, Hobart, Tasmania 7001, Australia; E-mail: Jane.Watson@utas.edu.au

Kelly, Ben A.
Faculty of Education, University of Tasmania, Private Bag 66, Hobart, Tasmania 7001, Australia

(Received December 12, 2006)

This study considers the performance of students in Grades 5 to 10 on four tasks assessing students' ability to evaluate data presented in numerical form, for example, in a list or table, or in graphical form, for example, in a frequency graph or scatter graph. The ability to tell a story from data or a graph is an important aspect of statistical literacy. The samples provide the opportunity to consider the association of two pairs of items, one from each type of interpretation, numerical and graphical. Educational implications for the outcomes and the classroom use of the items are considered.

Keywords: Graphs, Numerical Data, School Students, Statistical Literacy, Survey Data
ZDM Classification: K40, C30
MSC2000 Classification: 97C30

DEVELOPMENT OF NUMERICAL AND GRAPH INTERPRETATION SKILLS — PREREQUISITES FOR STATISTICAL LITERACY

Various skills are required when data are involved, often associated with the form in which the data are presented. Sometimes the requirement is to interpret raw data, or to translate data into a graphical form, or to interpret a given graph. At times the task is to create a graph to fit a verbal description or to speculate about what type of data set may have created the outcome. In the area of statistical literacy (Gal 2002) it is more than likely that people are required to interpret rather than create data sets and graphs and it is

[1] This research was funded by the University of Tasmania Internal Research Grant Scheme, grants no. W0011912 & W0014364. Annaliese Caney coded the data.

this type of task that this study considers. Three of the four questions employed involve presentations that might be found in the media, whereas the fourth presents a small data set to explore understanding of the basic statistical averaging tools that could be used to summarise a data set.

BACKGROUND AND LITERATURE

Although the school curriculum emphasises the complete process of a statistical investigation from collecting data to reaching a conclusion (National Council of Teachers of Mathematics 2000), once students leave the classroom for everyday life they are unlikely to carry out a complete investigation. More likely they are given a verbal statement, a graph, or a data presentation, probably in a table, out of which to make sense. In the first case, belief in the validity of the verbal statement may be the issue, with no other representation to consider. In the other two cases, however, it is necessary to translate the visual presentation involving numbers into a verbal statement, even if this is only kept in the mind rather than stated orally.

Gal (2002) addressed the issue of statistical literacy in everyday life when he laid out the requirements for adults in the following terms:

(a) 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) 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 critical skills of interpreting, evaluating, and communicating are assumed to take place in a myriad of contexts that are encountered in everyday life. As such they become important skills to develop during the school years, especially in a reform-based curriculum environment (*e.g.*, Department of Education Tasmania 2002) where essential learning elements are cross-disciplinary and "thinking" based on inquiry and reflection is the central core of the curriculum.

Assuming that the skills expected by Gal (2002) can be incorporated across many of the traditional curriculum areas, such as social science and science, as they are integrated into essential elements of the curriculum, the issue of assessment then arises. Assessment of statistical literacy can be built into the requirements for large scale investigations and

inquiries that students undertake. Systems, however, often have requirements for the assessment of skills across the board, for example in setting numeracy benchmarks across grade levels (*e.g.*, Watson 1998). The development of a battery of relatively short items that provide varying contexts for the assessment of statistical literacy skills has hence been a priority of statistics educators at the school level. Watson & Callingham (2003) began with a large set of 80 items from several studies and described a statistical literacy construct with six developing levels of understanding from "idiosyncratic" to "critical mathematical." The highest level incorporated the skills required by Gal for the statistically literate adult. To develop a useable instrument across grade levels, Callingham & Watson (2005; see also Watson & Callingham 2005) chose and adapted items from earlier instruments and tested two parallel forms in classrooms from Grade 5 to Grade 10. Their overall studies reported confirming the six levels of statistical literacy but for some of the items developed details of the rubrics and internal incremental understanding displayed were not provided. This paper is an opportunity to explore four items, two from each of the survey forms, in depth and consider potential classroom use in the development of statistical literacy skills at school level.

A brief summary of previous research related to the four items employed in this study is presented in order to set the scene for the research questions. The topics are students' understanding of measures of average, of information presented in two-way tables, and of information presented in graphical form.

Average

Student performance on tasks associated with calculating and understanding measured centre was one of the early foci in statistics education. This began with a basic study of school students' ability to handle mean, median, and mode (Goodchild 1988) and consideration of tertiary students' ability to calculate weighted averages (Pollatsek, Lima & Well 1981), and progressed to a study of abstract features of the mean (Mevarech 1983), of working the algorithm backward (Cai 1995), and of students' interpretation of the meaning of averages (Mokros & Russell 1995). Watson & Moritz (2000) provided a model of developing understanding supported by longitudinal data from students three or four years after initial assessment. Issues of procedural and conceptual knowledge (Hiebert & Carpenter 1992; Skemp 1986) and their relationship have featured across these research studies.

Two-way tables

Although basic reading and interpreting of two-way tables has been included in some studies of statistical literacy (*e.g.*, Watson & Callingham 2003), few individual studies

appear in the literature. Watson (1998) considered table reading and interpretation, for example, comparing or adding cell values as numeracy skills, and reported on performance by Grade 3 and 5 students. Success rates for reading single cell values and summing were 89% or higher, whereas for comparing two cell values for equality they dropped to 55% and 80% for Grades 3 and 5, and for comparing two sums of multiple cells they rose to 71% and 81% respectively.

Batanero; Estepa; Godino & Green (1996) considered two-way tables with conditional information and questions about association of variables. A complex classification of outcomes was presented, with a total of 16 different strategies observed, including three strategies identified as correct, 8 strategies identified as partially correct, and 5 strategies identified as incorrect in obtaining an appropriate response. Estepa; Batanero & Sanchez (1999) used similar two-way tables, one with a relationship and one without a relationship between the variables. Basically, the classification process was the same as for Batanero et al. with 14 levels related to strategies. Four strategies were identified as correct, four were identified as partially correct, and six were identified as incorrect in obtaining an appropriate response to the items presented. Except for the overall correctness criteria, the categories resulting from these classification procedures were not hierarchical. The two-way table item is this study was adapted from Batanero *et al.* (1996)

Graph interpretation

From the early work of Swan and his colleagues (*e.g.*, Bell, Brekke & Swan 1987) and later work of Curcio and her colleagues (*e.g.*, Curcio 1987; Friel, Curcio & Bright 2001), interest has grown in students' appreciation of specific types of graphs and their ability to read and interpret information from them. Watson and Kelly (2003) considered student interpretation of pictographs, following the work of Pereira-Mendoza & Mellor (1991). Pie graphs have been studied by Watson (1997) and scatter graphs by Moritz (2003; 2004). One of the issues in dealing with a structured curriculum is the complexity of graphical representations and their potential value in everyday settings. Both pie and scatter graphs are often seen in the media but due to the traditional view that they are relatively difficult to construct, they may not be introduced into the curriculum until well into the middle school years. One of the tasks used in the current study contains actual data from a study and also portrays a non-intuitive direction shown in a scatter graph similar to that used by Moritz.

Another issue that has not been explicitly associated with graph interpretation until recently is the appreciation of variation. Moore (1990) brought the omnipresence of variation to the attention of statistics educators and since then others have called for

research on students' understanding (*e.g.*, Green 1993). Students' ability to display variation in graphs they create has been considered in recent research (*e.g.*, Kelly & Watson 2002; Watson & Kelly 2005) but again there has been less attention directed at interpreting the variation shown in graphs produced by others. The other graph task discussed in this study follows the work adapted by Garfield & Chance (2000) who devised tasks specifically to assess students' appreciation of variation displayed in frequency graphs.

Conceptual and procedural knowledge

One of the concerns of mathematics educators when assessing students for benchmarking exercises, is the testing of both procedural and conceptual knowledge (Hiebert & Carpenter 1992; Skemp 1986). Procedural knowledge is that which can be organized in steps, often associated with algorithms, and carried out either with or without an understanding of the mathematical concepts underlying the procedures. An example from the current study is the algorithm used to calculate the arithmetic mean. Students may be able to use the "add-up-and-divide" algorithm to calculate a mean without having an appreciation that it is a value within the range of the data presented that represents the data set. They may hence be able to obtain the correct answer to a "Find the mean ..." problem but not be able to explain why it is not one of the elements in the set or what is the influence of an outlier.

Conceptual knowledge is that which includes relationships among the mathematical ideas involved and shows an understanding beyond the capability to carry out a procedure. In the context of this study for example, a student may have a conceptual understanding of the importance of central tendency in representing a data and further be able to distinguish the characteristics of the mean and median that make them appropriate measures for particular data sets. Although procedural knowledge is important, the major focus of the items in this study is the conceptual knowledge that is necessary to develop the critical interpreting, evaluating, and communicating skills required by Gal (2002). Arguments for or against data presented from external sources can only be made with an understanding of the underlying concepts. In this study they relate not only to measures of central tendency but also to proportional reasoning, variation, and the meaning of two types of graphical representation.

RESEARCH QUESTIONS

Although much previous research has taken place on these topics, the current study is interested in testing items that can be easily administered to large groups of students and

which acknowledge the importance of both procedural and conceptual knowledge for statistical literacy. For the items introduced in Figure 1, the following questions are asked.

1. What levels of development are observed for interpretation of numerical data?
2. What levels of development are observed for interpretation of graphical data?
3. What association if any exists between the levels of development observed for the two types of questions?

METHOD

Sample

The sample of students consisted of 673 students in Grades 5 to 10 from five Catholic schools in the Australian state of Tasmania. The number of students per grade is shown in Table 1.

Table 1. Number of Students per Grade

G5	G6	G7	G8	G9	G10	Total
136	123	112	98	105	99	673

Tasks

The tasks used in the study are given in Figure 1. The average items were adapted from those used earlier (Watson & Moritz 1999; 2000) in order to canvass opinions on all potential measures of average, not just those that students would choose to use on their own. The item with the 2 x 2 table asking about the relationships between smoking and lung disease was adapted from Batanero *et al.* (1996). The problem involving two graphs depicting the height of students in two schools was adapted from Garfield and Chance (2000, Figure 8), and the homework item was based on information presented in a graph in the Third International Mathematics and Science Study (TIMMS) report (Lokan, Ford & Greenwood 1996, p. 169). Using a slightly different scoring rubric the items were included in the comprehensive analysis completed by Callingham & Watson (2005) and Watson & Callingham (2005). The items were not detailed there, however, to show the progression of development of students across grades.

Average Item

Nine students in a science class weighed a small object separately on the same scales. 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

(a) The students had to decide on the best way to summarise these values. Ben said, "I'd use the most common value to get the mode. That's 6.3." Is Ben's way a good way to summarise the information? Explain your answer.

(b) Jane said, "I'd put them in order and use the middle value to get the median. That's 6.2." Is Jane's way a good way to summarise the information? Explain your answer.

(c) Ron said, "I'd add them all up and divide by 9 to get the mean. That's 7.18." Is Ron's way a good way to summarise the information? Explain your answer.

(d) May said, "I'd leave out 15.3 and use the mean of the others. That's 6.17." Is May's way a good way to summarise the information? Explain your answer.

(e) Which of the ways described above would you use? Why?

2 x 2 Table Item

The following information is from a survey about smoking and lung disease among 250 people.

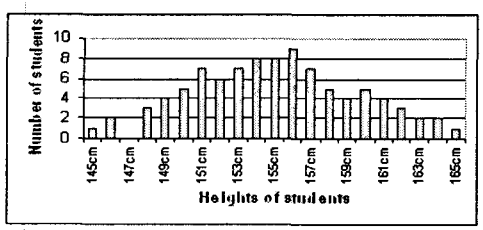
	Lung disease	**No lung disease**	**Total**
Smoking	90	60	150
No smoking	60	40	100
Total	150	100	250

Using this information, do you think that for this sample of people lung disease depended on smoking? Explain your answer.

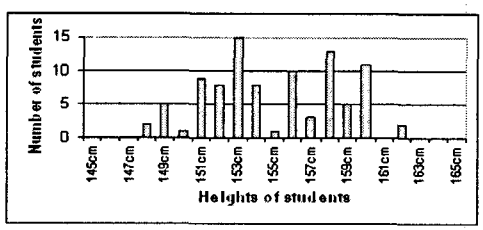
Height Item

The following graphs describe some data collected about Grade 7 students' heights in two different schools.

School A



School B



(a) How many students are 156 cm tall in each school? School A _____ School B _____

(b) Which graph shows more variability in students' heights?

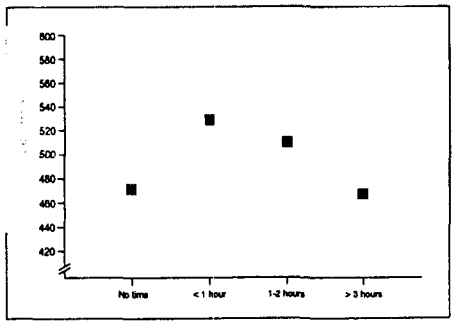
(c) Explain why you think this.

Figure 1. Items used in survey

Homework Item

A survey of Grades 7 and 8 students produced the information shown in the graph below. It shows the students' maths scores and the amount of time they spent on maths homework each day.

Hours per day spent on maths homework



(a) What is the maths score for students who spend 1-2 hours per day on mathematics homework?
(b) What does the graph tell you about maths homework time and maths scores?
(c) Gives some reasons why the graph has the shape it does?

Figure 1 (Continued). Items used in survey.

Analysis

The rubrics for scoring the four items were devised by the first author and Callingham based on structural complexity and statistical appropriateness. The SOLO model of Biggs & Collis (1982) played a role in the development of the rubrics. This model, derived from the neo-Piagetian theory, suggests that the structure of responses is likely to include elements relevant to the problem and its solution in the following fashion. Prestructural responses consider no such elements. Unistructural responses consider a single element with no linkage to other elements noted. Multistructural responses include two or more elements usually in sequence, whereas Relational responses link together multiple elements to obtain closure for the problem. In the area of interest for the items in this study, relational responses are likely to demonstrate both procedural and conceptual knowledge and how they are linked together.

The rubric for the Average item considers statistical appropriateness as well as structural complexity and as can be seen in Table 2, the lowest three levels of response are basically single statements of relatively more statistical relevance with increasing level, whereas at Level 3 procedural knowledge is accurate but not conceptually based as stated, and at Level 4 responses express contrasting evidence in an appropriate fashion

based on conceptual understanding.

Table 2. Rubric for the Average Item

	Explanation Parts (a), (b), (c), (d)	Explanation Part (e)
Level 0	No response No reason or idiosyncratic reason	No response No reason or idiosyncratic reason for preferred method
Level 1	Methodological choice without statistical reason Tautological, positive evaluation (except (d)) Recommendation or preference for other methods Preference expressed for method without reason ((b), (d))	Methodological implications Personal preference Tautological argument
Level 2	Claims of accuracy of method but with no statistical reason (except (d)) Claims of inaccuracy of method but with no statistical reason Tautological, negative evaluation ((d))	Reasons based on claims of accuracy without a statistical argument
Level 3	Statistical negative evaluation Statistical positive evaluation (except (d)) Statistical uncertainty – positive evaluation with an element of doubt ((d))	Statistical argument
Level 4	Statistical and contextual response incorporating both positive and negative aspects of method (focusing on outlier for (c) and (d))	Reasons based on statistical and contextual information

The rubric for the Two-way Table Item reflects both structure and appropriateness in two successive sets of three levels. The first three show an increasing appreciation of the context of the problem, from none, to personal belief, to a questioning of statistical aspects of the investigation. Although Level 2 responses could be considered conceptually based with respect to the context of the item, they do not answer the question asked. The second three levels then address the specific question of association using the data in the table. Level 3 responses are unistructural in nature using only one element (cell) from the table in support of a conclusion. Level 4 responses are multistructural in employing two (or three) cells, whereas Level 5 responses are relational in employing all cells to reach a conclusion of no association between smoking and lung disease. These Level 5 responses demonstrate a conceptual understanding of the proportional reasoning required for the task.

Table 3. Rubric for the Two-way Table Item about Smoking and Lung Disease

	Explanation
Level 0	No response No reason Positive or negative statement with no justification
Level 1	Positive or negative statement justified with knowledge of the content area, but not based on the data Agreement or disagreement with statement, stating more or less chance, but without explicit reference to the data
Level 2	Critical analysis and limitations of the assumed survey method and data collection techniques in the study
Level 3	Agreement or disagreement with statement citing evidence from a single cell in the table
Level 4	Agreement or disagreement with the statement, citing evidence from the table focusing within, between or across columns (2 or 3 cells)
Level 5	Critical examination of all information presented in the table (including all cells) and/or correct statement of proportion and/or percentages

The rubric for scoring the Height item was devised by the first author and Callingham based on structural complexity or statistical appropriateness. For Part (a) of the Height Item, considered to be procedural in terms of graph reading, three codes are based on whether the student answered incorrectly, Level 0 (incorrect values or no response); partially correctly, Level 1 (School A = 9 or School B = 10); or correctly, Level 2 (School A = 9 & School B = 10). For Part (b), the rubric again distinguishes the ability of students to identify the school with the most variation in heights.

At Level 0, responses are idiosyncratic, at Level 1 students respond with School B or state that both schools have a degree of variation (*e.g.*, "the same"), and at Level 2 students respond with School A. The rubric for Part (c) of the Height item, asking for an explanation for the response to Part (b), is given in Table 4. An increasing focus on the information in the graphs is shown in the levels of the rubric, with the final two making the right choice but Level 4 reflecting an explicit consideration across specific features of the two graphs. These two highest levels demonstrate increasing conceptual understanding of variation and its representation in graphical form.

Table 4. Rubric for Part (c) of the Height Question with for Students in Two Schools

	Explanation (Part (c))
Level 0	No response / No reason Misreading of data, misinterpretation of question, or unjustified statements Personal preference for graph usually based on aesthetic appearance Inappropriate focus on the context of the data
Level 1	Misapplied notion of variability or focus on an average height or "more" data
Level 2	Focus on the number of individual bars in either graph with no regard for what the bars represent Focus on the size of the individual bars in either graph with no regard for what the bars represent
Level 3	Implicit mention of the wide range of heights
Level 4	Explicit mention of the wide range, spread and/or the variety of heights of students

In the Homework item, Part (a) asks students to read the graph and determine what the mathematics score is for students doing 1–2 hours of homework. This procedural task is treated as a right or wrong response with two levels of coding. At Level 1, responses are correct in the range 510–520; Level 0 includes any other response apart from the range specified. The rubrics for Parts (b) and (c) are presented in Table 5. For Part (b) increasing levels of response reflect taking into account more of the information provided in the graph. At Level 1 only one variable is mentioned, whereas at Level 2 both variables are noted but no association or an inappropriate association is suggested. At Level 3, causation between variables is assumed, whereas at Level 4 the appropriate association is interpreted from the graph without assumption. The conceptual understanding here is related to the reporting of an observed association without assuming a cause. For Part (c) similar levels are identified in terms of developing a hypothesis about the information from the graph, but only four levels are identified. The conceptual understanding of forming a hypothesis is shown at the highest level.

A research assistant scored responses for all four questions based on the rubrics in the text and Tables 2 to 5. Results for each item are presented by grade and level of response for the four questions, two associated with each of Research Questions 1 and 2. For the Average, Height, and Homework questions the parts of the question are tallied to give a Total score for each. For some of the items, associations between pairs of parts are considered with indicative correlations. For the pairs of items where the same group of students completed both, associations are also considered with indicative correlations in relation to Research Question 3.

Table 5. Rubric for Parts (b) and (c) of the Homework Item with Mathematics Scores in a Line Plot Graph

	Explanation (Part (b))	Explanation (Part (c))
Level 0	No response Idiosyncratic response Misinterpretation of data	No response Idiosyncratic response Comment relating to the type of graph and/or appropriateness
Level 1	Single focus response from the graph	Comment made about the variables but with no hypothesis stated
Level 2	No association identified between the variables Incorrect association identified between the variables	Comment involving data reading but with no clearly identifiable hypothesis
Level 3	Causation between the variables assumed	Creation of a hypothesis from the graph based on student abilities or other possible contributing factors influencing the data
Level 4	Appreciation of association between variables through basic data reading Appreciation of association between variables with implications and conclusions drawn from the data	NA

RESULTS

Research Question 1: Levels of development for interpretation of numerical data

Table 6 below shows the number of students in each grade that responded to each of the Average questions relating to the mode, median, mean including outlier, and mean excluding outlier as being a good way to summarise the information.

The percents of students responding at Level 4 on any of the parts of the Average Item were extremely low with the greatest percent being 7% of responses to Part (d). There is a decrease in the number of Level 0 responses with increasing grade, which is encouraging, however, there is a low percent of Level 4 responses to Parts (a) and (b) of this item, relating to the use of the mode and median. The average level of response generally increased over the range of grades but there was a plateau across the middle years. That the mean scores for Parts (c) and (d) were higher than for Parts (a) and (b) may reflect an emphasis in the curriculum.

Table 6. Number of Students per Grade and Level for Each Part of the Average Item

Part (a) Mode

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	29	10	3	7	0	49	0.76
G6	49	21	8	15	1	94	0.91
G7	23	14	5	7	0	49	0.92
G8	22	16	9	4	0	51	0.90
G9	21	12	16	13	1	63	1.38
G10	18	8	5	14	0	45	1.33
Total	162	81	46	60	2	351	1.03

Part (b) Median

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	33	12	2	1	1	49	0.47
G6	60	21	6	7	0	94	0.57
G7	26	15	6	2	0	49	0.67
G8	34	8	5	4	0	51	0.59
G9	24	18	16	4	1	63	1.05
G10	16	13	9	5	2	45	1.20
Total	193	87	44	23	4	351	0.74

Part (c) Mean including outlier

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	22	7	6	4	0	49	0.63
G6	52	23	3	15	1	94	0.83
G7	24	6	3	14	2	49	1.27
G8	22	8	7	12	2	51	1.29
G9	24	7	7	20	5	63	1.60
G10	13	3	7	11	11	45	2.09
Total	167	54	33	76	21	351	1.23

Part (d) Mean excluding outlier

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	36	2	6	4	1	49	0.61
G6	63	8	16	4	3	94	0.68
G7	27	5	12	4	1	49	0.92
G8	31	3	10	5	2	51	0.90
G9	30	2	16	11	4	63	1.32
G10	14	7	2	8	14	45	2.02
Total	201	27	62	36	25	351	1.02

Examples of responses for Levels 1–4 in Parts (a), (b), (c), and (d) of the Average Item are presented in Table 7. Level 0 responses were either idiosyncratic or were "non-responses" by students. Level 0 responses tended not to take into account any relevant idea of the task presented.

Table 7. Examples of Response for Average Items Parts (a), (b), (c), (d)

Level and Description	*(a) Mode*	*(b) Median*	*(c) Mean including outlier*	*(d) Mean excluding outlier*
Level 1 – *Methodological choice without statistical reason*	"Ben's is a good way to summarise his information it's not confusing."	"No because that is too long."	"Yes I think Ron's idea is a good one because it's easy to do."	"I don't like this way. Because you not checking what you're doing as much."
Tautological, positive evaluation	"I think it's a good way because it was the most common and it's the most likely one."	"That would work well because the median is the middle number."	"I think this is good because you are adding and dividing to get your answer."	
Recommendation or preference for other methods	"It's a pretty good answer but he should find an average."	"No because you need to get the average answer."	"I think Ben's is the best because it is a lot easier and this one is too confusing."	"No, it's just as bad as Ben's and Jane's solution."
Preference for method but without reason		"That is better than Ben's but still not 100%."		"This way is better than Ron's."
Level 2 – *Claims of accuracy, no statistical reason*	"It is a good way because it's fairly accurate."	"Yes because it is more accurate."	"It is a good way because you get an exact answer."	
Claims of inaccuracy, no statistical reason	"No because if they kept weighing there could have been a different common value."	"Not really because she is using the middle one and she doesn't know that it is the correct one."	"No you would get the wrong answer."	"No, not accurate."
Tautological, negative				"No she has missed 1 out."

(Cont.)

Table 7(Cont.). Examples of Response for Average Items Parts (a), (b), (c), (d)

Level and Description	*(a) Mode*	*(b) Median*	*(c) Mean including outlier*	*(d) Mean excluding outlier*
Level 3 – *Statistical negative evaluation*	"No because one answer is way out."	"This is not so good. It only shows one score and does not involve all scores."	"No because you will get a higher answer than you are supposed to."	"No because 15.3 is a result and should be used when summarising the values."
Statistical positive evaluation	"Yes because it is the majority."	"Yes as it would use all the info and still be correct."	"Yes because he is getting the average of the measurements which includes all the answers."	
Statistical uncertainty, positive evaluation				"Yes and no, 15.3 could be truthful or made up but she should use all answers."
Level 4 – *Statistical and contextual response incorporating positive and/or negative aspects of method*	"Yes, it would be close to the objects actual value and doesn't take into account the 15.3. It still would be inaccurate though."	"This is a good way, as it does not make too much use of the very high score of 15.3."	"No, because there is a 15, which is most likely a mistake."	"Yes because 15.3 is an outlier."

Table 8. Number of Students per Grade and Level for the Final Part of the Average Question

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	35	10	0	4	0	49	0.45
G6	49	28	4	11	2	94	0.82
G7	25	10	9	5	0	49	0.88
G8	20	14	9	7	1	51	1.12
G9	24	14	12	11	2	63	1.25
G10	13	8	5	10	9	45	1.87
Total	166	84	39	48	14	351	1.03

Table 8 shows the number of students in each grade that responded to Part (e) of the Average question relating to which method presented earlier (mode, median, mean

including outlier, or mean excluding outlier) the student would use to summarise the information. For this part the average score increases monotonically with grade.

At Level 0 students could select a method they would use to summarise the information, however, had trouble justifying their choice, did not see any reason to do so, or supplied idiosyncratic reasons with no apparent logic.

- Jane's. [Grade 7]

- D (May), because it is good. [Grade 9]

At Level 1, responses were somewhat more articulate but lacked statistical ideas, instead relying on personal preferences of which is better, stating basic methodological implications of performing the task, or simply restating the method thereby providing a tautological response.

- I would use Ben's because I like the way he did it. [Grade 5]

- I like Ron's idea because it's easier. [Grade 6]

- I would use Ben's the most common mode. [Grade 6]

Level 2 responses, although lacking an explicit statistical argument, seemed to imply some knowledge of the area, with students claiming the methods they would use to summarise the data were more "accurate" or "exact." An implicit appreciation seemed evident.

- Ben's because it's more accurate. [Grade 8]

- Jane because it is a more precise way. [Grade 7]

Students responding at Level 4 provided responses that were both statistical in nature and took into account the context of the problem. Instead of reciting learnt algorithms, applying the mean, or ordering data, responses reflected more critical thinking relating to the context of the question (*i.e.*, weighing an object).

- The best way is D (May), as it is the most logical way, to get the answer as 15.3 was obviously a mistake when we see the other answers. [Grade 10]

- I would use D, as May is going to leave out 15.3. This result is most likely to be a mistake, and should be classed as an outlier. That leaves only similar results, so the mean would be more accurate. [Grade 10]

The scores across all five parts of the Average questions were summed to make a Total score for this item. The total possible score was 20. Table 9 shows the number of students with each possible Total score, as well as the average score for each grade. It is seen that the average score rose monotonically but there was little change over Grades 7 and 8. (See Table 9 on p. 275.)

Table 9. The Distribution of Students per Grade for the Total Score over the Average Items.

Level / Grade	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	Total *n*	Average Score
G5	23	2	4	3	4	2	3	3	2	0	0	0	0	1	1	0	1	0	0	0	49	2.92
G6	28	12	8	6	8	2	6	7	1	7	2	1	3	1	0	0	0	1	1	0	94	3.82
G7	12	4	3	6	2	3	4	3	1	1	2	5	1	0	2	0	0	0	0	0	49	4.65
G8	12	6	3	2	3	2	2	5	3	4	4	2	3	0	0	0	0	0	0	0	51	4.80
G9	14	2	5	1	3	0	3	2	5	4	6	6	7	3	1	0	0	0	1	0	63	6.60
G10	7	1	2	2	1	1	2	2	2	4	5	3	1	1	1	4	2	2	1	1	45	8.51
Total	96	27	25	20	21	10	20	22	14	20	19	17	15	6	5	4	3	3	3	1	351	5.05

Although a relatively large percent of Level 0 responses were recorded for each part of this item (46%, 55%, 48%, 57%, and 47%, respectively over the 5 items), 71.5% of students had a Total score above 0 indicating that most students attempted or gave a non-idiosyncratic response to at least one of the Average items. Only 21.5% of students scored between 10 and 20, with no student scoring the maximum possible.

The ten correlations between the five parts of the Average question are given in Table 10. They are relatively strong as would be expected. The amount of variance explained ranged from 20% to 41% for the pairs of parts.

Table 10. Indicative Correlations and Percent of Variance Explained between Parts of the Average Question

	Part (a)	Part (b)	Part (c)	Part (d)	Part (e)
Part (a)	1	41%	28%	24%	20%
Part (b)	0.638	1	33%	30%	29%
Part (c)	0.533	0.571	1	28%	36%
Part (d)	0.488	0.545	0.533	1	26%
Part (e)	0.451	0.537	0.599	0.511	1

Table 11. Number of Students per Grade and Level for the Two-way Table on Smoking and Lung Cancer

Level Grade	0	1	2	3	4	5	Total (*n*)	Average Score
G5	31	28	0	7	1	0	67	0.79
G6	10	12	0	4	2	1	29	1.28
G7	20	22	1	8	10	2	63	1.56
G8	16	18	1	6	6	0	47	1.32
G9	8	14	2	5	13	0	42	2.02
G10	10	16	5	8	14	1	54	2.06
Total	95	110	9	38	46	4	322	1.47

Table 11 shows the numbers of students in each grade and level who responded to the Two-way Table item asking about an association between smoking and lung cancer. For this question there was a lower percent of Level 0 responses than for the previous question. Again there was an increase in average score with grade level except for Grade 8, and a plateau at Grade 10.

At Level 0 responses generally provided no reason for the claims made. Some students, however, provided a general statement with no justification in relation to the question. These students seemed not to understand the intent of the task.

- I don't understand, but I hate people who smoke, cigarettes and smoking, I would hate to smoke. Yuk. [Grade 5]

- I think you should not smoke. [Grade 6]

At Level 1 responses were somewhat more articulate but still seemed to have difficulty comprehending what the task was asking. General knowledge of the content area (i.e., smoking and lung disease) at this level interfered with reasoning, with many students supporting their claim based on prior knowledge of what they knew about smoking and lung disease and not from the data provided. Some students may have considered the data and used limited chance language in supporting their claims (*e.g.*, more/less chance), however, there was no explicit reference made to the table or data, and therefore this assumption about responses not made.

- Yes because smoking is not good for you. [Grade 5]

- I would say smoking because I've read some books and they say that a main cause of lung disease is smoking. [Grade 6]

- Well you'd have more chance of getting lung disease if you smoke but if you don't you've still got some chance of getting it but not as much. [Grade 5]

- No not really but smoking can increase chances of forming lung disease. [Grade 7]

Level 2 responses showed an appreciation of survey methods by providing critical analysis of the methods employed and pointing out limitations to the possible method used to collect the data presented in the table. Often, responses focused on the unequal sample size of smokers to non-smokers not recognising the need to convert the data into percent for a fair comparison.

- Sort of. They need to use the same amount of people in each of the categories otherwise its not really fair. [Grade 9]

- I don't think that this sample is sufficient because there are more smokers than non-smokers. [Grade 10]

At Level 3, students appeared to extract data from the table in order to make a decision and support their claims. Responses at Level 3 looked at only one cell when doing this, therefore supporting the claims based on limited evidence. Student responses tended to either agree or disagree with the statement. At Level 3 it appears students were beginning to appreciate the intent of the task.

- No because more people that didn't smoke got lung disease. [Grade 6]

- Yes because it is a lot higher than no smoking. [Grade 7]

Level 4 responses took the analysis one step further citing evidence for the claims by focusing within, between or across columns. Once again, students tended either to agree or disagree with the statement and support this claim by looking at two or three cells.

- No because 60/150 of smokers didn't get lung disease and 40/100 no smokers didn't get lung disease. There is a lot of people that get lung disease that don't smoke. [Grade 7]

- Yes because with smoking people who had lung disease was 60. Non-smokers who didn't have

the disease was dropped to 40. Smokers who had lung disease was up at 90 while non-smokers was dropped to 60. [Grade 8]

Level 5 responses tended to examine critically all the information supplied in the table and correctly state proportions, ratios or percents to back up their claims. These students disagreed with the statement.

- No lung disease doesn't depend on smoking because it is not at a higher rate, 90-60 is 2/3 and 60-40 is 2/3. [Grade 6]
- No because the figures look bigger but there were more smokers used for this information. So if you take into account the more people there are the figures are even. [Grade 7]

Research Question 2: Levels of development for interpretation of graphical data

The item relating to the Heights of students in two schools depicted in two graphs had three parts (see Figure 1). In Part (a) most students could respond appropriately, reading from the graphs how many students in each school were 156cm tall. Seventy-eight percent of students overall were able to answer correctly. The highest rate of incorrect response for this item was in Grade 5 where 32% of students answered either incorrectly or only partially correctly giving only one correct response (*e.g.*, for School A or B, not both). Similarly, Part (b) was also answered appropriately by the majority of students with 62% overall giving the appropriate response of School A having more variability in students' heights. Thirty percent of students answered inappropriately suggesting they were either "both the same" or School B displayed greater variation. Only 7% of students failed to understand the term variability by giving idiosyncratic responses or no response this item. The percent correct varied from 54% in Grade 10 to 72% in Grade 5.

Table 12 below shows the number of students in each grade and level that responded to Part (c) of the Height item asking students to explain their reasoning for choosing the school with the most variability. The average scores for Part (c) fluctuate and it should be noted that the Grade 6 sample for the Height item is quite small. The decreasing ability with grade to explain may be related to the increasing knowledge of graphing interfering with an appreciation of the type of variation displayed.

Table 12. Number of Students per Grade and Level for Part (c) of the Height Question

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	34	29	13	9	2	87	1.03
G6	2	5	9	10	3	29	2.24
G7	22	9	18	11	3	63	1.43
G8	22	7	12	5	1	47	1.06
G9	16	7	12	4	3	42	1.31
G10	21	12	11	5	5	54	1.28
Total	117	69	75	44	17	322	1.30

At Level 0 responses did not acknowledge the intent of the task and failed to mention any type of variation in either graph. Responses were likely to focus solely on the content of the data without regard for variation, to focus on the aesthetic appearance of the graph, quite often citing personal preferences for lay-out, or to misread the data or misinterpret the question giving unjustified and irrelevant statements.

- Because School B has 10 people and because it goes up in fives. [Grade 5]

- (School B) Because it's more spaced out and easy to read and not so crowded like School A. [Grade 10]

- (School B) Because people mightn't have had their growth spurts. [Grade 7]

At Level 1 responses misapplied the notion of variability and instead focused on the average height or "more."

- Because School A has more pupils than School B. [Grade 6]

- Because in school A, a lot of people are around the same height but in School B they are not. [Grade 5]

Level 2 responses focused on the bars within each graph, but not explicitly on the variability within the data. Responses focused either on the numbers of bars within each graph or the sizes of the bars within each graph, but without any acknowledgement of what they actually represented.

- Graph A has more shaded in bars. [Grade 6]

- (School B) Because there are many ups and downs in the graph. [Grade 8]

At Level 3, students chose School A and implicitly acknowledged the number of different heights of students in School A compared to School B.

- By looking at the graphs I think it's A because there are lots of different heights. [Grade 5]

- (School A) Because there are more recorded heights. [Grade 7]

Level 4 responses made explicit mention of the wide range, spread or variety of heights within the context of the graphs in choosing School A. Some students went as far as to compare and contrast the data within each graph with the other to support their claims of more variation.

- Because School A has a large variety of heights and School B does not have many. [Grade 5]

- Because school A's heights vary from 145 cm to 165 cm and school B only varies from 148 cm to 162 cm. [Grade 7]

- Because on Graph A they have every height on the graph except for 147 cm, but on graph 2 there is more than 1 height that misses out. [Grade 6]

The percent of students per grade responding at Level 4 fluctuated with increasing grade, from 2% in Grade 8 to 9% in Grade 10 and 10% in Grade 6 (a small sample).

Overall there was no clear improvement with grade for Part (c), which may reflect students' lack of experience with consideration of variation shown in graphical representations over the middle school years.

The scores across all three parts of the Height questions were summed to produce a Total score for this question. The total possible score was eight. Table 13 shows the number of students in each grade with each possible Total score. Although 36% of students responded at Level 0 for Part (c) only 3% of students had a Total score of 0 or 1 over the three parts indicating that most students attempted at least one of the easier questions (Part (a) or (b)). Thirty-eight percent of students scored between 5 and 8, with 7% of students scoring the maximum possible total of eight. Again there is no overall trend for improvement in average score over the six grades.

Table 13. The Distribution of Students per Grade for the Total Score over the Height Question

Level Grade	0	1	2	3	4	5	6	7	8	Total (*n*)	Average Score
G5	0	3	11	19	19	17	8	8	2	87	4.17
G6	0	0	2	4	0	7	3	10	3	29	5.62
G7	4	3	3	8	10	16	8	9	2	63	4.44
G8	0	0	3	12	11	11	4	5	1	47	4.43
G9	0	1	3	4	13	7	8	4	2	42	4.71
G10	2	1	4	9	15	12	2	4	5	54	4.37
Total	6	8	26	56	68	70	33	40	15	322	4.49

Correlation between the three pairs of items shown in Table 14 is not as strong as one would expect. For example, of 250 students who responded at Level 2 for Part (a), 32% misinterpreted the way that variation was displayed in Part (b). The low association of Parts (b) and (c) indicates that some students probably guessed the correct response to Part (b) but could not justify the choice in a statistically appropriate manner in Part (c).

Table 14. Indicative Correlations between Parts of the Height Question

	Part (a)	Part (b)
Part (b)	0.083	1
Part (c)	0.188	0.295

The item relating to the association between the amounts of time spent on Homework and mathematics scores depicted in a graph also had three parts (see Figure 1). In Part (a) most students could respond appropriately, reading from the graph the mathematics score for students who spend 1-2 hours per day on homework. Seventy-five percent of students overall were able to answer correctly. The highest rate of incorrect response for this part was in Grade 7, where 35% of students answered incorrectly.

Table 15 shows the number of responses in each grade to Part (b) and (c) of the

Homework Questions relating to what the graph tells about the association (Part (b)) and giving reasons why the graph has the shape it does (Part (c)). As might be expected, relatively speaking Part (b) was easier than Part (c), with the overall mean score for Part (b) (2.56) representing 64% of the total possible, whereas for Part (c) the mean score (0.74) represented 25% of the total possible.

Table 15. Number of Students per Grade and Level for Part (b) and (c) of the Homework Question

Part (b)

Level Grade	0	1	2	3	4	Total (*n*)	Average Score
G5	20	7	9	5	8	49	1.47
G6	23	7	15	13	36	94	2.34
G7	12	2	5	4	26	49	2.61
G8	12	4	3	9	23	51	2.53
G9	11	1	5	4	42	63	3.03
G10	5	0	0	3	37	45	3.49
Total	83	21	37	38	172	351	2.56

Part (c)

Level Grade	0	1	2	3	Total (*n*)	Average Score
G5	46	1	0	2	49	0.14
G6	77	5	4	8	94	0.39
G7	36	3	3	7	49	0.61
G8	36	4	1	10	51	0.71
G9	34	5	5	19	63	1.14
G10	16	3	3	23	45	1.73
Total	245	21	16	69	351	0.74

The percent of students responding at Level 4 for Part (b) shows an increasing appreciation of the association shown in the graph with increasing grade from 16% at Grade 5 to 82% at Grade 10. The percents of Level 0, Level 1, and Level 2 responses tend to decrease with increasing grade, hence showing a general improvement over the years. Examples of responses for each level of response for Part (b) can be found in the left section of Table 16.

The percent of Level 3 responses for Part (c) asking for reasons to explain the shape of the graph increased with increasing grade. Fifty-one percent of Grade 10 students, 30% of Grade 9, and 20% of Grade 8, could create a hypothesis from the data presented to describe the shape of the graph, whereas only 14% of Grade 7, 9% of Grade 6 and 4% of Grade 5 students could do so. Examples of responses for each level of response for Part (c) can be found in the right section of Table 16.

Table 16. Example of Responses for Parts (b) and (c) of the Homework Question

Part (b)		*Part (c)*	
Level Description	*Example*	*Level Description*	*Example*
Level 0		***Level 0***	
Idiosyncratic	"The test would have been hard."	*Idiosyncratic*	"Because it just does."
Misinterpretation	"That 1 hour is the most common."	*Focus on graph type*	"Easy to see and read the numbers and dots."
Level 1		***Level 1***	
Single Focus response	"It tells me about maths homework and scores."	*Comment about the variables but no hypothesis*	"So people can see how long students spend on homework."
Level 2		***Level 2***	
No association identified	"It really does not matter how long you do homework for."	*Comment relating to data reading, with no identifiable hypothesis*	"One hour is a satisfactory amount of time to spend on maths homework."
Incorrect association between variables	"If you put more time and effort it gets you better scores."		
Level 3		***Level 3***	
Causation between variables assumed	"People who just spend an hour on homework get better marks."	*Hypothesis created from the data based on student abilities or other possible contributing factors*	"Because if you are good at maths you'll spend less time doing homework. But if you have a problem with it you are going to spend more time, which is why '> 3 hour' people have lower scores anyway."
Level 4			
Appreciation of association – basic data reading	"That working more than 3 hours is the same as no time."		

The scores across all three parts of the Homework question were summed to make a Total score for this question. The total possible score was eight. Table 17 shows the number of students with each possible Total score.

Although 70% of students responded at Level 0 for Part (c), only 23% of students had a Total score of 0 or 1 over the three items, suggesting that three quarters of the students attempted at least one of the Homework questions.

Table 17. The Distribution of Students per Grade for the Total Score over the Parts of the Homework Question

Level Grade	0	1	2	3	4	5	6	7	8	Total *n*	Average Score
G5	4	18	6	8	7	5	0	0	1	49	2.35
G6	5	16	12	14	12	22	3	6	4	94	3.53
G7	6	6	3	3	6	15	3	4	3	49	3.88
G8	5	7	2	6	7	12	5	3	4	51	3.92
G9	3	8	1	4	5	16	5	8	13	63	4.95
G10	1	3	1	1	2	9	3	10	15	45	5.98
Total	24	58	25	36	39	79	19	31	40	351	4.04

Approximately half of the students scored between 5 and 8, with 11% of students scoring the maximum possible Total of eight. Correlations between the three pairs of parts of the question, shown in Table 18, are not strong, except between Parts (b) and (c) where 15% of the variance is explained. The lack of association of Part (a) with the other two parts is likely the result of the high success rate with the reading of value asked for in Part (a) but the lack of experience in dealing with the type of questions in Parts (b) and (c).

Table 18. Indicative Correlations between Parts of the Homework Question

	Part (a)	Part (b)
Part (b)	0.094	1
Part (c)	-0.077	0.386

Research Question 3: Association of responses across types of questions

A total of 351 students answered both questions: the numerical Average task about an appropriate average to use with a small data set and the graphical Homework task about the association displayed between the two variables, mathematics score and time spent on homework. Only 14 students had a total score of zero on both sets of questions. The Homework task was easier for students overall than the Average task. The correlation of .460 between the Total scores represented 21% of shared variance between responses to the two tasks.

A total of 302 students answered both questions: the numerical Two-way Table task relating lung disease and smoking information and the graphical Height task based on the variability in student heights in two separate schools. Only six students failed to score on both questions and one scored the highest score for each. The Height task was easier for students to than the Two-way Table question and there was little association of scores with a correlation of 0.152, representing two percent of shared variation.

DISCUSSION

For most statistics educators, the performance on the four tasks used in this study would be rather disturbing. It must be acknowledged, however, that the students in the study had experienced no specific instruction related to the topics covered. Their learning was associated with the school's curriculum, which although roughly aligned with the Australian state of Tasmania (Department of Education and the Arts Tasmania 1993) was not mandated and hence very much left to the discretion of individual teachers.

As noted the issues of statistical literacy addressed in the tasks were at a level expecting critical thinking and decision making (Average, Height, Homework questions) or expecting high level proportional reasoning (Two-way Table question). As suggested by Watson (1997), there are three tiers of involvement in such tasks: first understanding the terminology involved, second appreciating the context in which the task is set, and third being able to evaluate critically the claims made. In the context of the tasks used in this study it was apparent that some students did not have a first tier understanding of terms such as median, mode, and variation. This made it impossible to justify responses to the types of questions asked. In relation to the second tier some students had further difficulty with relating the information given to the context of the task, for example smoking and lung disease, or homework time and exam mark. At the third, critical thinking tier many students could not question outliers or accommodate associations that did not fit their preconceived beliefs.

Comparing the expectations of Watson's (1997) statistical literacy with the two types of knowledge, procedural and conceptual (Hiebert & Carpenter 1992; Skemp 1986) shows not only some similarities of aim but also some differences in detail. The link of procedural knowledge, as shown for example in identifying the definition of mode, or in appreciating what a value in a cell of a two-way table represents, can be made to the terminology associated with Watson's first tier of statistical literacy. The understanding of terminology assumed for the overall tasks, however, goes beyond the procedural aspects and becomes conceptual in dealing with the higher order questions that are the goals of the tasks. The tasks are mainly aimed at the conceptual understanding required to tie together terminology, context, and consideration of critical aspects of the questions.

Generally students appeared more willing to attempt the graphing questions than the numerical questions. The fact that the first part of each of the graphing questions was based on basic graph reading, a procedural task, may have contributed to this. The visual presentation may have been another factor encouraging students to become engaged with the graph-based tasks. In each of the numerical tasks, students were required to take in the entire numerical presentation before tackling the questions. The cognitive load may

have been too much for some students. The fact that few students had scores of zero over combined parts of items or over both items attempted indicates that students were indeed attempting the questions and not leaving them blank. The percent of Level 0 responses to individual parts indicates mainly idiosyncratic responses, rather than non response, and a need to focus more directly on the subject matter of the questions in the classroom. Students were perhaps unaware of the inherent difficulty in the tasks.

The usefulness of the tasks reported in this study is related to their relatively short length and yet their setting in contexts that students should be familiar with: measuring mass, smoking and lung disease, height, and study time and mathematics scores. Each task, however, offers an element that requires critical conceptual knowledge to achieve the highest scores. The Average task, for example, includes an outlier in the data set, whereas both the Two-way Table lung disease question and Homework study time task represent an association that is likely to be counter intuitive. The Height task based on graphing heights presents two types of "variation" pictorially, which must be distinguished. Procedural knowledge may be displayed in some parts of the tasks and levels of the rubrics, indicating that students have made some progress on the tasks. Such information from intermediate levels or initial parts of the questions can assist teachers in planning classroom interventions to assist students in reaching higher levels.

The tasks may be used individually as part of assessment of specific topics such as average or proportional reasoning, or put together as an overall assessment as done by Callingham & Watson (2005) (cf. Watson & Callingham 2005). The low level of association between the two types of tasks, numerical and graphical, suggests different innate characteristics to the two types of tasks and that it is not possible to judge potential success on one type from success on the other. The somewhat higher association for the Average and Homework tasks may be a result of both being more closely associated with traditional topics in the mathematics curriculum. The Two-way Table and Height tasks would both appear to require less familiar concepts, those of proportional reasoning and variation represented graphically, which may have resulted in more nearly random behaviour for each and in relation to each other.

All four tasks provide means of assessing Gal's (2002) requirements for interpreting, evaluating, and communicating about statistical information. The requirement in all of the tasks to explain reasoning is significant both in judging conceptual understanding and in assessing the ability to communicate, the important "literacy" component of statistical literacy.

REFERENCES

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. MATHDI **1997a.**00636

Bell, A.; Brekke, G. & Swan, M. (1987): Misconceptions, conflict and discussion in the teaching of graphical interpretation. In: J. D. Novak (Ed.), *Proceedings of the second international seminar: Misconceptions and educational strategies in science and mathematics.* Vol. 1 (pp. 46–58). Ithaca, NY: Cornell University.

Biggs, J. B. & Collis, K. F. (1982): *Evaluating the quality of learning: The SOLO taxonomy.* New York: Academic Press. MATHDI **1985c.**03244

Cai, J. (1995): Beyond the computational algorithm: Students' understanding of the arithmetic average concept. In: L. Meira & D. Carraher (Eds.), *Proceedings of the Nineteenth Annual Meeting of the International Group for the Psychology of Mathematics Education* Vol. 3 (pp. 144–151). Recife, Brazil: Universidade Federal de Pernambuco. MATHDI **1997c.**02130

Callingham, R. A. & Watson, J. M. (2005): Measuring statistical literacy. *Journal of Applied Measurement* **6(1),** 19–47.

Curcio, F. R. (1987): Comprehension of mathematical relationships expressed in graphs. *Journal for Research in Mathematics Education* **18,** 382–393. MATHDI **1988d.**01311

Department of Education and the Arts Tasmania. (1993): *Mathematics Guidelines K–8.* Hobart: Curriculum Services.

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

Estepa, A.; Batanero, C. & Sanchez, F. T. (1999): Students' intuitive strategies in judging association when comparing two samples. *Hiroshima Journal of Mathematics Education* **7,** 17–30. MATHDI **1999e.**03592

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.

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

Garfield, J. & Chance, B. (2000): Assessment in statistical education: Issues and challenges. *Mathematical Thinking and Learning* **2,** 99–125. MATHDI **2001f.**05452

Goodchild, S. (1988): School pupils' understanding of average. *Teaching Statistics* **10,** 77–81. MATHDI **1989i.**02399

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?* (pp. 219–239). Voorburg, The Netherlands: International Statistical Institute.

Hiebert, J. & Carpenter, T. P. (1992): Learning and teaching with understanding. In: D.A. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 65–97). New York: NCTM and MacMillan.

Kelly, B. A. & Watson, J. M. (2002): Variation in a chance sampling setting: The lollies task. In: B. Barton, K.C. Irwin, M. Pfannkuch, & M.O.J. Thomas (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 26th Annual Conference of the Mathematics Education Research Group of Australasia, Auckland, Vol. 2, pp. 366–373). Sydney, NSW: MERGA.

Lokan, J.; Ford, P. & Greenwood, L. (1996): *Maths and science on the line: Australian junior secondary students' performance in the Third International Mathematics and Science Study.* Melbourne: Australian Council for Educational Research. MATHDI **1997f.**04543

Mevarech, Z. R. (1983): A deep structure model of students' statistical misconceptions. *Educational Studies in Mathematics* **14,** 415–429. MATHDI **1984e.**11098

Mokros, J. & Russell, S. J. (1995): Children's concepts of average and representativeness. *Journal for Research in Mathematics Education* **26,** 20–39. MATHDI **1995f.**03985

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

Moritz, J. B. (2003): Interpreting a scatter graph displaying counterintuitive covariation. In: L. Bragg, C. Campbell, G. Herbert & J. Mousley (Eds.), *Mathematics education research: Innovation, networking, opportunity* (Proceedings of the 26th Annual Conference of the Mathematics Education Research Group of Australasia, Geelong, pp. 523–530). Sydney, NSW: MERGA.

Moritz, J. B. (2004): Reasoning about covariation. In: D. Ben-Zvi & J. Garfield (Eds.), *The challenge of developing statistical literacy, reasoning and thinking* (pp. 227–256). Dordrecht: Kluwer Academic Publishers.

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

Pereira-Mendoza, L. & Mellor, J. (1991): Students' concepts of bar graphs — Some preliminary findings. In: D. Vere-Jones (Ed.), *Proceedings of the Third International Conference on Teaching Statistics: School and general issues* Vol. 1 (pp. 150–157). Voorburg, Netherlands: International Statistical Institute.

Pollatsek, A.; Lima, S. & Well, A.D. (1981): Concept or computation: Students' understanding of the mean. Educational Studies in Mathematics **12,** 191–204. MATHDI **1981k.**00533

Skemp, R. R. (1986): The psychology of learning mathematics. Suffolk: Penguin Books. MATHDI **1986i.**03547

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. (1998): 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, QLD: MERGA.

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

Watson, J. M. & Moritz, J. B. (2000): The longitudinal development of understanding of average. *Mathematical Thinking and Learning* **2(1&2),** 11–50.

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

Watson, J. M. & Kelly, B. A. (2003): Inference from a pictograph: Statistical literacy in action. In: L. Bragg, C. Campbell, G. Herbert & J. Mousley (Eds.), *Mathematics education research: Innovation, networking, opportunity* (Proceedings of the 26th Annual Conference of the Mathematics Education Research Group of Australasia, Geelong, pp. 720–727). Sydney, NSW: MERGA.

Watson, J. M. & Callingham, R. A. (2005): Statistical literacy: From idiosyncratic to critical thinking. In: G. Burrill & M. Camden (Eds.), *Curricular Development in Statistics Education. International Association for Statistical Education* (IASE) Roundtable, Lund, Sweden, 2004 (pp. 116–162). Voorburg, The Netherlands: International Statistical Institute.

Watson, J. M. & Kelly, B. A. (2005): The winds are variable: Student intuitions about variation. *School Science and Mathematics* **105,** 252–269.