Statistical Literacy: From Idiosyncratic to Critical Thinking

Jane M. Watson
University of Tasmania
Australia

Rosemary A. Callingham
University of New England
Australia

Abstract

This paper follows earlier research using a survey instrument devised to measure statistical literacy understanding at the school level. Based on partial credit Rasch analysis, the performance of 673 students in Grades 5 to 10 is reported both overall and for three subgroups of items reflecting strands within statistical understanding. The three strands are the basic measurement of average and chance, the related ideas of sampling and inference, and the representation of data and variation. A hierarchy of six levels of understanding is presented, with differing trends across the grades discussed and an example of individual student performance at each level given. Some of these examples illustrate student differences in understanding for the different strands. Implications for the school curriculum are considered with respect to potential development across the years of schooling and realistic expectations for students at various grade levels. Issues for further consideration and research are raised in the final section.

Introduction

Statistical literacy has become an issue for educators in the light of the overwhelming access to information in today's society (Steen, 1997), the changes to the school mathematics curriculum in many countries since 1990 (e.g., Australian Education Council [AEC], 1991, 1994; National Council of Teachers of Mathematics [NCTM], 1989, 2000), and the expectations for school curricula to integrate learning areas in preparing students to meet the needs of society when they leave school (Madison & Steen, 2003; Watson, 2004). Wallman (1993) and Gal (2002) have provided descriptions of the statistical literacy requirements for adults, including the ability to interpret statistical results and evaluate these critically, and the ability to communicate reactions and concerns as required.

Accepting that statistical literacy is a goal for the school curriculum, very little research has been carried out to document the progress made by students as they progress through the compulsory part of their schooling in developing both statistical techniques and critical evaluation skills. In addition, it is not well understood how the goals of statistical literacy might be related to the curriculum in terms of development of statistical concepts. This study, based on a data set of 673 students in Grades 5 to 10, helps establish a set of expectations for students across these years in terms of statistical skills, appreciation of contexts within which statistical understanding is required, critical thinking when required, and the ability to communicate understanding in writing.

Background and Literature Review

Growing out of the overall concern to develop statistical literacy skills in students before they leave school, the specific previous work leading to this study includes the work of Watson (1997) in suggesting a three-tiered Statistical Literacy Hierarchy. The hierarchy consists of (a) an understanding of statistical terminology, (b) an understanding of this terminology in social contexts, and (c) and the ability to question claims made without proper statistical justification. The hierarchy, with the addition of a disposition to become involved in the issues, is consistent with the objectives of Gal (2002) for adults.

Watson and Moritz (2000) considered the development of student understanding of sampling in terms of the specific skills related to statistical literacy. Using survey items reflecting the definition of sample, the appreciation of sampling in a context such as a decision to purchase a car, and the critical analysis of a media article inappropriately claiming an outcome for a population from a sample, improvement with respect to the three-tiered hierarchy was observed across grades. Although across the grades students could increasingly identify individual features relevant to sampling, by Grade 11, only 20% of students performed at an optimal level in relation to Gal's (2002) adult expectations. Students' creation of graphs from information provided in the media and interpretation of graphs presented in the media were other aspects of statistical literacy explored for school students. Again responses were observed indicating hierarchical levels of performance in relation to pie graphs (Watson, 1997), the creation of a graph to show association (Watson, 2000), and the interpretation of a bar graph (Watson & Chick, 2004).

In considering students' drawing of inferences from a pictograph, Watson and Kelly (2003a) found a leveling of performance in the middle years with an indication of some interference from ideas on pattern from elsewhere in the mathematics curriculum. At a basic level of terminology, Watson and Kelly (2003b) considered students' understanding of the terms sample, random, and variation, finding that by Grade 9, 37 percent, 31 percent, and 30 percent of students, respectively, could give reasonable but not necessarily highly sophisticated meanings for the three terms. Aspects of basic chance measurement and average have also been considered as contributing to statistical literacy (Watson & Moritz, 1998, 1999). Although teaching interventions were not part of these studies, increased levels of performance across grades were observed reflecting general curriculum implementation. The identification of hierarchical levels, however, points to sequencing that could take place in relation to classroom or individual instruction based on initial observation of student performance. Conventional measurement techniques were employed in these analyses.

Rasch (1980) modelling was first introduced in the work of Watson, Kelly, Callingham, and Shaughnessy (2003) to measure the understanding of statistical variation within the chance and data curriculum of 746 students in Grades 3, 5, 7, and 9. Statistical literacy was the focus of some items used in this study but Watson and Callingham (2003) carried out further work with larger archived data sets and more items. The students (n=3852) in this later study were in Grades 3, 5, 6, 7, 8, and 9 in government-run Australian schools that were expected to follow a curriculum consistent with that of the AEC (1991, 1994). The study was based on survey items devised to measure student understanding with respect to six aspects of the chance and data curriculum: chance, average, graphs and tables, sampling, variation, and inference. The items placed the skills associated with these topics in various settings reflecting abstract contexts (e.g., tossing dice), familiar contexts (e.g., surveying a school population), and less familiar social contexts (e.g., media articles). Based on the coding of responses in a hierarchical fashion reflecting structural complexity and statistical appropriateness, the data were fitted to the Partial Credit Model (Masters, 1982). Qualitative analysis of the scale indicated six hierarchical levels of measurement of understanding of the construct: Idiosyncratic (Level 1), Informal (Level 2), Inconsistent (Level 3), Consistent Non-critical (Level 4), Critical (Level 5), and Critical Mathematical (Level 6). These levels are described in Table 1.

Continued on Next Page

Table 1
Statistical literacy construct (adapted from Watson & Callingham, 2003)

Level	*Brief characterization of levels*
6 Critical Mathematical	Critical, questioning engagement with context, using proportional reasoning particularly in media or chance contexts, showing appreciation of the need for uncertainty in making predictions, and interpreting subtle aspects of language.
5 Critical	Critical, questioning engagement in familiar and unfamiliar contexts that do not involve proportional reasoning, but which do involve appropriate use of terminology, qualitative interpretation of chance, and appreciation of variation.
4 Consistent Non-critical	Appropriate but non-critical engagement with context, multiple aspects of terminology usage, appreciation of variation in chance settings only, and statistical skills associated with the mean, simple probabilities, and graph characteristics.
3 Inconsistent	Selective engagement with context, often in supportive formats, appropriate recognition of conclusions but without justification, and qualitative rather than quantitative use of statistical ideas.
2 Informal	Only colloquial or informal engagement with context often reflecting intuitive non-statistical beliefs, single elements of complex terminology and settings, and basic one-step straightforward table, graph, and chance calculations.
1 Idiosyncratic	Idiosyncratic engagement with context, tautological use of terminology, and basic mathematical skills associated with one-to-one counting and reading cell values in tables.

As an example of the sort of responses expected at various levels of the scale, consider a task asking for interpretation of a stacked dot plot (item TWNA in the Appendix) with data on how many years 22 families had lived in a town. The horizontal axis had the number of years but with no gaps for unused values. The response category appearing at the Informal level (2) contained responses that understood the question but responded inappropriately with comments like, "A lot of people live in the town." At the Inconsistent level (3), the response category indicated basic data reading from the plot, such as "3 and 12 have the most," whereas at the Consistent Non-critical level (4), the response category reflected data summaries, such as "they range from all years."

Following the work of Watson and Callingham (2003), deletions, changes, and additions of items were made to develop two parallel surveys used to produce the data set described in this study, with the aim of producing a more workable instrument for practical use in schools. The smaller number of items meant that there were too few to develop reliable scales for six sub-groupings, as in the earlier study, and instead, items were combined into three subgroups or strands. The combination of items into subgroups was based on their coherence in relation to the chance and data curriculum and the aims of statistical literacy. The AC strand contains items related to the measurement of average and chance as reflected within the mathematics curriculum in most western countries. The SI subgroup comprises items on sampling and inference due to the close association of these ideas in statistical decision making. The third strand, GV, contains items related to graphing and variation reflecting the predominant use of graphs to display variation. The survey questions used are presented in the Appendix.

Technical aspects associated with establishing the scales and interpretation of the underlying construct are reported in Callingham and Watson (2005). This report extends the earlier work by considering the performance of students across Grades 5 to 10, for both the overall scale and the three subgroups of items. Aggregated performance across grades is presented and this is further elaborated by a consideration of typical performances for individual students at each of the six hierarchical levels. In the Implications and Discussion, issues are raised related to the school curriculum, classroom instruction, and further research.

Methodology

Sample and Procedure

The sample consisted of 673 students in Grades 5 to 10 from five Catholic schools in the Australian state of Tasmania. For Grades 5 to 10, the numbers of students, respectively, were 136, 123, 112, 98, 105, and 99. Classroom teachers administered surveys during class time following instructions in an Administration Manual provided by the authors. The authors prepared a Code Book, and a research assistant coded all responses, occasionally asking questions of one of the authors. The Code Book is incorporated with the Appendix.

The coding schemes were hierarchical based on the structure of the response (Biggs & Collis, 1982, 1991), and the statistical appropriateness of the response (Watson, 1997). As an example based on statistical appropriateness, consider the GV subgroup item VAR1, which asked, "What does variation mean?" A code of 0 was given to idiosyncratic or tautological responses. Code 1 responses provided an example or an isolated idea, such as "lots of choices." Code 2 reflected a simple definition based on difference between things, whereas Code 3 reflected subtle change, such as "slight change or difference." These responses do not show increased structural complexity with increasing code but more specific understanding of the term variation.

Another example, from the AC subgroup, which illustrates both increasing appropriateness and structure, is item DI1B, the explanation for determining whether a die is more likely to land on a 1 or a 6, or equally likely to land on either. A code of 0 was given if no justification of the choice of outcome was made. Code 1 responses were idiosyncratic beliefs, for example related to methods of rolling the die. Code 2 responses were single statements of chance and hence not knowing the outcome. Code 3 reflected the equal likelihood of outcomes in a qualitative fashion, for example, all sides having the same size and weight. Finally Code 4 reflected responses that related the equal likelihood to a mathematical calculation, for example stating a probability of 1 in 6 for all outcomes.

Finally from the SI subgroup, consider the item MV11, which asks for a judgment about a sampling method based on pulling the names of 60 students from a hat to sample a school of 600. A Code 0 response was a misinterpretation or no reasoning about the method. Code 1 responses rejected the method for a variety of reasons, for example that the sample size was too small or that some students would not get a chance to be selected. Code 2 responses accepted the method but for peripheral reasons, such as a good sample size or an easy way to select people, whereas Code 3 reflected an appreciation of the random nature of the process and/or the potential to obtain a range of opinions. The codes hence represent hierarchical steps in addressing the tasks.

Analysis and Interpretation

The data were analyzed using Rasch (1980) measurement techniques, which allowed both students' performances (termed ability) and item difficulties to be measured using the same metric, and placed on the same scale. The Quest computer program (Adams & Khoo, 1996) was used to apply the Partial Credit Model (Masters, 1982) and obtain a variable map showing the placement of students and items along the scale. This map is shown in Figure 1. The horizontal lines indicate points at which there is a qualitative change in the demands of the items, and qualitative analysis confirmed the original description of the variable (Callingham & Watson, 2005; Watson & Callingham, 2003). The subgroups of items were analyzed separately but anchored to the statistical literacy scale so that direct comparisons could be made across the three subgroups. The same level thresholds are applied across each subgroup, so that performances across grades and by individuals on each subgroup of items are comparable.

The mean of the item difficulties is set at zero in the analysis. The match of item difficulty with student ability is used as an indication of the suitability of the instrument for the student cohort. Inspection of the distributions of both students and items in the variable map in Figure 1 shows that in general the distributions match well. This is confirmed by the mean of the student abilities at –0.04 logits, indicating a good match between the item difficulties and student abilities. The presence of items apparently well below and well above the bulk of student abilities is desirable, since it avoids a "ceiling" or "floor" effect. The instrument appears to provide a suitable means of measuring the understanding of statistical literacy of the students tested (Callingham & Watson, 2005).

Continued on Next Page

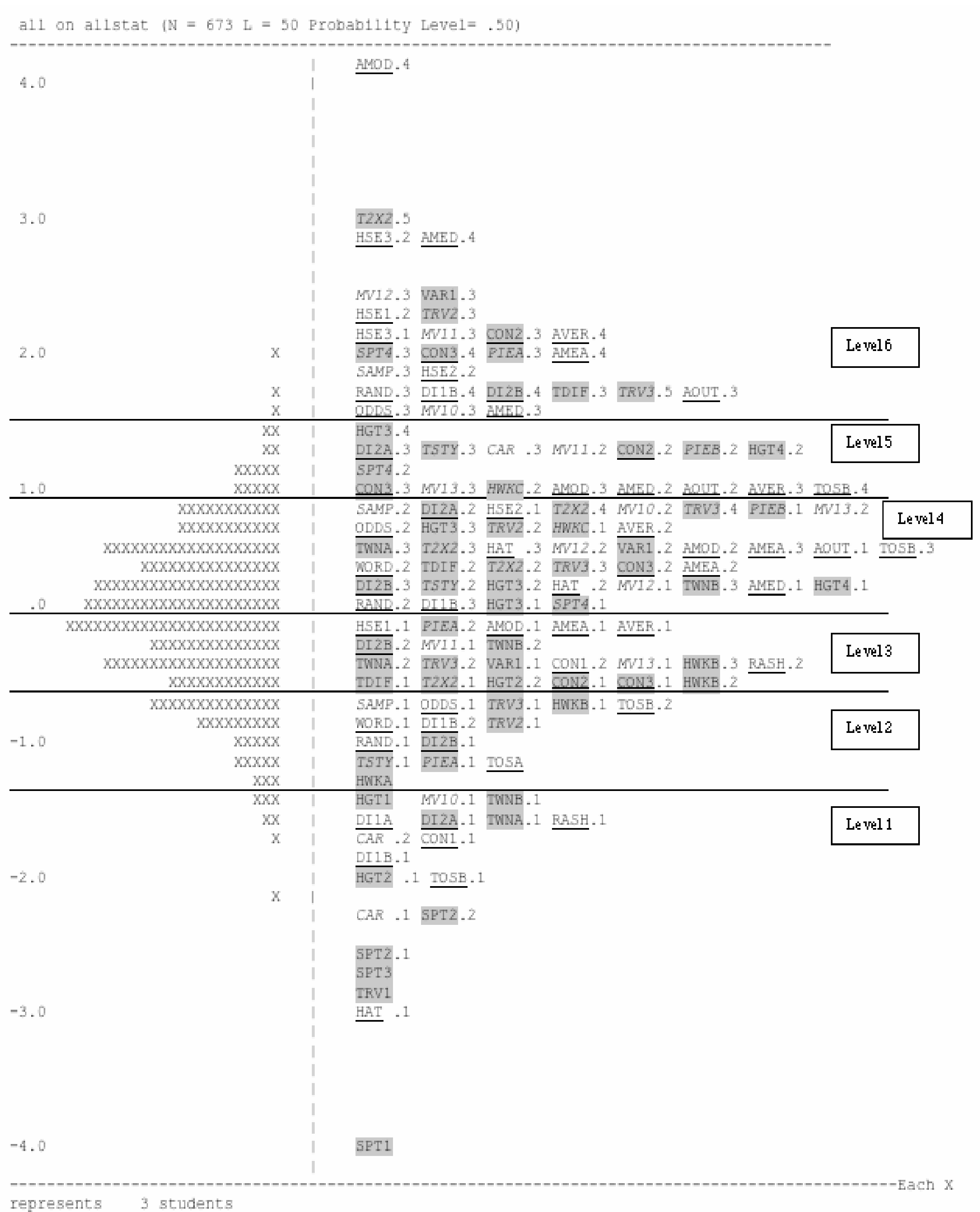


Figure 1. Variable map for the statistical literacy construct based on the items in the Appendix (underline indicates an AC item; italics, an SI item; highlight, a GV item; multiple signs indicate two strands)

The model was evaluated by a consideration of fit of all three subgroups of items, AC, SI, and GV, as well as of the overall statistical literacy scale. Fit is acceptable if it lies between values of 0.77 and 1.3 logits (Keeves & Alagumalai, 1999), and has an ideal value of 1.0 logit. Table 2 shows the fit to the model of the overall scale and related subgroups of items, together with the standard error of measurement (sem). Misfit is shown in italics. The standard error of measurement is small in all instances. Fit to the model is satisfactory across all grades and all subgroups of items, with the exception of the SI scale, which shows small misfit at the Grade 5 level. In general, there is little to suggest that any of the scales is behaving differently for different grade groups of students, suggesting that "opportunity to learn" or curriculum effects do not appear to be strong.

Table 2
Fit to Partial Credit Model of statistical literacy scale & related item subgroups

GRADE	G5	G6	G7	G8	G9	G10	All
STATLIT	0.93	0.98	1.02	1.00	1.07	1.14	1.02
STATLIT sem	0.03	0.03	0.03	0.03	0.04	0.04	0.01
AC	0.86	0.89	0.86	0.88	0.87	0.99	0.89
AC sem	0.03	0.05	0.04	0.04	0.05	0.05	0.02
SI	*0.74*	0.85	0.92	1.02	1.01	1.09	0.92
SI sem	0.03	0.04	0.05	0.05	0.05	0.05	0.02
GV	0.89	0.89	1.05	0.99	1.02	1.07	0.98
GV sem	0.03	0.03	0.05	0.04	0.04	0.05	0.02

That the association of the hierarchical coding system with the spread of items across levels is reasonable can be illustrated with the three items, VAR1, DI1B, and MV11, described above. The item providing for the greatest range of difficulty across the statistical literacy construct is DI1B. Code 1 responses at Level 1 reflect idiosyncratic reasoning. Code 2 responses at Level 2 show a single "anything can happen" chance view. Code 3 responses at Level 4 indicate a consistent non-critical approach to equal likelihood, and Code 4 responses at Level 6 provide a proportional mathematical expression for the probability.

The VAR1 item presents a task with more difficult demands for students. Code 1 appears at Level 3 indicating a vague appreciation of variation appearing along with other inconsistent notions, for example, in relation to items on average. Code 2 appears at the next level suggesting a more consistent appreciation of change, whereas Code 3 appears at Level 6 indicating attention to the critical detail of subtle change or difference in the definition of variation. The MV11 item demands are similar to those for VAR1, with Code 1 appearing at Level 3 indicating the inconsistency of the incorrect alternative, with Code 2 appearing at Level 5 indicating appropriate qualitative reasoning, and with Code 3 appearing at Level 6 indicating use of more sophisticated statistical language. Specific examples of responses for these codes are given in the Appendix.

Descriptive statistics and analysis of variance are used to compare the performance of different grade levels on the overall scale and on the subgroups of items. Performances are displayed visually based on mean values (Figure 2), by percents in each grade at each level (Figure 3), and by box-and-whisker plots to show variation by grade (Figure 4). To illustrate performance typical of the six levels of the variable, kidmaps (Adams & Khoo, 1996) are produced for individual students near the middle of each level. Some of these maps also indicate the variation in performance for students on the different subgroups of items. These are part of the basis for the discussion of the implications of the results for the school curriculum.

Results

The results are presented in two parts. First, grade level performances are compared for the overall statistical literacy variable and the AC, SI, and DV subgroups of items. Second, typical performance of students at each level is discussed in relation to the overall variable and the subgroups of items.

Performance by Grade

Figure 2 presents the mean performance in terms of logits for each of the grades for the statistical literacy variable and for the AC, SI, and GV subgroups of items. The trend generally shows increasing performance as the grade increases, with a slight drop from Grade 7 to Grade 8 in some cases. An ANOVA indicates that overall the change across grades is significant for both statistical literacy generally and the subgroups of items ($F(5, 667) = 18.6$, 16.6, 15.8, and 8.7, respectively, for the groups of items as shown in Figure 2; $p<.0005$ for all tests).

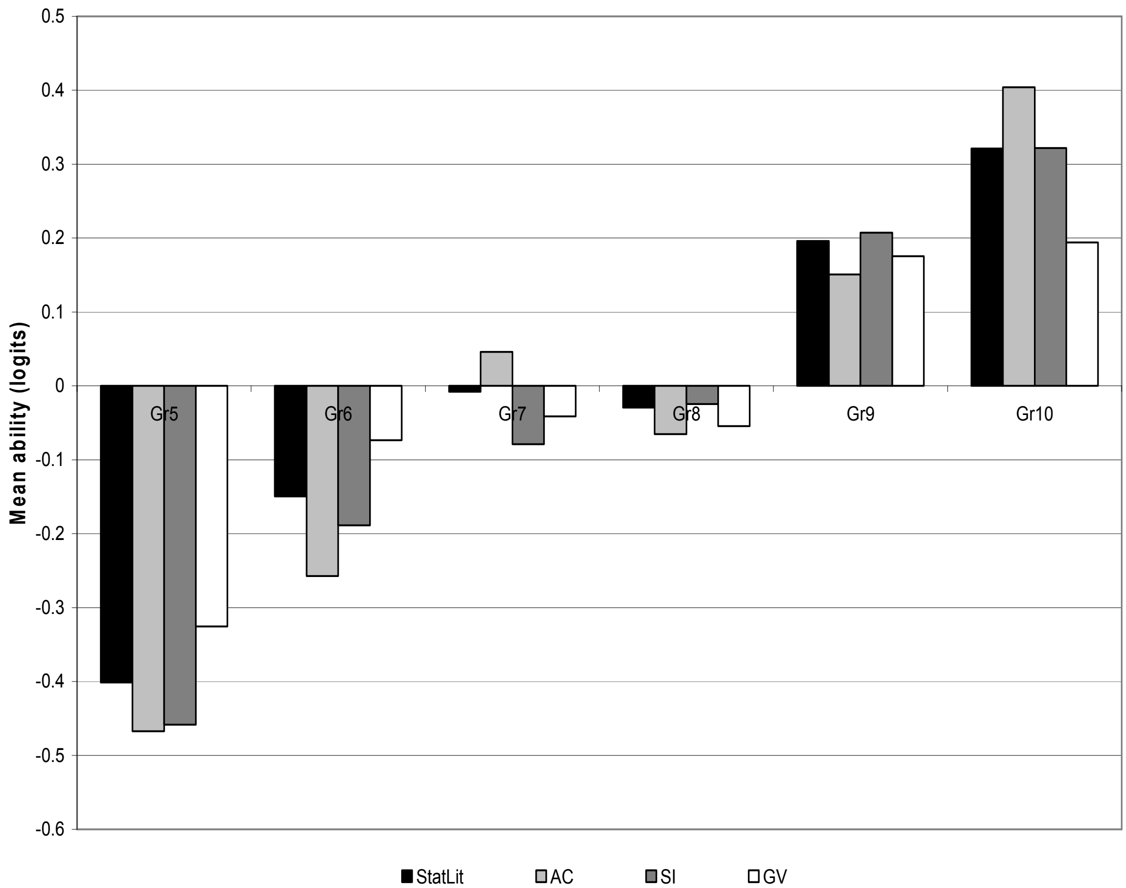
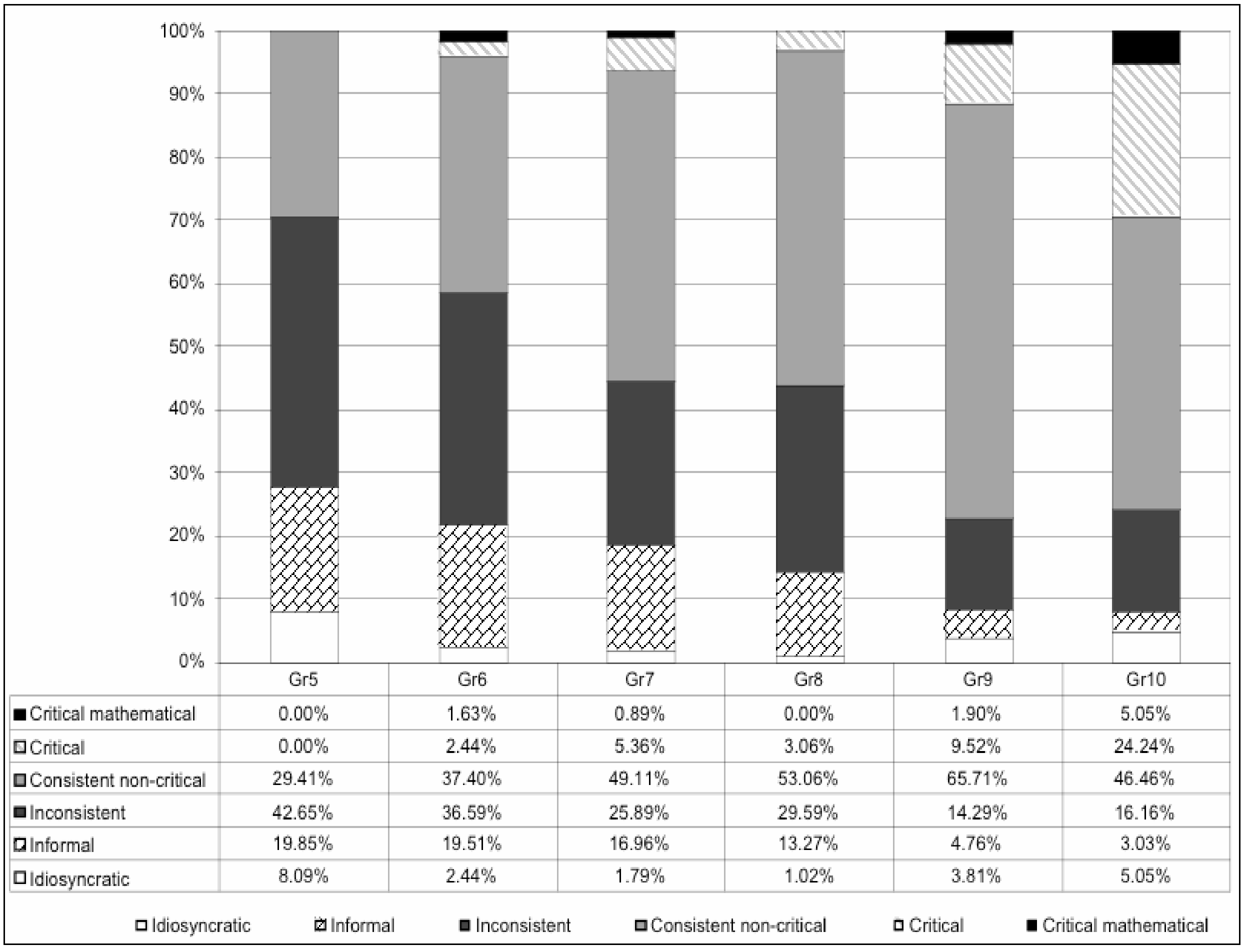


Figure 2. Mean ability (logits) for each grade on statistical literacy and item subgroups (AC, SI, and GV).

The percent of students at each grade performing at the six levels of the statistical literacy hierarchy is shown in Figure 3. The largest percent is at Level 4, Consistent Non-critical, for Grade 6 and above, an issue raised in the Discussion.



	Gr5	Gr6	Gr7	Gr8	Gr9	Gr10
Critical mathematical	0.00%	1.63%	0.89%	0.00%	1.90%	5.05%
Critical	0.00%	2.44%	5.36%	3.06%	9.52%	24.24%
Consistent non-critical	29.41%	37.40%	49.11%	53.06%	65.71%	46.46%
Inconsistent	42.65%	36.59%	25.89%	29.59%	14.29%	16.16%
Informal	19.85%	19.51%	16.96%	13.27%	4.76%	3.03%
Idiosyncratic	8.09%	2.44%	1.79%	1.02%	3.81%	5.05%

Figure 3. Percents of students at each of the six levels of statistical literacy for each grade.

The distributions of ability measures for statistical literacy generally and the item sub-groups separately are also of interest. Figure 4 shows the box-and-whisker plots for all grades for the complete statistical literacy variable. This indicates a lack of growth from Grade 7 to 8, shown by the medians and an apparent dip at the 90^{th} percentile level. Figure 5 contains the box-and-whisker plots for the three subgroups of items and as can be seen a similar dip appears at Grade 8 for the AC items; however the median is flat for the SI subgroup and rises slightly for the GV items for Grade 8.

It is also of interest to observe the behavior of the 90% bars in the plots. For example, in considering the pattern of higher performing students at each grade, it is seen that for AC items there is a dip reflecting the median grade performance, but for the SI items there is a monotonic increase again mirroring that for the median. For the GV items there is a shallow dip at the 90% level in Grade 8 whereas the median continues to rise. In all cases the pattern for students at the lower end shows improvement to Grade 9 and then diminishing performance again at Grade 10. Again this is a focus in the Discussion.

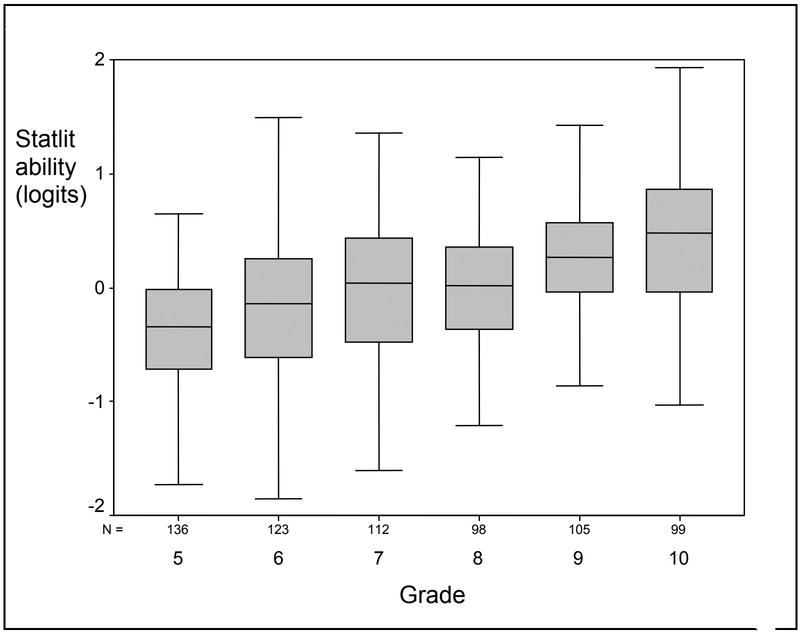


Figure 4. Box-and-whisker plots for statistical literacy by grade.

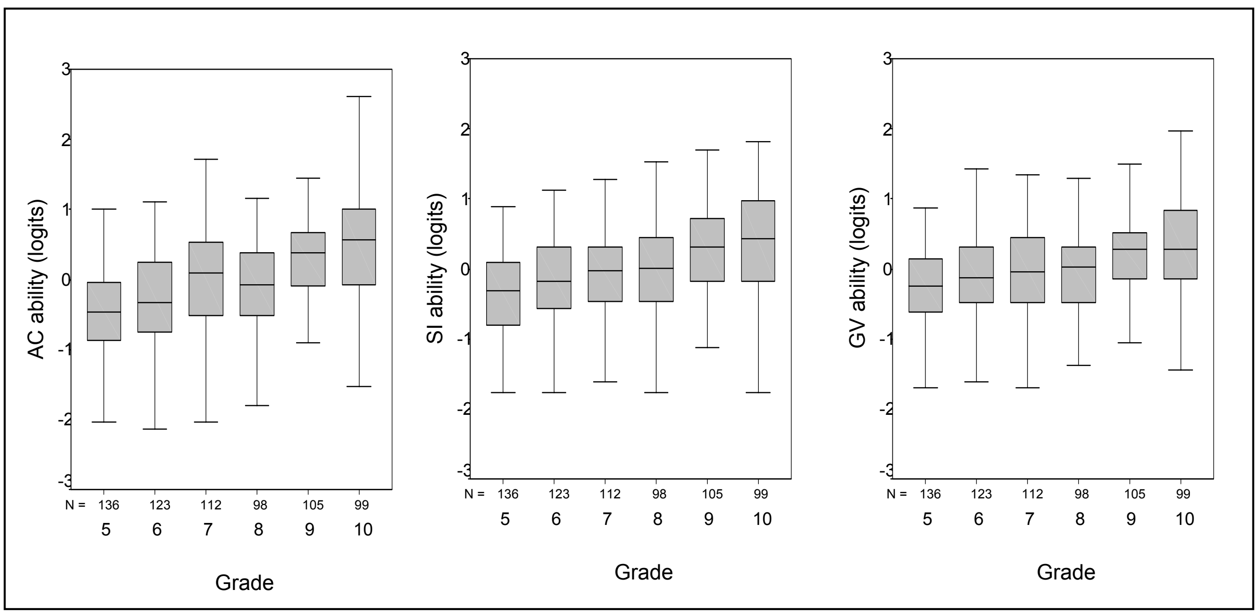


Figure 5. Box-and-whisker plots for AC, SI, and GV items by grade.

Performance at Levels

To explicate the performance of students at each of the six levels of statistical literacy, a kidmap is presented, illustrating typical performance overall and at times the difference or similarity for the three subgroups of items. A kidmap presents the items on the survey form completed by a particular student (see Appendix for identification of items), with the XXX representing the position of the student on the variable map in Figure 1.

The dotted lines in the figure indicate the range of difficulty where the student has approximately a 50% chance of success on items appearing within those limits. For items below the lower line the chance of success is greater than 50% and for those above the higher line it is less than 50%. Items found in the upper left or lower right of a kidmap, hence, represent somewhat anomalous performance based on the expectation of the model. The reported student ability values, measured in logits, indicate the student's position on the variable map in Figure 1.

Level 1: Idiosyncratic

At this level students generally achieve success on a small number of items as seen in Figure 6. This success is usually with regard to providing idiosyncratic explanations for chance outcomes (DI1A) or completing basic table or graph reading tasks (SPT1, TRV1).

The Grade 5 student (105) in Figure 6 (overall ability = –2.05 logits), who answered Form 2, illustrates idiosyncratic thinking on nearly all items but an unexpectedly good response on a graphing item based on a stacked dot plot (TWNB), reflected in an ability measure on the GV items of –1.37. For the AC items (ability = –3.16) the student believes either a one or a six as more likely to occur on a die and cannot interpret a 15% chance of getting a rash (RASH). For the SI items (–2.28), the student, for example, does not recognize any method for selecting a car, inappropriate or appropriate. This student, despite low performance overall, is close to a Level 2 performance on the GV scale, suggesting that the student as a better than expected understanding of data presented in graphical form.

Continued on Next Page

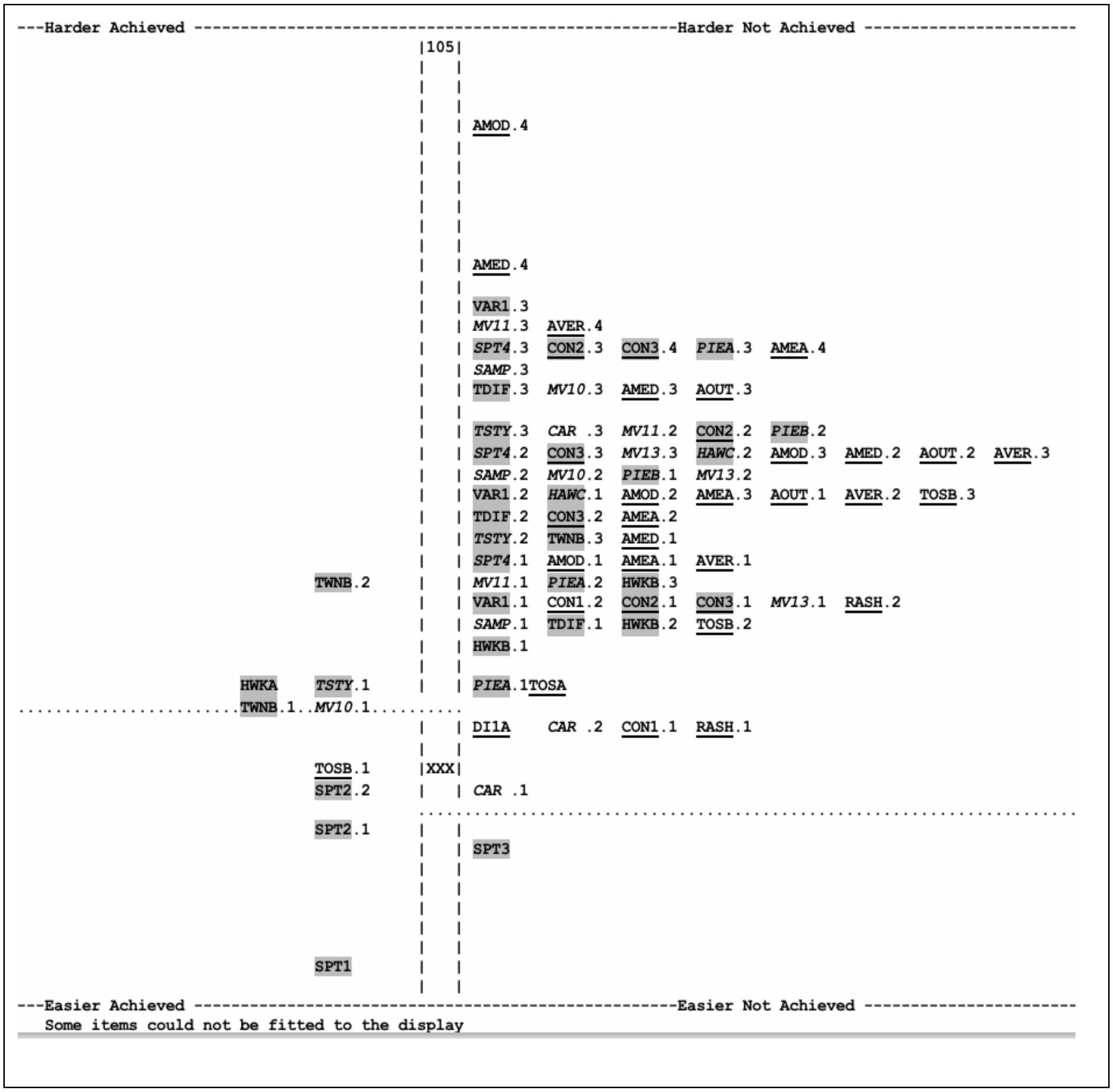


Figure 6. Kidmap at Level 1 – Idiosyncratic

Level 2: Informal

At the Informal level students are beginning to become involved in the context of some questions and, for example as seen in Figure 7, express uncertainty equivalent to "anything can happen" in explaining dice outcomes (DI1B). They attempt to explain the meaning of a stacked dot plot, although inappropriately (TSTY), may give an intuitive reason to explain an observation in a pictograph (TRV3) and are able to provide a single idea associated with the concept of sample (SAMP).

Student 713 in Figure 7 is a Grade 6 student (overall ability = –1.04) who, in answering Form 2, appears to find all three subgroups of items equally difficult (–1.03 for AV, –1.13 for SI, and –1.14 for SV). This performance is typical of the Informal level, with the unexpected responses lying close to the boundaries of the 50 percent success rate.

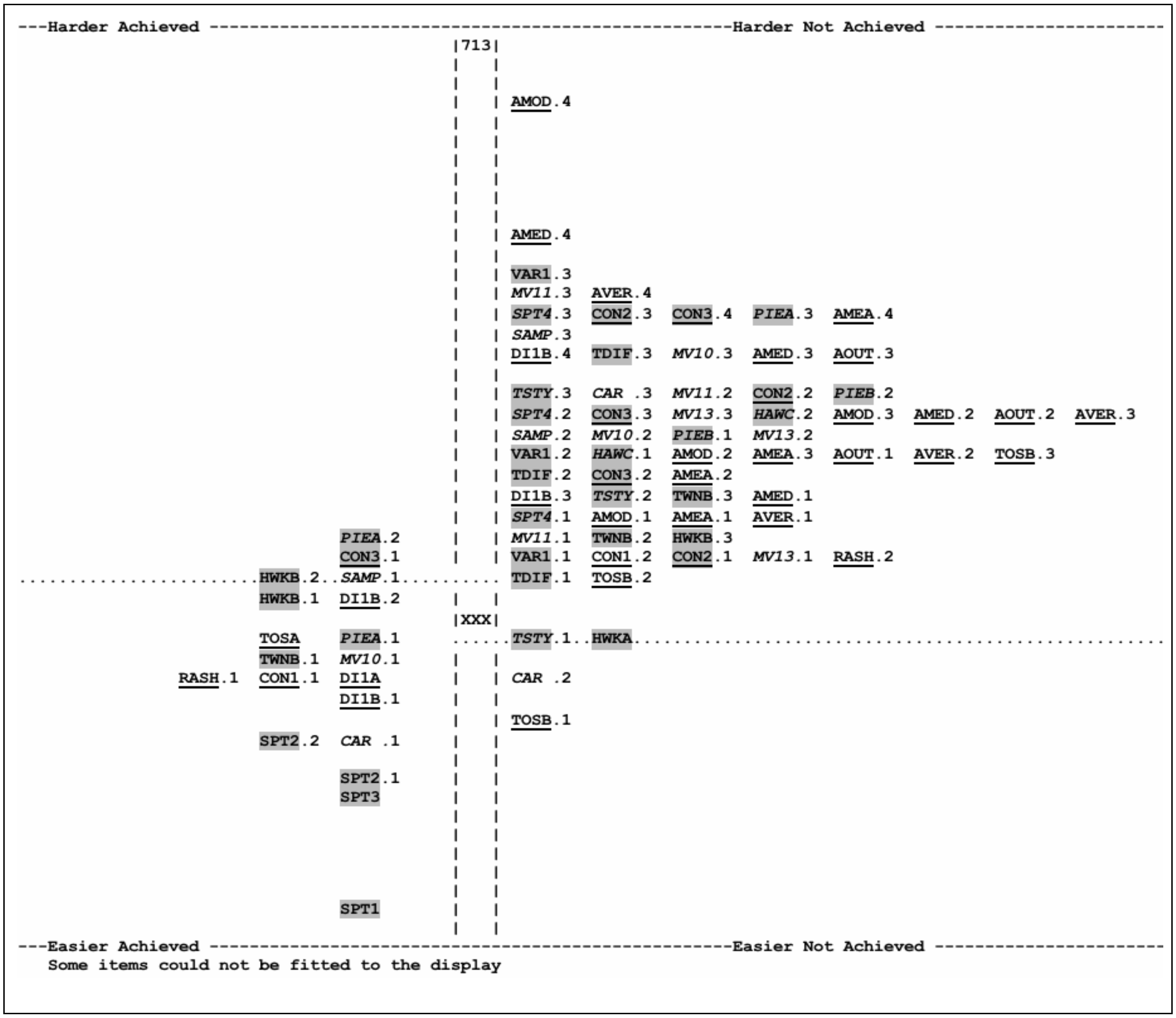


Figure 7. Kidmap at Level 2 – Informal

Level 3: Inconsistent

At the Inconsistent level students may be able to identify in a supportive setting (such as a multiple choice item) the meaning for "15% chance" (RASH). They also are likely to apply content knowledge without statistical reasoning in interpreting data in a two-way table (T2X2) but be unable to criticize an inappropriate sampling method (MV13). In considering a newspaper article about house prices, the students are likely to have a single idea about average not related to the context of median house price in the article (HSE1).

Student 218 in Figure 8 is in Grade 7 (overall ability = –0.43). Responses to Form 1 show a very different relative performance on one of the three subgroups of items, being better on the AC items (0.53) compared to the other two (–0.91 for SI and –0.82 for GV). The better performance on AC items is seen for example in a basic quantification of chance (HAT and DI1A) and ordering of chance newspaper headlines (WORD). Performance on SI items is inconsistent, as seen in the items on inference in the lower right of the kidmap (TSTY, MVE10, TRV2), whereas unexpectedly good responses are seen to CAR and T2X2. A similar mixture of performance is seen on GV items. This student appears to respond at a higher level to items more specifically related to sampling (CAR) and where data are presented in a table form (T2X2) rather than graphs (HGT2).

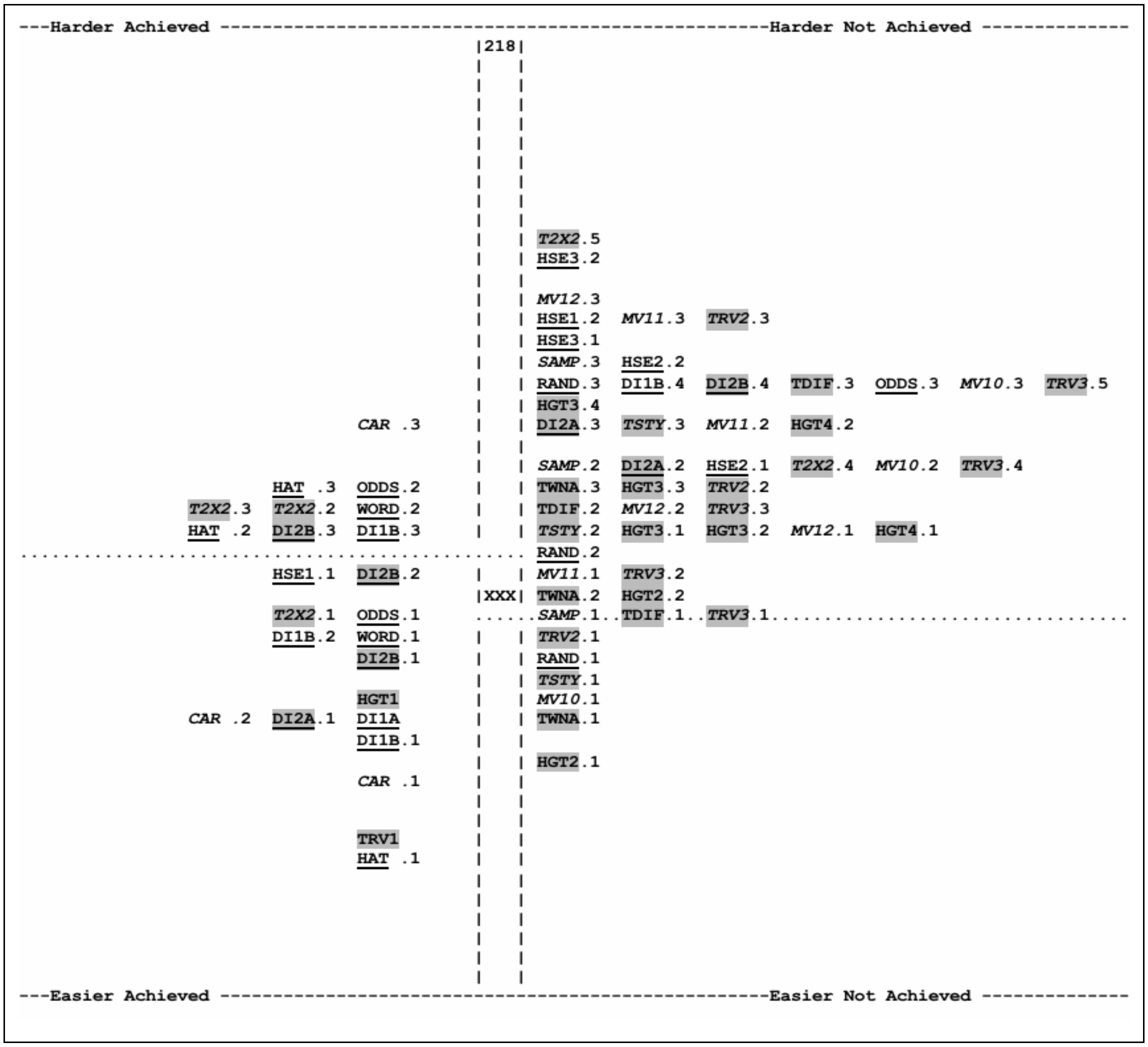


Figure 8. Kidmap at Level 3 – Inconsistent

Level 4: Consistent Non-critical

At this level students are likely to be successful on items that do not require critical analysis or to provide contextual but not critical responses to them. They are likely for example to provide an appropriate summary of the data in a stacked dot plot (TWNB), to order newspaper headlines involving chance appropriately (WORD), to use evidence from a single cell of a two-table to justify a claim (T2X2), and to provide two elements to the definition of "sample" (SAMP).

Student 231 (overall ability = 0.26 logits) is a Grade 7 student whose performance on Form 1 on the SI (0.07) and AC (0.85) items varies from that on the GV items (0.29). This is shown in Figure 9. The items in the lower right portion of the kidmap are all SI and GV items. This student appears to have particular difficulties with describing samples (MV10 and MV11) and explaining differences between graphs in familiar contextual settings (TSTY and TDIF) but can deal with information in some media settings in expected ways for the ability level (HSE).

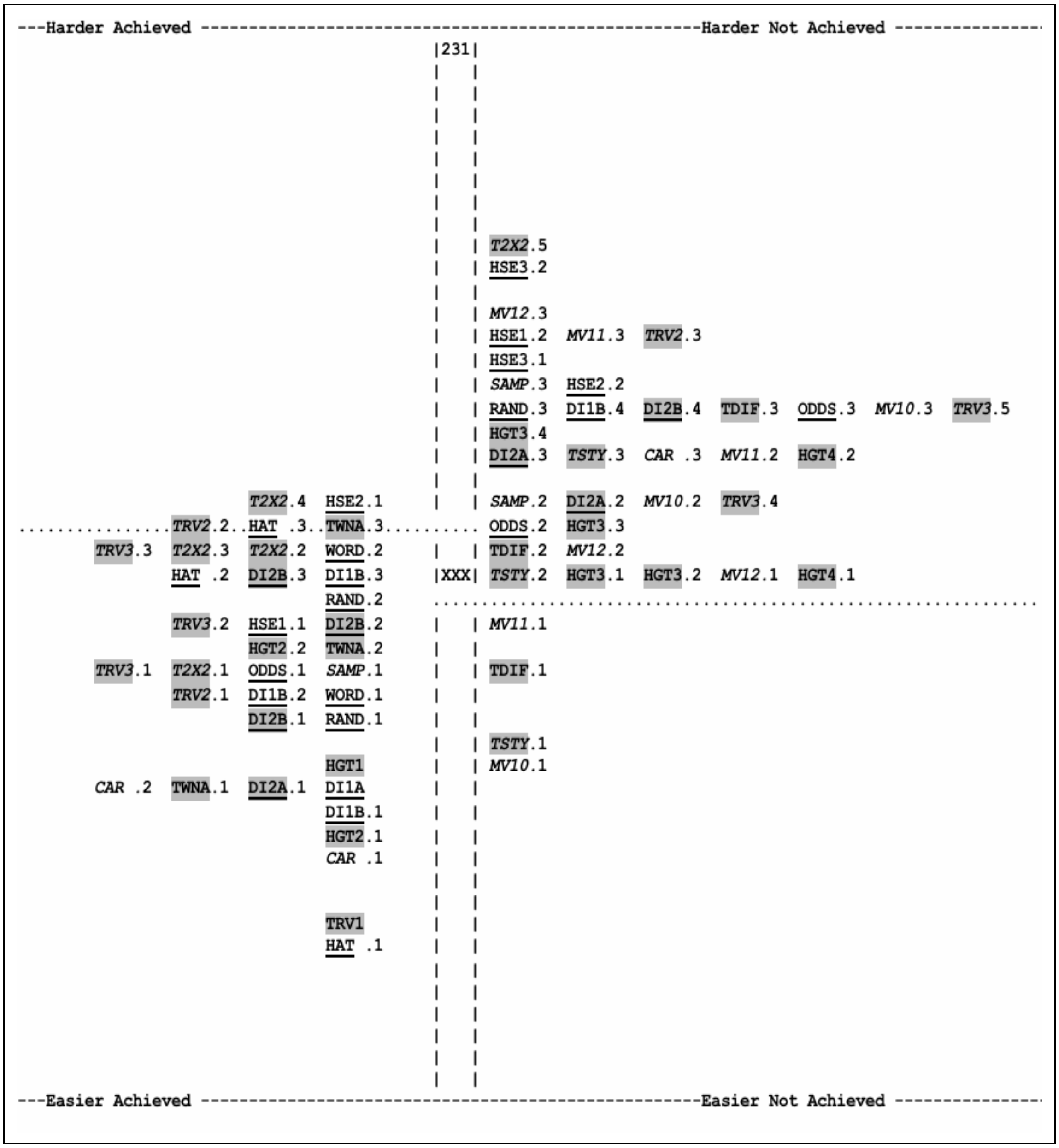


Figure 9. Kidmap at Level 4 – Consistent Non-critical

Level 5: Critical

At the Critical level items reflect students' ability to think critically but without the necessity to perform sophisticated mathematical calculations. Responses are likely to include appropriate criticism of a voluntary survey method (MV13), appropriate selection of a method for buying a car (CAR), and implicit but not formal understanding of probability in explaining a distribution for many tosses of four coins (CON3).

The Grade 10 student (590) in Figure 10 (overall ability = 1.25), shows a somewhat higher level of performance on Form 2 for the AC items (1.64) than for SI (1.20) and GV (1.05). This student is unexpectedly unable to read a result from a curvilinear graph (HWKA) but demonstrates a good understanding of average (AVER). This may be related to curriculum or classroom emphases. Otherwise the performance is typical of this level.

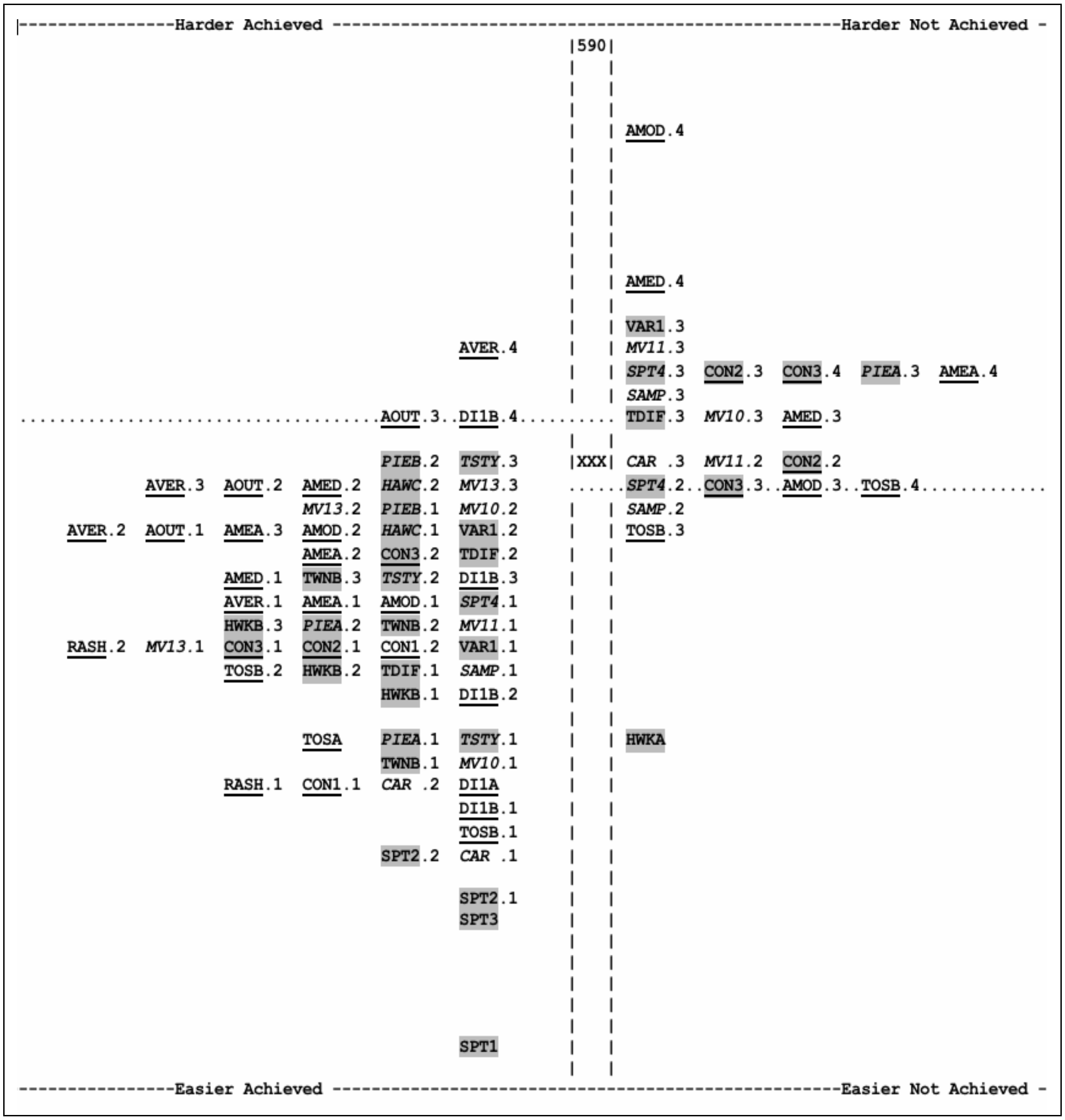


Figure 10. Kidmap at Level 5 – Critical

Level 6: Critical Mathematical

The performance of Level 6 students is quite strong on all aspects of the survey, for example, making a prediction involving uncertainty based on a pictograph (TRV3), making random and representative suggestions for a survey (MV10), explaining the meaning of a pie graph (PIEA), and handling questions about different averages (AOUT, AVER, AMEA) appropriately.

Student 447, a Grade 10 (overall ability = 1.66 logits) answering Form 1, shows similar performance across all subgroups of items (AC = 1.58, SI = 1.83, and GV = 1.69). The kidmap is shown in Figure 11. Across the three types of items, however, this student has difficulty providing written definitions of a standard equivalent to other more numerical responses. When the student's unexpected responses are considered it seems that the student may have difficulty explaining ideas in writing.

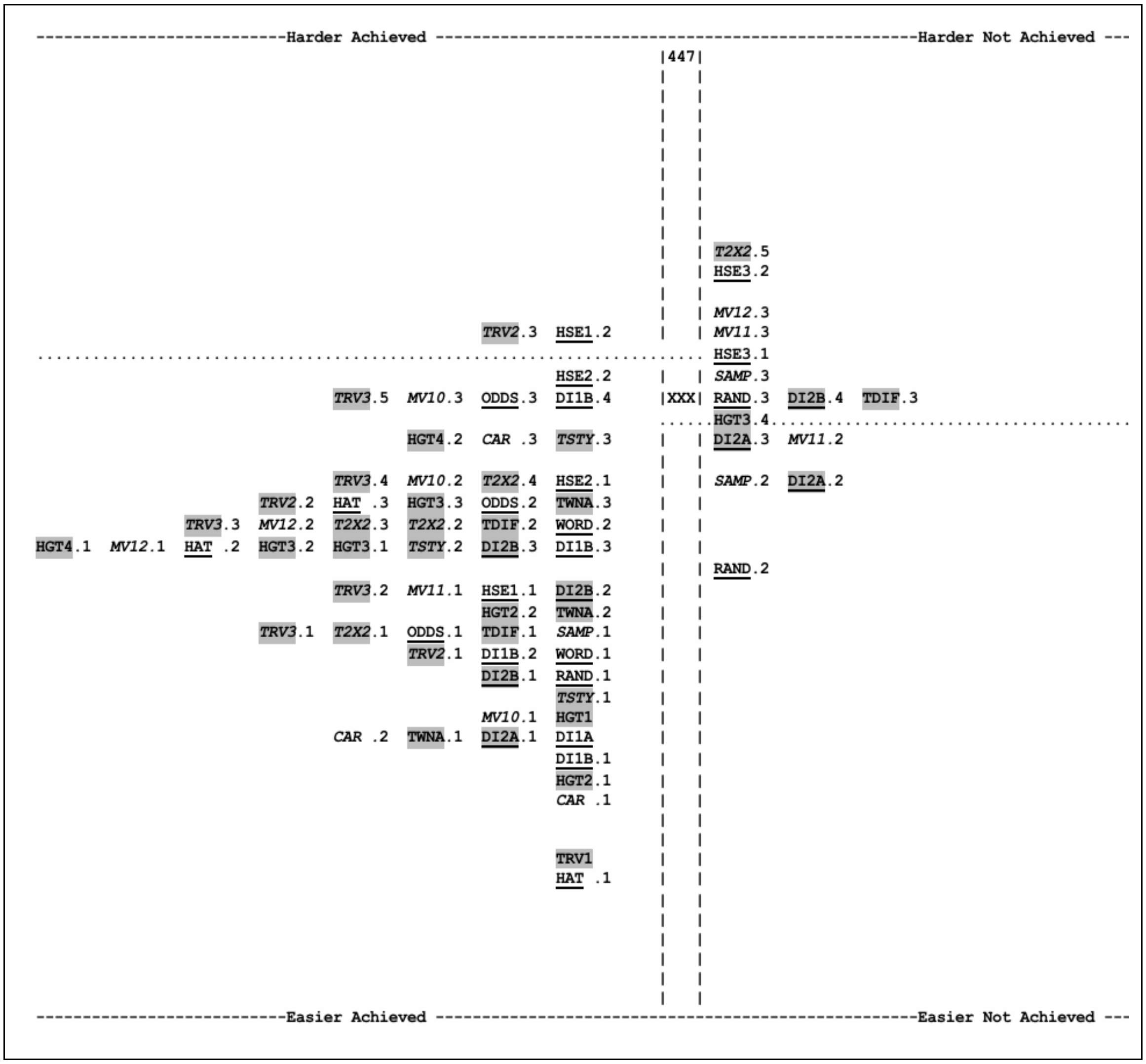


Figure 11. Kidmap at Level 6 – Critical Mathematical

Implications and Discussion

<u>*Trends across Grades*</u>

The overall scale of statistical literacy shows growth in mean performance from Grade 5 to 10 (Figure 2) as would be expected. From Grade 7 to Grade 8, however, there is little change. On some groups of items the mean performance actually appears to fall. This plateau effect has been reported elsewhere in different contexts (see, for example, Callingham & McIntosh, 2002; Hill, Rowe, Holmes-Smith, & Russell, 1996). As shown in Figure 3, however, it is clear that by Grade 10 students are beginning to reach the two highest levels of the hierarchy, Critical and Critical Mathematical, with one-quarter of all students reaching these levels. In Grade 5 the highest level reached is Level 4, Consistent Non-critical, and nearly three-quarters of all students are at the Inconsistent level or below.

Across all grades, however, Level 4 behavior dominates. This Consistent Non-critical level is characterized by appropriate, but unquestioning, engagement with context, and straightforward application of statistical skills associated with the calculation of simple probabilities and means and graph reading. These skills would seem to be appropriate for the lower years of high school. In Grade 10, however, nearly half of all students are still exhibiting Consistent Non-critical behavior. This suggests that more opportunities need to be created for students to question critically statistical claims from media sources or other real-world contexts in order to develop the analytical habits of mind that are needed to respond critically to quantitative claims.

In terms of the findings presented here, it would be reasonable to expect students in Grade 5 to be able to engage with quantitative information in familiar contexts, and to discuss this information in qualitative terms, with little justification. These are characteristics of the Inconsistent level. In Grade 6, attempts should be made to broaden the contexts in which students interact with data, discussion of variation within chance settings becomes important, and students should be introduced to more formal statistical skills, such as calculation of means and more complex data displays in the form of tables and graphs. These topics provide a bridge to the more complex understanding of Level 4. In Grades 7 and 8, further development of the more formal ideas of statistics can begin, and these need to be presented in increasingly unfamiliar contexts. Teachers may need to model critical questioning behavior during these years, and should encourage students to begin to acquire these skills in order both to consolidate the understanding shown in Level 4, Consistent Non-critical, and to provide a bridge to Level 5, Critical.

Although the students in this study were undertaking an Australian curriculum, the expectations of the mathematics curriculum are similar to those in other countries' curriculum documents. In comparing the outcomes observed for students in this study with curriculum expectations, for example in documents such as the NCTM *Standards* (2000), it is important to note subtle differences in the goals of statistical literacy and the goals of the school curriculum. The ability to think critically, for example questioning claims, and to use proportional reasoning mathematical skills are important for high levels of statistical literacy. Whereas proportional reasoning is acknowledged by the NCTM from the Standard for Grades 6-8, the ability to think critically in assessing the claims of others does not receive the same attention. The focus for the NCTM is rather on students creating their own statistical investigations and following them through. As indicated in the discussion above, teachers may need to model the questioning necessary for students to develop high levels of skill in statistical investigation. An inability to reflect critically on statistical claims is likely to have an impact on the quality of statistical investigations initiated and undertaken by students.

The modelling and encouragement of questioning appears to be essential in Grade 9 to provide a basis for further development. Unless students are provided with appropriate opportunities at this stage, their development may remain at a non-critical level. Reaching the higher levels of the hierarchy presented in this study is unlikely to be simply a function of developmental maturity but will be related to appropriate and sensitive teacher intervention supported by a curriculum that values critical, questioning behavior justified by appeal to appropriate quantitative data. By the time students reach the NCTM Standard for Grades 9-12, the expectation for knowing methodologies and specific techniques would lead one to infer that critical examination of others' work could be expected. The high level of content knowledge suggested for the Grade 9-12 Standard, however, is not tested in the survey of statistical literacy used in this study.

The other issue that arises when comparing outcomes like those described in this study, where partial credit is given, with standards that are set in curriculum documents, is that curriculum documents state final outcomes and do not suggest steps along the way where students may display partial understanding. As an example consider the NCTM Grade 3-5 Standard, which says "use measures of center, focusing on the median, and understand what each does and does not indicate about the data set" (p. 176). The equivalent comment from the Grades 6-8 Standard is "find, use, and interpret measures of center and spread, including mean and interquartile range" (p. 248). What was observed in this study is that many students develop partial appreciation of these ideas (e.g., AMED and AMEA; see Appendix for codes describing partial success) but do not move to critical levels of understanding. What is disappointing in terms of the outcomes observed in this study is that many middle school students do not appear able to "use proportionality and a basic understanding of probability" as suggested for Grades 6-8 (NCTM, 2000, p. 248). This is an area that clearly requires further work.

Some interesting differences can be seen among the three item subgroups. The box plots of the overall statistical literacy variable (Figure 4) across grades show a wide spread of achievement in Grade 6, whereas that of Grades 8 and 9 is more contracted. For the three subgroups of items, Average/Chance (AC), Sample/Inference (SI) and Graphing/Variation (GV) (Figure 5), the medians show a similar pattern across grades as do the means, as expected from distributions that are approximately normal.

The performance of students at the extremes of the distributions, however, is somewhat different. Higher ability students tend to show a "dip" in performance from Grade 7 to Grade 8 on the overall statistical literacy scale, and this is quite marked on the AC items, and also seen on the GV items (Figure 5). On these items lower ability students, in contrast, show a continued improvement at least until Grade 9, which explains why the distributions are contracted. The exception to this observation is the SI subgroup. On this subgroup the higher ability students' performance appears to increase almost monotonically, whereas that of the lower ability students, other than a rise in Grade 9, varies very little.

Although it is not possible to explain these patterns on the basis of the data presented, some conjectures can be made on which further research might be based. The AC items, particularly, and also the GV items to some extent, address aspects of the chance and data curriculum currently found in schools in Australia, and in other countries as suggested by the NCTM (2000) Standards quoted above. These include measures of central tendency such as mean, median and mode, and graph and table reading. Appreciation of sampling and its importance in drawing inferences, however, is less likely to be included in the curriculum as it is enacted. A similar observation can also be made about the recognition of the importance of variation.

Patterns of response may, thus, be related not only to grade level and experience but also to students' opportunity to learn. In this instance the underlying construct is described not only by the mathematical demands of the items but also by the interaction with increasingly complex contexts. Opportunities to experience both content and context are necessary for students to progress. Drawing inferences, in particular, requires an appreciation of the context to which the data relate. This may explain the lack of growth among lower ability levels from grade to grade on the SI items. This kind of explanation, however, does not address the issue of the depressed performance of higher ability students from Grade 7 to Grade 8 on the GV, and particularly the AC, items.

One possibility is that students in the early years of high school, Grades 7 and 8, are introduced to more complex mathematical aspects of statistics and probability. This is consistent with the shift from Level 3, Consistent, to Level 4, Consistent Non-critical, in Grades 7 and 8 shown in Figure 3. One characteristic of the Consistent Non-critical level is the ability to undertake somewhat sophisticated mathematical procedures but to apply these uncritically to a limited range of contexts.

The depressed performance of higher ability students from Grade 7 to Grade 8 may reflect students' partial understanding of these more sophisticated mathematical skills. In contrast, students demonstrating less ability may have had less opportunity to develop the necessary mathematical skills, particularly if they are streamed into ability groupings in these years. Such students may show this shift later. The percentages of students in Grades 9 and 10 in the Inconsistent level, Level 3, are very similar to those in Grades 7 and 8 in the Informal level, Level 2. This conjecture could be explored by a consideration of the curriculum and by a consideration of data obtained through interviews.

The Statistical Literacy Hierarchy describes a general developmental progression through which most students will pass. As shown by some of the individual students' responses described in this paper, however, progress along the hierarchy is not necessarily linear. At different times students may show more or less understanding on particular kinds of items. It is possible that a particular student, for example, demonstrates Level 2 performance on SI items, but Level 3 performance on AC or GV. This is seen in the range of performances illustrated by the kidmaps in Figures 6 and 8. This may well depend on the classroom experiences to which the student has had access. At present, most curriculum documents still emphasize mathematical skills rather than inferential processes, and this could affect the performances of students.

Suggestions for the Classroom and Research

The range and variation of performance of students across Grades 5 to 10 means that a variety of statistical activities with wide potential to produce results at various levels of the hierarchy is required at all grade levels. Many possibilities can be developed based on the items used in the survey. Based on the task related to stacked dot plots of how many years families have lived in a town described early in this paper, it would be appropriate to have students complete such an activity, indicating how long their own families had lived in their town, discuss various scales as more or less meaningful, and write reports summarizing the information and telling the story in the graphs.

Context was found to be an important factor affecting the difficulty of items for students, and the curriculum, both in mathematics and other subject areas, should be structured to include opportunity for students to meet data, uncertainty, and prediction in many contexts throughout the school years. By the time they leave school, students should be able to criticize inappropriate media reports with confidence.

Traditional text book questions are unlikely to fulfill the need of providing motivating contexts to challenge students' critical thinking. It is important for teachers to exploit current issues, for example from television or local newspapers. Text books could introduce an example or two from the media, which were relevant and did not become dated, and the teacher might develop exercises involving the finding and interpreting of current articles. Students should, one way or another, become motivated to become media detectives, bringing examples to class.

More recent books do focus on activities that integrate the specific skills of data and chance into investigations as suggested by curriculum documents, for example the *Navigating* (e.g., Bright, Brewer, McClain, & Mooney, 2003; Bright, Frierson, Tarr, & Thomas, 2003) or *Exploring Statistics* (e.g., Bereska, Bolster, Bolster, & Scheaffer, 1999) series. Activities such as one based on considering data for treatment of migraine headaches should assist students in developing statistical literacy skills (Bright, Brewer et al., 2003) but in assessing outcomes teachers will need to develop rubrics that reflect partial understanding of the processes involved, similar to those described in the Appendix of this paper.

For students, learning to communicate their understanding and concerns, for example by writing letters to newspaper editors, government officials, or company bosses, is an important aspect of critical statistical literacy as required for adults by Gal (2002). The difficulties that some students have with this aspect is indicated by the results of the student whose kidmap is shown in Figure 11. The student has apparently high levels of numerical skill, but the responses to written ideas are unexpectedly poor. Specific suggestions for classroom activities based on chance, data, variation, and inference in media contexts that have grown out of earlier research are suggested in Watson (1999, 2002).

Although the contexts used in this survey were Australian, they would be expected to be common in the media and social milieu of most countries. Certainly the media examples presented by Gal (2002) from Israel and the United States are of a similar nature. The use of media reports on sporting events popular in a certain region could be motivating for some students, but care must be taken to cater for all groups surveyed, for example both boys and girls. The authors predict that similar outcomes to those achieved in this study would be achieved in other western countries and await with interest the research of others in the field.

One of the purposes for developing the statistical literacy survey used in this research was to provide two parallel forms of an instrument that would be practical for use in a range of middle school classrooms. A Code Book was developed as part of the project, which is incorporated in the Appendix, but in this study the coding was carried out by a research assistant not a classroom teacher. It would be of interest for others to use the survey in classrooms with marking by teachers. Although the production of kidmaps would require data entry and Rasch analysis, the descriptive checking of codes for a student with reference to a variable map such as shown in Figure 1 would be useful in determining students' strengths and weaknesses. It is the authors' view that the survey offers rich potential for documenting and following students' progress in the area of statistical literacy over the middle school years.

Acknowledgements

This research was funded by the Australian Research Council through grants W0011912 and DP0208607. Data were entered by Judith Deans and coded by Annaliese Caney.

References

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

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.

Bereska, C., Bolster, C.H., Bolster, L.C. & Schaeffer, R. (1999). *Exploring statistics in the elementary grades. Book Two (grades 4-8)*. White Plains, NY: Dale Seymour.

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

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

Bright, G.W., Brewer, W., McClain, K. & Mooney, E.S. (2003). *Navigating through Data Analysis in Grades 6-8*. Reston, VA: National Council of Teachers of Mathematics.

Bright, G.W., Frierson, D. Jr., Tarr, J.E. & Thomas, C. (2003). *Navigating through Probability in Grades 6-8*. Reston, VA: National Council of Teachers of Mathematics.

Callingham, R. & McIntosh, A. (2002). Mental computation competence across years 3 to 10. In B. Barton, K.C. Irwin, M. Pfannkuch & M.O.J. Thomas, (Eds.), *Mathematics education in the South Pacific* (Proceedings of the 25th annual conference of the Mathematics Education Research Group of Australasia, pp. 155-163). Sydney: MERGA.

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

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

Hill, P.W., Rowe, K.J., Holmes-Smith, P. & Russell, V.J. (1996). *The Victorian Quality Schools Project: A study of school and teacher effectiveness. Report (Volume 1).* Melbourne: Centre for Applied Educational Research, University of Melbourne.

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). New York: Pergamon.

Madison, B.L. & Steen, L.A. (Ed.). (2003). *Quantitative literacy: Why numeracy matters for schools and colleges*. Princeton, NJ: The National Council on Education and the Disciplines.

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

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.

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

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

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

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. (1999). Lessons from chance and data research II: For the classroom. In K. Baldwin & J. Roberts (Eds.), *Mathematics – The next millennium* (Proceedings of the 17th Biennial Conference of the Australian Association of Mathematics Teachers Inc., pp. 311-320). Adelaide, SA: AAMT, Inc.

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

Watson, J.M. (2002). Lessons from variation research II: For the classroom. In M. Goos & T. Spencer (Eds.), *Mathematics – making waves* (Proceedings of the Nineteenth Biennial Conference of the Australian Association of Mathematics Teachers Inc., Brisbane, pp. 424-432). Adelaide, SA: AAMT, Inc.

Watson, J.M. (2004). Quantitative literacy in the media: An arena for problem solving. *Australian Mathematics Teacher, 60*(1), 34-40.

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

Watson, J.M. & Chick, H.L. (2004). What is unusual? The case of a media graph. In M. Johnsen-HØines & A. B. Fuglestad (Eds.), *Proceedings of the 28th annual conference of the International Study Group for the Psychology of Mathematics Education* (Vol. 2, pp. 207-214). Bergen, Norway: PME.

Watson, J.M. & Kelly, B.A. (2003a). 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. & Kelly, B.A. (2003b, December). The vocabulary of statistical literacy. In *Educational Research, Risks, & Dilemmas: Proceedings of the joint conferences of the New Zealand Association for Research in Education and the Australian Association for Research in Education* [CD-ROM]. Auckland, New Zealand.

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. (1999). The development of concepts of average. *Focus on Learning Problems in Mathematics, 21*(4): 15-39.

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

Paper Appendix: Code Book for Scoring

This appendix includes all items on the surveys and the Code Book used for scoring. Numbers in square brackets refer to the form and item placement on that form; e.g., [F1, 13c] refers to Form 1, Item 13c. Some items appeared on both forms, e.g., [F1,1; F2,1] refers to Item 1 on both Form 1 and Form 2.

<u>Underline</u> indicates an item in the Average and Chance (AC) subgroup.
Italics indicates an item in the Sampling and Inference (SI) subgroup.
Highlight indicates an item in the Graphing and Variation (GV) subgroup.
Multiple indicators relate to two subgroups.

<u>AMOD</u> [F2, 11a]

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

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.

Code	**Description**	**Examples**
4	Statistical and contextual responses incorporating both positive and negative aspects of method	*"Yes, because it is using the most common weight for the item, however, it does not look at the other weights and if the most common weight was an extreme value it would be inaccurate"*
3	Statistical response – positive evaluation	*"Yes, majority of times it was weighed at 6.3"*
	Statistical response – negative evaluation	*"No, doesn't take into account the other weights"*
2	Claims of inaccuracy but with no statistical response – negative evaluation	*"No, the mode might weigh more than the others"* *"No, it's not accurate"* *"No, three people might have weighed wrong"*
	Claims of accuracy but with no statistical response – positive evaluation	*"Yes, it's the most accurate"* *"Yes, it's the average weight"*
1	Recommendation of other methods	*"No, he should have added them up and divided by 9"*
	Tautological but positive evaluation based on majority or "most common"	*"Yes, because he is using the most common"*
	Methodological reasons – positive and/or negative evaluations	*"Yes, it's easy"* *"No, too much calculating"*
0	No reason or apparent logic regardless of evaluation	
	No response	

- Scored independently of AMED, AMEA, AOUT and AVER

AMED [F2, 11b]

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.

Code	**Description**	**Examples**
4	Statistical and contextual responses incorporating both positive and negative aspects of method	*"Yes because it uses all the information presented, however, it does not take into account the extreme values when working it out"*
3	Statistical response – positive evaluation	*"Yes, it is using all the information"* *"Yes, it won't be too big or not too small"*
	Statistical response – negative evaluation	*"No, there's only one 6.2"*
2	Claims of inaccuracy but with no statistical response – negative evaluation	*"No, it's not very accurate"*
	Claims of accuracy but with no statistical response – positive evaluation	*"Yes, it's the most accurate way to do it"*
1	Recommendation or preferences for other methods	*"No, not as good as Ben's"* *"No, need to find the average"*
	Preference for the median over others without reason	*"Yes, better than Ben's"*
	Tautological but positive evaluation based on ordering or the "middle value"	*"Yes, because she's using the middle value"* *"Yes, because she's putting them in order"*
	Methodological reasons – positive and/or negative evaluations	*"Yes, as long as you do it right"* *"No, too slow to do"*
0	No reason or apparent logic regardless of evaluation	
	No response	

- Scored independently of AMOD, AMEA, AOUT and AVER

AMEA [F2, 11c]

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.

Code	**Description**	**Examples**
4	Statistical & contextual response recognising the outlier and the need to exclude it	*"Yes, but there is an extreme measurement"* *"It is not that good as it needs to get rid of the 15.3"*
3	Statistical response – positive evaluation	*"Yes, he is getting the average"* *"Yes, it includes all information"*
	Statistical response – negative evaluation	*No, cause it is too high* *No, cause the answer is too large*
2	Claims of inaccuracy but with no statistical response – negative evaluation	*"No, there is no 7.18"*
	Claims of accuracy but with no statistical response – positive evaluation	*"Yes, it is accurate"* *"Yes, cause it gives an exact answer"*
1	Recommendation or preferences for other methods	"No, the others are better"
	Tautological but positive evaluation based on the algorithm	*"Yes, because he is adding and dividing"*
	Methodological reasons – positive and/or negative evaluations	*"Yes, it is quick and easy"*
0	No reason or apparent logic regardless of evaluation	
	No response	

- Scored independently of AMOD, AMED, AOUT and AVER

AOUT [F2, 11d]

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.

Code	Description	Examples
3	Statistical and contextual responses recognising the outlier as a mistake	*"Yes, because 15.3 is nowhere near the others"* *"Yes, because she took out the extreme value"*
2	Statistical response – negative evaluation	*"No, needs to be done with all the information"* *"No, it'll make the average lower"*
	Statistical uncertainty – positive response with a somewhat uncertain explanation	*"It's okay, but there's no 6.17"* *"Good, but it wouldn't be very accurate"* *"Good, but should remove a 6.0 as well"*
1	Claims of inaccuracy but with no statistical response – negative evaluation	*"No, wouldn't add up right"* *"No, it won't work"* *"No, it's not accurate"*
	Tautological but negative evaluation based on excluding the outlier	*"No, because she left one out"* *"No, you can't just leave one out"*
0	Recommendation or preferences for other methods, comparison with others	*"No, I think the majority is still better"* *"No, just as bad as Ben and Jane"*
	Preference for mean (excluding outlier) over others with no reason	*"Yes, this is the best one"*
	Methodological reasons – positive and/or negative evaluations	*"No, it's too hard"*
	No reason or apparent logic regardless of evaluation	
	No response	

- Scored independently of AMOD, AMED, AMEA and AVER

AVER [F2, 11e]

Which of the ways described above would you use? Why?

Code	Description	Examples
4	Reason based on statistical and contextual information	*"May, because it is obvious that the 15.3 was a mistake when you look at the others"*
3	Statistical reasons only	*"Ben, he is using the majority"* *"Ron, he is getting the average – totally half"*
2	Reason based on claims of accuracy and exactness without a statistical argument	*"Ron, it is accurate"* *"Ron, he is doing it properly"* *"Ron, his way is the correct way to summarise"*
1	Reasons based on a tautological argument	*"Ben, he is finding the most common"* *"Jane, she is finding the middle value"*
	Reasons based on a personal preference of which is better	*"Jane, it is better than the others"*
	Reasons based on methodological implications	*"Ben, it is more convenient"* *"Jane, it's easier to understand"*
0	Idiosyncratic reason or no apparent logic	*"Jane, it's a good way"* *"May, because she has some left over"*
	Preferred method stated, no reason	*"Ben"*
	No response	

- Scored independently of AMOD, AMED, AMEA and AOUT

CAR [F1, 8; F2, 7]

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

Which car should Mrs. Jones buy?

A. 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. ☐

B. 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. ☐

C. 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	**Description**
3	B – Honda …
2	C – It doesn't matter …
1	A – Toyota
0	No response

CON1 [F2, 5a]

Imagine you are playing a game where you throw a coin 4 times. How many tails do you think might come up?

Code	**Description**
2	2 tails or 50%
1	Any other number
0	No response

- Scored independently of CON2 and CON3

CON2 [F2, 5b]

Imagine you are playing a game where you throw a coin 4 times. Imagine that 100 people played the game. In the table below, fill in how many people you think will get each number of tails.

Number of tails	Number of people getting the number of tails
0	
1	
2	
3	
4	
Total	100

Code	Description	Examples
3	Appropriate variability displayed incorporating probability and distribution	*Does not give a probabilistic example like Code 2* *All predictions made within the desired ranges below* *Prediction for 2 tails=30-45; Prediction for 1 and 3 tails=18-33; Prediction for 0 and 4 tails=3-10* *"10, 20, 35, 25, 10"*
2	Too narrow or no variation - extreme probabilistic outcome	*Prediction for 2 tails = 37 or 38 (37.5)* *Prediction for 1 and 3 tails = 25* *Prediction for 0 and 4 tails = 6 or 7 (6.25)*
	Primitive understanding of proportion – 50% chance for 2 tails	*"0, 25, 50, 25, 0" or "5, 20, 50, 20, 5" or "10, 15, 50, 15, 10"*
1	Assumes equality for all options	*"20, 20, 20, 20, 20, 20"*
	Seemingly random prediction	*"10, 30, 40, 1, 19"*
0	Does not add to 100 – Possible misinterpretation	*"10, 10, 10, 10, 10"*
	No response	

- Scored independently of CON1 and CON3

CON3 [F2, 5c]

Explain why you think these numbers are reasonable.

Code	Description	Examples
4	Reasoning reflecting aspects of chance and probability	*"Because out of 100 people not very many are going to get 0 tails or 4 out of 4 tails, it is most likely they'll get 2 tails and 2 heads because there's 2 different faces and 4 chances"*
3	Implicit understanding of chance and probability, sometimes mentioning 50% or 1/2 chance	*"I think it would most likely be even because there's a 50% chance it will come up"*
2	Reasoning reflecting an even or equal chance for all numbers	*"Because they all have equal opportunity"*
	Anything can happen, chance and luck	*"It's random so no one knows what will come up"*
1	Idiosyncratic reasoning and personal beliefs	*"Because tails never fails"*
0	No reason	

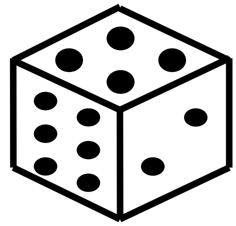
- Scored independently of CON1 and CON2

DI1A [F1, 4a; F2, 4a]

Consider rolling a normal six-sided dice.

Tick the box which shows which is easier to get:

A. 1 ☐
B. 3 ☐
C. 6 ☐
D. All the same. ☐



Code	Description
1	All the same chance
0	1, 3, 6

- Scored independently of DI1B

DI1B [F1, 4b; F2, 4b]

Explain your answer.

Code	Description	Examples
4	Includes quantitative explanation using of ratios, percentages, and odds	*"You have a 1 in 6 chance of rolling any number"* *"There is a 16.5% chance of getting it"* *"The probability is the same, 1:6"*
3	Equal likelihood aspect	*"You have an equal chance"* *"All the numbers are equally easy to roll"* *"Because all the sides are the same weight and size"* *"Because there is only one of each number on the dice"*
2	Component of uncertainty without the expression of equality	*"There's no saying which one it might land on"* *"It's all chance"*
1	Idiosyncratic beliefs	*"Because there is less dots, the 6 automatically goes down"* *"It's a lower number"* *"Just from the way I roll it"* *"When I roll the dice I get them about the same"* *"The number 1 will have more chance"*
0	No reason	

- Scored independently of DI1A

DI2A [F1, 5a]

Imagine you threw a dice 60 times. In the table below, fill in how many times you think each number came up.

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

Code	Description	Examples
3	Appropriate variability	*Predictions add to 60; Spread between 4 and 12 inclusive* *Does not use large singular number >21* *Does not use multiples of five to add to 60* *"9, 8, 11, 12, 10, 10"*
2	Too narrow or too wide variability	*Predictions add to 60; Spread <4 or >12* *"6, 8, 10, 20, 11, 5"*
	Strict probability outcome	*"10, 10, 10, 10, 10, 10"*
	Uses multiples of five to add to 60	*"5, 15, 10, 15, 10, 5"*
1	Adds to 60 but uses large singular number >21	*"5, 40, 3, 4, 2, 5"* *"25, 10, 10, 6, 4, 5"*
0	Does not add to 60 – Possible misinterpretation	*"31, 5, 10, 29, 10, 10"* *"1, 60, 0, 0, 0, 0"*
	No response	

- Responses that add to "58-62" may be classified as adding to 60
- Scored independently of DI2B

DI2B [F1, 5b]

Explain why you think these numbers are reasonable.

Code	Description	Examples
4	Reasoning reflecting aspects of chance and variability	*"Because they are random numbers"* *"Because they're all pretty equal"* *"They are all about the same number"*
3	Reasoning reflecting strict probability and classical chance or reasoning reflecting aspects of die geometry	*"They all have the same ratio"* *"They all have the same chance of coming up"* *"There is a 1 in 6 chance of getting each number"* *"The die is even on all sides"* *"All the numbers have one each"* *"They are on a square cube that has 1cm sides"*
	Computational reasoning	*"Because 60 divided by 6 equals 10"* *"Because they all equal 60"* *"10 is 1/6 of 60"*
2	Anything can happen	*"50-50 chance, luck of the draw"*
1	Personal beliefs and idiosyncratic reasoning	*"It's the way I roll the dice"* *"Because some of the dots have more weight"* *"Because they look like they're reasonable"*
0	No reason	

- Scored independently of DI2A

<u>HAT</u> [F1, 12]

A mathematics class has 13 boys and 16 girls in it. Each student'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.

Tick **<u>one</u>** statement that best expresses what you think could happen:

A. A boy's name will be picked because boys are usually chosen in maths. ☐

B. It is likely that a girl's name will be picked because there is a 16/29 chance of this. ☐

C. Anything could happen – it's just luck what comes out. ☐

D. A girl's name will be picked because there are 16 girls and 13 boys. ☐

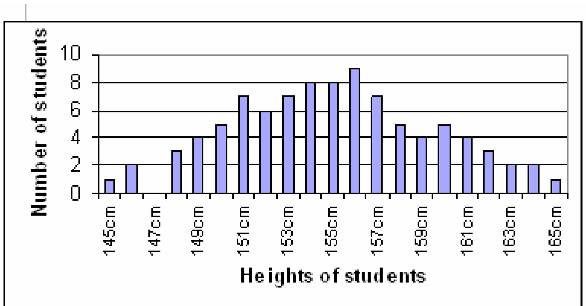
E. There's a 50:50 chance of getting a boy or a girl. ☐

Code	**<u>Description</u>**
3	B – 16/29 (girl more likely)
2	D – 16 versus 13 (girl)
1	E – 50:50 chance (either)
	C – Anything could happen
0	A – Boy's name usually chosen
	No response

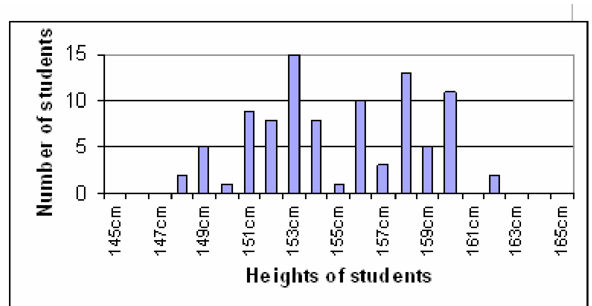
HGT1 [F1, 11a]

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

School A



School B



How many students are 156 cm tall in each school?

School A ______________________ School B _______________________

Code	**<u>Description</u>**
1	Reads correct values from graph, A=9 and B=10
0	Any other value/s or written responses
	No response

• Scored independently of HGT2 and HGT3

HGT2 [F1, 11b]

Which graph shows more variability in students' heights?

Code	Description
2	School A
1	School B
	The Same
0	Any other written response
	No response

- Scored independently of HGT1 and HGT3

HGT3 [F1, 11c]

Explain why you think this.

Code	Description	Examples
4	Mentions explicitly the wide range/spread and/or variety of heights	*"School A has more of each height. School B has lots of one".*
3	Mentions implicitly the wide range/difference of heights	*"A, because they have at least one person in very height except 147 cm"*
2	Focuses on the size of the individual bars without regard to what they represent	*"B, because students are all different heights"*
	Focuses on the number of individual bars without regard to what they represent	*"A, because the graph in School A shows more lines"*
1	Misapplies notion of variability and focuses on an average height	*"A, a lot of people are around about the same height in that school"*
0	Focuses on the content of data	*"A, more people are bigger"*
	Aesthetic appearance and personal preference and graph lay out	*"A, easier to see which is taller"*
	Misreads data or misinterprets question, unjustified statements	*"There are more people in school A"*
	No reason	

- Scored independently of HGT1 and HGT2

HGT4 [F1, 11b and c] [a combined code for the two parts]

Code	Description
2	Recognises variability and can apply this in context
1	Misapplies variability focussing on height of columns or graph features
0	No reason

HSE1 [F1, 9a]

This article appeared in a newspaper.

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.

What does "average" mean in this article?

Code	**Description**	**Examples**
2	Describes the central tendency for a data set or the method of obtaining the average from a data set (sometimes related to context)	*"The word average means that out of all wage earners this is the wage most people earn"* *"Average, a lot of numbers added up and divided by how many numbers there were at the start"*
1	Single idea not related to context	*"Average means about the same as everyone or anything"*
0	No idea of central tendency, often tautological	*"The average wage earner in Australia"* *"An average means it is small and good for little families"*
	No response	

- Scored independently of HSE2 and HSE3

HSE2 [F1, 9b]

What does "median" mean in this article?

Code	**Description**	**Examples**
2	Describes the central tendency for a data set or the method of obtaining the average from a data set (sometimes related to context)	*"The median means that the middle range house is around $88,200"* *"Middle score in size order, e.g., 3 houses cost $50,000 + $100,000, + $150,000; median equals, $100,00"*
1	Single idea not related to context	*"Median means normal"*
0	No idea of central tendency, often tautological	*"Median is when something rises"*
	No response	

- Scored independently of HSE1 and HSE3

<u>HSE3</u> [F1, 9c]

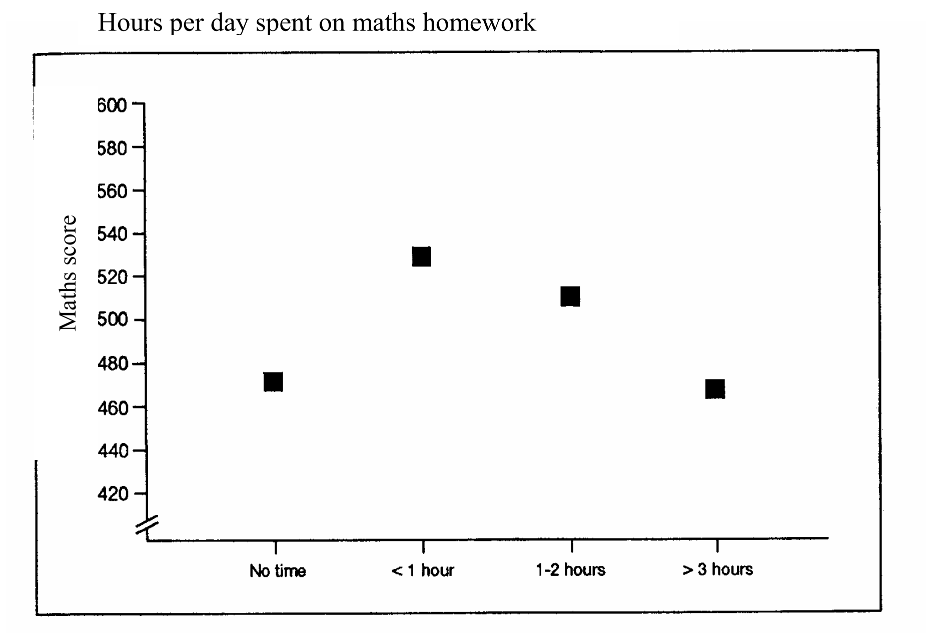
Why would the median have been used?

Code	Description	Examples
2	Mention of outliers or extreme values	*"Because there may be one or two very expensive houses down in Hobart which would take the average out of proportion"*
1	Usefulness or fairness (without explicit mention of outliers)	*"It is more accurate than the mean"* *"Because it is most representative of the type of house the average Australian family can afford"*
0	Response that does not refer to question (e.g., language/price)	*"So it can sell more things"* *"To point out a statistic"*
	No response	

- Scored independently of HSE1 and HSE2

HWKA [F2, 10a]

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.



What is the maths score for students who spend 1-2 hours per day on mathematics homework?

Code	Description
1	Correct response (510-520)
0	Any other written response
	No response

- Scored independently of HWKB and HWKC

HWKB [F2, 10b]

What does the graph tell you about maths homework time and maths scores?

Code	**Description**	**Examples**
3	Appreciation of association between variables with implications and conclusions drawn from the data	*"Those better at maths finished early"* *"You are more likely to get good grades if you do an hour"* *"Doing 3 hours can be just as bad as doing 1 hour"*
2	Appreciation of association between variables without implication – basic data reading	*"Those who did 1 hour got the highest scores"* *"Scores go up to 1 hour and then decrease after that"*
	Causation between variables assumed	*"If you do 1 hours homework you will get a better score"*
1	Incorrect association between variables	*"More homework you do the better your scores"*
	No association identified	*"It doesn't matter how much homework you do"*
	Uni-dimensional	*"It shows the total maths scores / total homework time"*
0	Misinterpretation of data	*"More students studied for an hour"*
	Idiosyncratic	*"People like to take time"*
	No response	

- Scored independently of HWKA and HWKC

HWKC [F2, 10c]

Gives some reasons why the graph has the shape it does?

Code	**Description**	**Examples**
2	Creates a hypothesis from the graph based on student abilities or other possible contributing factors	*"The struggling students take more time to do their homework"* *"The drop in scores is due to stress and frustration"*
1	Comment involving data reading – no identifiable hypothesis	*"The people that studied 1 hour got the highest mark, those that did not study or did too much got a lower mark"*
	Comment about the variables – no hypothesis	*"The shape is a result of the different times"*
0	Comment about graph type, and /or its appropriateness	*"Easy to read"* *"It's the most appropriate type of graph"*
	Idiosyncratic response	*"Because it is a square graph"*
	No response	

- Scored independently of HWKA and HWKB

MV10 [F1, 13a; F2, 9a]

A class wanted to raise money for their school trip to Movieworld. They could raise money by selling raffle tickets for a Nintendo Game system. Before they decided to have a raffle they wanted to estimate how many students in the whole school would buy a ticket. 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.
How many students would you survey? How would you choose them? Explain your answers.

Code	**Description**	**Examples**
3	Random and representative methods	*"Probably 10 from each class. I would choose them by picking out of a hat"* *"300 students, 50 from each class chosen randomly"*
	Random methods	*"Put all 600 student names into a hat and draw out 65 names"* *"About 300 out of a hat"*
2	Representative methods (no random device)	*"I would survey 60 students, 10 from each grade and pick 5 boys and 5 girls"* *"I would choose 50 girls and 50 boys"*
1	Non-representative methods	*"I would survey 10 from each class. I would choose them by the richest in the class"* *"20 people from each class, grades 4-6, random selection on the computer"*
	Method only, no sample	*"Pull their names out of a hat"* *"You would choose by kindness"*
	Sample only, no method	*"I would survey 30 people"* *"I would survey 300 because it is half of 600"*
	Entire population	*"All of them"* *"Take 100 from each class"*
0	Misinterpretation (selling tickets)	*"I would sell raffle tickets to each and everybody in the class"*
	No response	

- Scored independently of MV11, MV12, and MV13

MV11 [F1, 13b; F2, 9b]

Shannon got the names of all 600 students in the school and put them in a hat. Then she pulled out 60 names. What do you think of Shannon's survey? Explain your answer.

Code	**Description**	**Examples**
3	Appropriate statistical appraisal focusing on random and range	*"Good, because he/she is picking them out randomly"* *"Good, she can get a whole range of students"*
2	Appropriate, non-central appraisals focusing on fairness, sample size and methodology	*"Good, that is a fair way of doing it"* *"Good, 60 is not too many people or too less"* *"Good, it would not take too long"*
1	Inappropriate criticisms focusing on being too random, inaccurate, small sample size, fairness and methodology	*"Not sure, because they could all be in the same grade"* *"Bad, because it wouldn't give an accurate answer"* *"Bad, because it is not really a big enough sample size"* *"Not sure, because it wasn't fair for the rest of the kids"* *"Bad, because it's too time consuming"*
0	No reason or apparent logic, regardless of opinion	*"Good, because it was a good way to do it"* *"Not sure, because it could be good or bad"*
	Misinterpretation (ticket selling)	*"Bad, because you only have one Nintendo"*
	No response	

- Scored independently of MV10 and MV12

MV12 [F1, 13c]

Jake asked 10 students at an after-school meeting of the computer games club. What do you think of Jake's <u>survey</u>? Explain your answer.

Code	Description	Examples
3	Appropriate and fully integrated criticism focusing on sample size and non-representativeness	*"Bad, because he got a small opinion from a group of people with the same interests"* *"Bad, not enough people and they're game fanatics"*
2	Appropriate criticisms, focusing on sample size or non-representativeness (not both)	*"Bad, wasn't enough people"* *"Bad, because most of them would say the same thing"* *"Bad, because he only asked 10 people"*
	Statistical uncertainty focusing on appropriate criticisms	*"Not sure, because not many different people would go there"* *"Not sure, because he only asked 10 people"*
	Appropriate, non-central criticisms focusing on fairness, methodology and implications	*"Bad, because not everyone has got a chance"* *"Bad, because they might not do the survey"* *"Bad, they won't tell you the majority"*
1	Inappropriate appraisals focusing on creating bias, sample size, methodology and implications	*"Good, he picked people who would be interested"* *"Good, he didn't ask so many people"* *"Good, it was quick and easy and not taking too long"*
0	No reason or apparent logic, regardless of opinion	*"Bad, because I think it is a bad idea"* *"Good, because Jake was being nice"*
	Misinterpretation (ticket selling)	*"Good, because he's starting to raise money"*
	No response	

- Scored independently of MV10 and MV11

MV13 [F2, 9c]

Claire set up a booth outside 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 <u>survey</u>? Explain your answer.

Code	Description	Examples
3	Appropriate criticism focusing on range, variation and overall non-representativeness	*"Bad, they could all be in the one grade or the same sex"* *"Bad, not everyone goes to the tuckshop"* *"Bad, only children interested would fill out the survey"*
2	Statistical uncertainty focusing on appropriate criticisms	*"Not sure, cause only people who could be bothered did it"* *"Not sure, it would depend on who came"*
	Appropriate, non-central appraisals focusing on sample size	*"Good, that is a good number of people"* *"Good, a lot of answers"*
1	Inappropriate appraisals focusing on the assumption of range and variation, fairness, free choice and methodology	*"Good, she should have a range"* *"Good, because everyone gets a chance"* *"Good, she is not forcing anyone"* *"Good, the tuck shop would have a lot of people"*
	Inappropriate criticism focusing on sample size	*"Bad, she could probably get more people"* *"Bad, not enough people"*
0	No reason or apparent logic, regardless of opinion	*"Bad, truckies wouldn't have time"* *"Good, because she had a better system"*
	Misinterpretation (ticket selling)	*"Good, because they would have money"*
	No response	

- Scored independently of MV10 and MV11

ODDS [F1, 7]

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

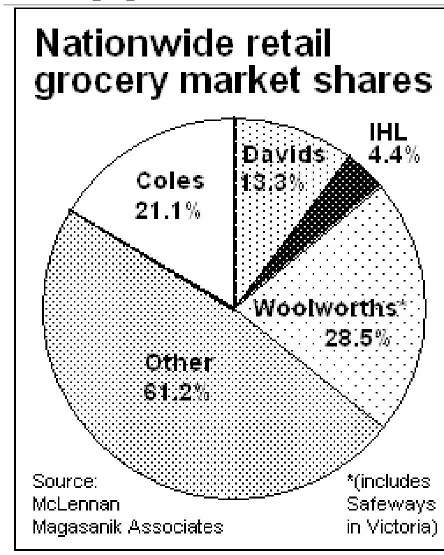
What does odds at 7:2 mean in this headline about the North against South football match? Give as much detail as you can.

Code	**Description**	**Examples**
3	A ratio measurement of chance identified with an appropriate part-part ratio and an expression of the direction favoured in interpreting the odds	*"North hasn't got a good chance, like a ratio 7:2. A 2/9 chance to win"* *"Out of nine in a scale from one, there is a 7 chance that the other team will win and 2 chances that North will win"*
	Identification of odds as part-part ratio, and a conversion to an appropriate part-whole ratio	*"Experts believe North has a 7/9 chance of winning and South has 2/9. This means that North is likely to win, but South still has a small chance"*
2	Quantified measure as percent, chance or a part-whole ratio but inappropriate	*"7:2 means they have a 2 out of 7 chance to beat South"* *"Basically it means in 7 games South are likely to win 5, and North to win 2"*
	Unconventional quantified ratio measure	*"It means 7 out of 2 chances of winning or 3.5 out of 1"* *"7/10 chance that South will win and 2/10 chance North will win, there is 1/10 there will be a draw"*
	Uncertainty interpreted in a primitive quantitative manner	*"North has 7% chance to win and 2% chance to lose"* *"North has 2 chances of winning and South has 7 chances"* *"North has a 7 in 2 chance of beating South"*
	Single idea of uncertainty expressed as chance	*"That's North's chances of winning"* *"It means chances are 7 to 2"* *"7:2 means the chances of North winning isn't very high"*
1	Tautological responses or idiosyncratic answers	*"I think just 7:2 and that it's the coach saying to win"* *"It means like 7 to 2"* *"7am – 2pm"*
0	No response	

PIEA [F2, 8a]

This pie chart appeared in a newspaper.

What is the newspaper trying to tell its readers in this pie chart?



Code	Description	Examples
3	Appropriately interprets the information on the chart and understands proportions	*"People have most shares in 'other' companies"* *"It shows you what companies people invest most in"*
2	Inappropriately interprets the chart but understand the proportions	*"How many people shop there"*
1	Draws extremely basic and non-central conclusions about the intent of the graph	*"Shows different stores"* *"Woolworth's is going well"*
	Tautological response or reads percentages direct from chart	*"Grocery market shares"* *"IHL has 4.4%, Woolworth's has …" etc.*
0	Idiosyncratic responses and literal interpretations about pies	*"Other stores sell more pies"*
	No response	

- Scored independently of PIEB

PIEB [F2, 8b]

Is there anything unusual about the pie chart?

Code	Description	Examples
2	Detects error	*"There is more than 100% there's 128.5%"*
1	Focuses on non-central issues	*"I thought Coles and Woolworths were bigger than that"*
0	Focuses on presentation or fails to give an example	*"It has got other shaded but IHL is black and its writing is on the outside of the graph"*
	Does not notice anything unusual	*"No it looks fine"*
	No response	

- Scored independently of PIEA

RAND [F1, 2]

What does "random" mean?

Code	Description	Examples
3	Sophisticated definition reflecting lack of structure, pattern and/or uncertainty	*"To do something that isn't planned or organised"* *"Something isn't picked by choice, but by chance"* *"Picked without order or any distinct pattern"*
2	Simple definition emphasising chance and unpredictability	*"Completely out of the blue"* *"Not specifically chosen"*
	Simple definition with an emphasis on "any"	*"To just pick anything"* *"Picking stuff out in no/any order"*
	Simple definition emphasising "no thinking" or "no looking"	*"Not knowing what is going to happen"* *"Just pick something without thinking"*
1	Single idea emphasising on picking or choosing something	*"To take a non-accurate pick or guess"* *"To just pick someone"*
	Single idea emphasising "guessing"	*"Just a guess"* *"You don't know what's going to happen"*
	Example only	*"Tattslotto"* *"Random Breath Testing"*
0	Intuitive example - tautological	*"Where you pick something from random"* *"Not always happens"*
	Idiosyncratic	*"To come quickly or someone's held at random"* *"One thing which is something"*
	No response	

RASH [F2, 13]

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	**Description**
2	D – 15 out of 100
	D & A or D & E (appropriate dual selection)
1	E – Hardly any chance
	A – Good chance
0	B – 15% of dosage
	C – 15% of skin
	Multiple selections excluding those mentioned in Code 2
	No response

SAMP [F1, 1; F2, 1]

What does "sample" mean?

Code	**Description**	**Examples**
3	Small part representing a whole, purpose to test	*"To take something out of something bigger to test"* *"To take a part of something and look into it"* *"To take a small piece of something and test it"*
2	Small part, not all	*"You have a small piece of something"* *"To take / have a portion of something"*
1	Single idea – A part or a test	*"Trying it out"* *"It means a little amount"*
	Example only	*"Sample means when you take a blood sample"* *"Getting a sample of water"*
0	Attempting to explain - tautological	*"A kind of example"* *"It's a sample, not to write on it or use it"*
	Idiosyncratic	*"When you look through something"* *"That something is easy"*
	No response	

SPT1 [F2, 3a]

A primary school had a sports day where every student 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

How many girls chose tennis?

Code	**Description**
1	Correct response (15)
0	Any other written response
	No response

- Scored independently of SPT2, SPT3, and SPT4

SPT2 [F2, 3b]

What was the most popular sport for boys?

Code	**Description**
2	Identification of two modes (Soccer & Tennis)
1	Identification of one mode (Soccer or Tennis)
0	Any other written response
	No response

- Scored independently of SPT1, SPT3, and SPT4

SPT3 [F2, 3c]

How many children were at the sports day?

Code	Description
1	Correct addition of values (125)
	Correct table reading (50 boys, 75 girls)
0	Any other written response
	No response

- Scored independently of SPT1, SPT2, and SPT4

SPT4 [F2, 3d]

One of the tennis players was late.
Was this player a boy or a girl? Explain your answer.

Code	Description	Examples
3	Statistical prediction based on chance with an element of uncertainty	*"Most likely boy, more play tennis"* *"Maybe a girl, there is more"* *"Cannot be sure but it could be a boy/girl – there's more"*
	Statistical prediction based on the majority – no uncertainty present	*"Boy, because more play tennis"* *"Girl, there are more girls overall"*
2	Prediction based on inverse majority or balancing	*"Girl, because they have a smaller group (tennis)"* *"Boy, there are less boys overall"*
	Anything can happen, sometimes with no graph interaction	*"Could be either, 50/50 chance"*
	Non-statistical predictions based on patterns (uncertainty may be present)	*"Girl, to make their numbers even for tennis"* "Boy, they have an even number"
1	Story telling and assumptions based on context	*"Boys are always late"* *"She slept in"*
	Reluctant to predict, or interact with graph (sometimes with no reasoning)	*"Not enough information"* *"The graph does not carry that information"* *"Can't tell …"*
0	Misinterpretations and idiosyncrasies	*"Girl because 20 is the highest amount"*
	Prediction but no reason	*"Boy"*
	No response	

- Scored independently of SPT1, SPT2, and SPT3

T2X2 [F1, 10]

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.

Code	**Description**	**Examples**
5	Critically examines all information (implies look at all cells) and/or correctly states proportions and percentages	*"No, because if they surveyed 150 people who smoked and didn't smoke it would be the same – 90-60 each"*
4	Agrees or disagrees with statement, citing evidence from the data focusing within, between or across columns (2 or 3 cells)	*"60 no smoking lung disease, 60 smoking no lung disease"* *"Difference between 60 (smokers) and 40 (no smokers) is big"* *"Lung disease – 60 no smoking, 90 smoking, difference not much"*
3	Agrees or disagrees with statement citing evidence from a single cell in the data	*"90 who smoked got lung disease"* *"There is a lot of non-smokers with lung cancer"*
2	Critical analysis of potential survey methods and the limitations of the methods used to collect the data	*"We don't know how they collected their sample. It could be biased."*
1	Agrees or disagrees with statement, stating more/less chance, but with no explicit reference to the data	*"Yes, there is more chance if you smoke"*
	Positive or negative statements justified with knowledge of the content area, but not based on data	*"No, it could be hereditary."* *"Yes, smoking is known to cause lung cancer – it is a fact."*
0	Positive or negative statement with no justification	*"Yes, because smoking is bad."*
	No reason	*"Yes"*
	No response	

TOSA [F2, 12a]

The captain of a sports team chose heads each time and lost 8 out of 8 coin tosses. What should he choose for the next match?

Tick one box to show your answer.

A. Heads ☐
B. Tails ☐
C. Doesn't matter ☐

Code	**Description**
1	Doesn't matter
0	Heads, Tails, No response

- Scored independently of TOSB

TOSB [F2, 12b]

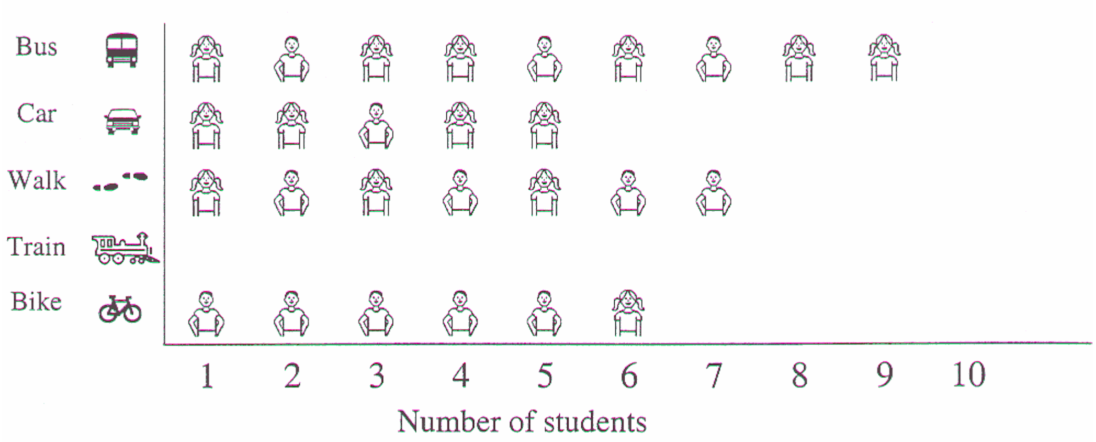
Explain your answer.

Code	Description	Examples
4	Includes quantitative explanation using of ratios, percentages, and odds	*"C, it's a 50% chance on landing on heads"*
3	Equal likelihood aspect	*"C, both sides have an equal chance"*
2	Component of uncertainty without the expression of equality	*"C, it's just the luck of the draw, it's not his fault if he lost every time"*
1	Idiosyncratic beliefs	*"A, because if he changes he might get it wrong"*
0	No reason	

- Scored independently of TOSA

TRV1 [F1, 14a]

How children came to school one day



How many children walked to school?

Code	Description
1	Correct response (7)
0	Any other written response
	No response

- Scored independently of TRV2 and TRV3

TRV2 [F1, 14b]

A new student came to school by car. Is the new student a boy or a girl? Explain your answer.

Code	Description	Examples
3	Statistical prediction based on chance with an element of uncertainty	*"You can't know for sure, but 4 out of 5 students that come by car are girls, so it is probable"* *"Girl, because there is a 4/5 chance that it was a girl"*
2	Statistical prediction based on the majority – no uncertainty present	*"Girl, the graph shows more girls go by cars than boys"* *"Girl, because there are 14 girls and 13 boys"*
	Prediction based on inverse majority or balancing	*"Boy, because there are already 4 girls that come by car"* *"Boy, because then there would be an even class"*
1	Non-statistical predictions based on patterns (uncertainty may be present)	*"It could be a girl, more likely a boy because of the pattern"* *"Boy, the graph goes – girl, girl, boy, girl, girl …"* *"Girl, she was the last one on the row"*
	Anything can happen, with no graph interaction	*"You can't tell, because it could be either"* *"Boy or girl, it's a 50/50 chance"*
0	Inappropriate or idiosyncratic response, sometimes with no reason	*"The graph doesn't give enough information"* *"Girl, because she is nervous about her first day at school"* *"Boy, I just guessed"*
	No response	

- Scored independently of TRV1 and TRV3

TRV3 [F1, 14c]

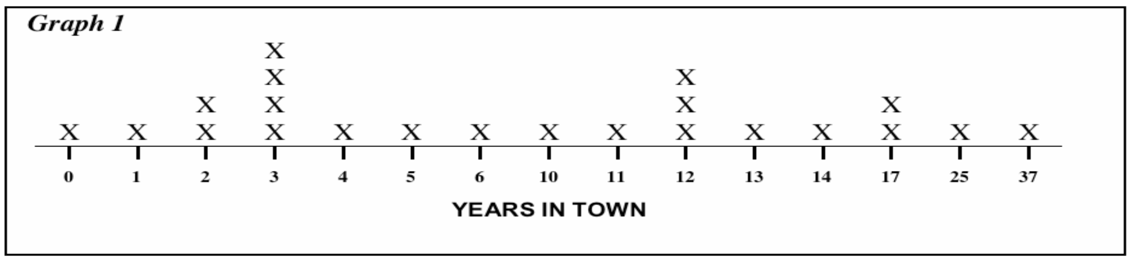
Tom is not at school today. How do you think he will come to school tomorrow? Explain your answer.

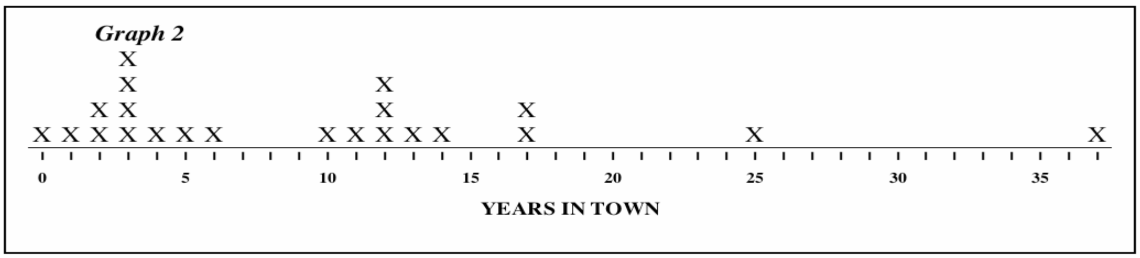
Code	Description	Examples
5	Statistical prediction based on chance with an element of uncertainty	*"Most likely bus, most children catch the bus"* *"Most likely by bike, because most of the boys ride to school"* *"Probably bus, because 1/3 of the children caught it today"*
4	Statistical prediction based on the majority – no uncertainty present	*"Bike, more boys use bikes to get to school"* *"Bus, because the majority go by bus"*
3	Prediction based on inverse majority or balancing	*"Train, because there is no one on the train today"* *"Car, because there is only one boy"*
	Anything can happen, sometimes with no graph interaction	*"You don't know because the graph is only for one day"* *"It could be anything, depends on how he wants to get there"*
2	Non-statistical predictions based on patterns (uncertainty may be present)	*"Walk, he's the only boy at the end of the walk line"* *"Car or bus, because they fit the patterns"* *"Bus, because there is 10 numbers and 9 people"*
	Story telling and assumptions based on context	*"Train, because he lives a long way away"* *"Car, so he doesn't catch a cold again"*
1	Reluctant to predict, or interact with graph (sometimes with no reasoning)	*"The same way he does any other day"* *"I don't know, because there is no information about Tom"* *"Walk, because I just think so"*
0	Egocentricism or misinterpretations	*"Bus, because that's how I get to school"* *"Yes, he will be feeling better after a day off"*
	No response	

- Scored independently of TRV1 and TRV2

TWNA [F1, 6a]

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.





What can you tell by looking at Graph 1?

Code	Description	Examples
3	Summarises data	*"They range from all years"* *"Not many families have stayed there for the same time"* *"The population is building each year"*
2	Basic data reading	*"3 and 12 have the most"* *"1 family had lived there 37 years"* *"The numbers along the bottom tell you how many years"*
1	Inappropriate interpretations	*"The town must not be well known"* *"They move after a couple of years"* *"There was a lot of people in that town"*
0	No response	

- Scored independently of TDIF and TSTY

TWNB [F2, 6a] [Order of graphs swapped in TWNA]

What can you tell by looking at Graph 1?

Code	Description	Examples
3	Summarises data	*"There are two main groups, 0-6 and 10-14"* *"The majority of people have lived there for up to 6 years"* *"That more people lived in town for 3 years"*
2	Basic data reading	*"It only goes in 5's"* *"4 families have been there for 3 years"* *"That someone has lived there for 37 years"*
1	Inappropriate interpretations	*"11 families lived in the town for 5.5 years"* *"They are likely to move after 15 years"* *"The number 3 lives there the longest"*
0	No response	

- Scored independently of TDIF and TSTY

TDIF [F1, 6b; F2, 6b]

What differences do you notice between Graph 1 and Graph 2?

Code	Description	Examples
3	Acknowledgement of the same data and/or correctly states the difference between the scales	*"There is no difference from graph 1 to graph 2 except that graph 2 shows the spaces where graph 1 doesn't"*
2	Attention to detail, comments about lay out, spread and accuracy	*"Graph 2 goes up in fives and Graph 1 doesn't"*
1	Aesthetic appearance, personal preferences	*"Graph 2 is harder to read because numbers are together, Graph 1 is easier to read because numbers are spread out"*
0	Reports incorrect differences about the data	*"There are more Xs in Graph 2"*
	Misinterprets question by reading data	*"Not many people have lived there for long"*
	No response	

- Scored independently of TWNA, TWNB, and TSTY

TSTY [F1, 6c; F2, 6c]

Which graph is better at presenting the information and "telling the story"? Explain your answer.

Code	Description	Examples
3	Statistically appropriate choice (Graph 2) with statistical reasoning	*"Graph 2 – you can see the difference between the years more clearly and the graph is more spaced out"*
2	Indifference (both the same)	*"They tell the same story – it's a different way of setting it out"*
	Statistically appropriate choice (Graph 2) with personalised reasoning	*"Graph 2 – because it is easier to understand"* *"Graph 2 – because they have set it out better"* *"Graph 2 – it is in a story-like way"*
1	Statistically inappropriate choice (Graph 1) with reasoning	*"Graph 1 – because they're all on the numbers"* *"Graph 1 – because it only has the time it needs"* *"Graph 1 – it's more understandable"*
0	Idiosyncratic reasoning or no reasoning	*"Graph 2 – because it's got more people"* *"Graph 1 – because they've lived here longer"* *"Neither – they both tell me nothing"*
	No reason / no response	

- Scored independently of TWNA, TWNB, and TDIF

VAR1 [F2, 2]

What does "variation" mean?

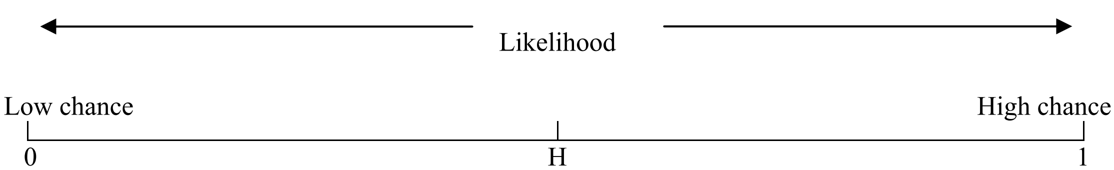
Code	**Description**	**Examples**
3	Sophisticated definition reflecting change, or a slight difference within something	*"There are different types of the same thing"* *"Slight change or difference"* *"When something doesn't stay the same all the time"*
2	Simple definition emphasising a difference between things	*"When something is different from something else"* *"Different things"*
1	Loose understanding with an attempt at a definition. Unable to fully grasp concept – isolated ideas	*"Wide range of something"* *"The answer can change"* *"Lots of choices"*
	Example only	*"It varies from bigger to smaller, taller to shorter"* *"Like it could vary in size, colour, type, style, etc."*
0	Intuitive example - tautological	*"A variation from something else"*
	Idiosyncratic	*"Something is hard"*
	No response	

WORD [F1, 3]

Here are eight chance words or phrases from newspaper headlines.

A	58 per cent success rate 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.



Code	**Description**	**Examples**
2	Comprehensive evaluation: Ordering and positioning on scale important	*Precise ordering -* *Lower 50% - B (<10%), E or G (<50%), G or E (<50%), H,* *Higher 50% - A (>50%, <70%), D, F or C, C or F (>85%)*
1	Partial evaluation	*At least 5 letters on the appropriate side of the 50% line (B must be lower than 25%) as shown in Code 3*
0	No evaluation	*Four or fewer letters on the appropriate side of the 50% line (B must be lower than 25%) as shown in Code 3*
	No response	

- Researchers' ordering is BEGHADFC