Implications of Differential Item Functioning in Statistical Literacy: Is gender still an issue?

Rosemary Callingham

University of Tasmania/ University of New England

Jane M. Watson

University of Tasmania

Abstract

Statistical literacy is a complex developmental construct requiring both mathematical skills and contextual understanding. The development of statistical literacy is an important objective of classrooms where the curriculum is approached through considering problems that require the active engagement of learners with relevant social material. Such approaches are often advocated for the middle years of schooling. Little attention has been paid, however, to the effects of these approaches on male and female students. This paper reports on a study that considers Differential Item Functioning (DIF) with respect to gender of questions on a statistical literacy scale derived from archived data. Multi-faceted Rasch models were applied to polytomous data to determine the interactions between gender and item. Three criteria were applied to the results: statistical significance, replicability and substantive explanation of DIF. The results suggested that although there was no overall difference in the average performance of male and female students, items requiring numerical responses or calculations were less difficult for male students and, conversely, items demanding written explanations were less difficult for female students. The implications of these findings for both assessment and teaching are discussed.

Introduction

New approaches to school curriculum, such as the "New Basics" in Queensland (Education Queensland, 2000) and the "Essential Learnings" in Tasmania (Department of Education, 2002) emphasise an integrated approach to the development of skills and knowledge. In order to be "active citizens" in the future, students are expected to be engaged "...in active participation in social, political and economic issues in communities, as well as in their school life and studies" (Education Queensland, 2000, p. 47). This implies that teaching and learning are contextually based, and that traditional academic skills need to be developed in such a way that connections are made between classroom experiences and the wider world (Freebody, Ludwig & Gunn, 1995; National Council on Education and the Disciplines (NCED), 2001).

The approaches to schooling suggested by this orientation may be particularly relevant in the middle years, typically grades 5 to 9, where there is concern about an increasing number of disaffected students (Eyers, Cormack & Barratt, 1993). A drop in performance in the middle years has been demonstrated in both literacy and numeracy (Hill, Rowe, Holmes-Smith and Russell, 1996). This has led to calls for improved approaches to curriculum focusing on relevance and activity based learning (Earl, 2000; Hill & Russell, 2000). Desired outcomes from schooling include critical thinking skills and the ability to analyse information presented through different media (Education Queensland, 2000).

Against this backdrop, the development of statistical literacy becomes a crucial element in students' education. Statistical literacy is a complex developmental construct that requires both mathematical skills and contextual understanding and, as such, makes demands on students' literacy and numeracy (Watson & Callingham, 2002). The need for students to develop an understanding of data presented in newspapers, on television, and on the Internet is now recognised as important by many (e.g., NCED, 2001; Steen, 1999; Gal, 1995). That this is not just a recent awareness is illustrated by a quote attributed to H.G. Wells at the end of the nineteenth century: "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write" (quoted in Castles, 1992, p.v). Since educators advocate the use of pedagogies that teach through social questions that are of interest and concern to middle-years students (Earl, 2000), the development of statistical literacy to assist understanding and critical thinking should be an important aspect of middle years' classrooms (Watson, 1998a).

Research into students' understanding in the area of statistical literacy has only recently emerged as a focus out of the research into students' understanding of the probability and statistics curriculum introduced in many countries around 1990 (e.g., Australian Education Council, 1991; National Council of Teachers of Mathematics (NCTM), 1989). An appreciation of the need to consider context together with curriculum content, as well as the implications beyond the mathematics classroom (e.g., Watson, 1995; 2000), has led to specific consideration of some curriculum topics in this way. Sampling, for example, is a crucial aspect of most statistical investigation, but is often ignored as an important mathematical idea. Using items that are also included in this current study, Watson and Moritz (2000) showed a developmental progression in understanding of sampling with respect to the appreciation of the importance of context and to the ability to identify bias in claims from the media. The contexts used in these items were the purchase of a car based on advice from a friend or from a consumers' report, a media report about school students in the United States based on a sample taken in Chicago, and a phone-in survey on legalisation of marijuana conducted by a teenage radio station.

Although there has been research into the development of statistical understanding in the school context, and much of this has occurred in the middle school grades, little consideration has been given to the effects of teaching through context and practical applications, or whether different aspects of statistical literacy impact differentially on males' and females' learning. This study aimed to explore some issues that may be related to these aspects through a further consideration of data obtained from studies of the development of statistical understanding.

Background

The data for this study came from several surveys undertaken during the period 1993 to 2000, addressing aspects of students' understanding of the chance and data curriculum, reasoning with respect to statistical information presented in newspapers, and

comprehension of statistical variation (Watson, 1994; Watson, Kelly, Callingham & Shaughnessy, in press). Students responded to open-ended questions and these responses were analysed using hierarchical cognitive developmental models (e.g., Biggs & Collis, 1982; 1991) and Watson's (1997) three-tiered structure of statistical understanding (Watson, 1997). These analyses provided qualitative numerical codings based on hierarchies of response that reflected hypothesized developmental levels. With respect to the Biggs and Collis model, for example, three levels occurred in cycles representing i) unistructural functioning based on single elements of the problem; ii) multistructural functioning, combining elements in a sequential fashion; and iii) relational functioning, based on integration of elements for a complete response within the cycle.

The initial interest of the current project was to map all of these responses onto a single hierarchical scale, using Rasch techniques based on common item equating (Rasch, 1980; Griffin, Callingham, Smith & Kays, 1998). Using the kind of hierarchical response sequence applied to these items combined with Rasch modelling techniques is becoming widely used to measure higher order thinking (e.g., Wilson, 1990; 1992; Griffin, 2000; 2001).

Some of the items were contextually based, using real newspaper articles and situations embedded in a "real-life" context, as illustrated earlier with the sampling examples, whereas others were curriculum based, using situations that were typical of classroom activities or textbook questions, such as rolling dice or reading graphs (e.g., Watson, Collis & Moritz, 1997). The nature of the items thus mirrored approaches to teaching recommended for middle school teachers, being based on activities, such as using spinners or dice, or using real and relevant contexts, such as media items about drugs and guns. The manner of response was varied, some items requiring a numerical approach, whereas others demanded written explanations. Findings from PISA (Programme for International Student Assessment) (Organisation for Economic Co-operation and Development (OECD), 2001) suggested that in Australia boys tended to do less well on questions demanding extended reading, as was required by some of the statistical literacy items. No attempts had been made in previous research, however, to consider item bias or gender effects on the items in this study. Hence, there was interest in seeing whether the scale obtained from the equating process was biased towards female students, particularly since both the nature of many items and the manner of response demanded literacy as much as mathematics skills.

One way of addressing potential bias is to consider Differential Item Functioning (DIF). DIF is said to occur when subgroups of students or test takers who have the same ability perform differently on particular items. In this situation, performance on each item is associated with some characteristic or property that is not related to achievement (Bolt & Stout, 1996). This may or may not have educational significance. In this study the characteristic of interest was gender.

Du (1995) proposes three criteria to consider in relation to DIF. The first is statistical significance of any DIF observed. This, Du suggests, is important but not sufficient on its own to require adjustment to the test. Second, is any DIF real or simply a statistical artifact? Replicability across different groupings of the sample is one way of testing this: DIF exhibited when the groupings are changed, randomly or systematically, is likely to be real rather than accidental. Finally, is there a substantive interpretation of the DIF where it is detected? By considering items that demonstrate DIF it may be possible to provide an explanation that affects the interpretation of the test results (Grimby, Andrén & Daving, 1998; Grimby, 1999).

Item Response Theory (IRT) provides a means of detecting the presence of DIF. If the probabilities of response to an item cannot be explained wholly by the ability of the student and the fixed difficulty parameters of the item, the item is considered to exhibit DIF (Wu,

Adams & Wilson, 1998). In this study, the items were mainly polytomous, having a number of hierarchical response categories. The theoretical model used for the initial analyses was the Partial Credit Model (PCM) (Masters, 1982). This model uses the interaction between test-takers and items to produce estimates of person ability and item difficulty on a single measurement scale. The generalized form of the model is

$$\frac{\pi_{ix}}{\pi_{i(x-1)} + \pi_{ix}} = \frac{\exp(\beta - \delta_{ix})}{1 + \exp(\beta - \delta_{ix})}$$

where

π_{ix} is the probability of a person responding in category x (x = 1, 2, ... m) of item i;

b is the person's ability in the domain being measured by this set of items; and

d_{ix} is the difficulty of the step threshold that governs the probability of the response occurring in category x rather than category x - 1.

This model is particularly suited to the type of items used in the earlier studies because it does not require the same step structure for every item. By using an estimate of "step difficulty" within each item in the assessment, the PCM locates a person on the underlying variable through a consideration of the number of steps that the person has made beyond the lowest level of performance. The points at which the likelihood of a higher-level response becomes greater than that of a lower level response are called thresholds.

When items are polytomous, that is they are scored using a series of steps, differences between subgroups, such as males and females, in the probabilities of achieving each threshold, that is scoring in any one category, may indicate DIF (Bolt & Stout, 1996). Under these conditions, the two-way interaction between persons and items becomes multi-faceted; that is, there may be interactions between gender and item, and among gender, item and step. To estimate these interactions, and thus determine the presence or otherwise of DIF, requires a multi-faceted IRT model. General comparative indices can be obtained from Quest software (Adams & Khoo, 1996) using the COMPARE command. This allows for comparison of two different sub-groups, in this case males and females, against the item parameter estimates. In order to model the more complex interactions between gender, item and step levels, Conquest software was used (Wu, Adams & Wilson, 1998). This program allows the estimation of multi-faceted models through the application of a generalised Rasch model. In a simple case of dichotomous items, for example, the main effects of interest would be gender and item and the gender*item interaction. In this situation the model has three terms - gender, item and gender*item, and two facets - gender and item. The software constructs every possible combination of gender and item to create a number of generalised items. Probabilities of response to each of these generalised items is then estimated using gender, item and gender*item as the main effects. The fit to this model provides indication of DIF. This process can be extended to polytomous items, as in this study (Wu, Adams & Wilson, 1998).

Method

The data were taken from the responses of 2811 students across grades 5 to 9 in Tasmanian schools to a series of surveys about statistical understanding undertaken from 1993 to 2000 (Watson, 1994; Watson et al., 2001). The sample is summarized in Table 1.

Table 1: Student sample by grade and gender

Grade	Male	Female	Total
5	224	197	421
6	427	454	881
7	140	98	239
8	108	99	207
9	545	519	1065
Total	1444	1367	2813

The surveys all had items in common that were used in pre-test/post-test models of evaluation, although no students had undertaken all items. The subset of data used in this study came only from students in grades 5 to 9 who were responding to items for the first time. No post-test data were used. A total of 80 items was included in the data set.

Of the 80 items, 44 had been used in a study of students' understanding of variation (Watson et al., 2001). These covered typical curriculum topics, such as spinner outcomes, and more contextually based items, such as sampling a school population. Forty of the items were used in surveys undertaken in 1993, 1995 and 1997. These included curriculum content items, such as aspects of basic probability (Watson, Collis & Moritz, 1997), of conditional and conjunction probabilities (Watson & Moritz, 2002a), of average (Watson & Moritz, 1999), and some that presented information in tables (Watson, 1998b). Ten items, the media items, were based on genuine newspaper articles. These addressed applied statistical understanding in context, such as median house prices (Watson & Moritz, 1999), sampling (Watson & Moritz, 2000), and aspects of graphing (Moritz & Watson, 1997; Watson, 2000). Four items were common across all administrations of the surveys. A summary of the items used is provided in Appendix A.

The data had all been previously coded using developmentally based models (Biggs & Collis, 1982; 1991), and the codings stored electronically. Rasch modelling techniques were used to prepare an anchor file from the four items that were common to all grades and all surveys (Adams & Khoo, 1996; Griffin, 1997). These four items provided 14 data points in all for linking. All items were then calibrated and equated onto one scale using the Quest v2.1 computer program (Adams & Khoo, 1996), anchored to the common item set (Griffin, 1997; Griffin et al., 1998). Through a consideration of fit to the model, it was established that the items were all measuring the same construct, although their qualitative interpretation covered different aspects of statistical literacy. From the full scale analysis, a second anchor file was written out for all items. This was used to anchor subsequent subscale analyses.

Two subscales were constructed. "Curriculum" consisted of 49 items that ranged from worded explanations of terms, such as sample, to graph reading questions. These items were considered to be typical of classroom activities, such as tossing dice, or textbook questions. "Context" consisted of 31 items based mainly on the media survey, although it also included questions based on real situations such as fish population estimates and sampling in a school environment. Case ability and fit, and item fit statistics were established

for each of these scales, anchored to the full item set. These established that each scale was measuring a single construct in a consistent way.

Of interest here, however, was how the items behaved with respect to gender. This was achieved using the Quest COMPARE command that allowed comparison between the responses of males and females. Initial estimation suggested that there was a difference in the ways in which some items behaved with respect to gender. In order to explore this further, the grouping variable was changed to gender by grade and comparisons of item functioning were undertaken for males and females in each of the grades from 5 to 9. The results were displayed in the form of a map and as a table showing the difficulty level of each item for male and female students in each grade, together with a chi-squared value of significance (Adams & Khoo, 1996).

To explore the nature of the DIF seen in these analyses, Conquest was used to provide estimates of DIF using multi-faceted models. Since many of the items used in this survey were polytomous, an additional interaction was included of step, or threshold level, as well as the item and gender terms. Models that were invariant with respect to gender - item*step - and models that included gender in the interaction - gender*item*step - were used to establish the nature of any DIF identified.

Results

Overall ability and fit to the model

To establish whether any DIF detected had come about because of differences in measured ability or fit to the model, estimated abilities and fit statistics were obtained for the overall scale of all items, and the Context and Curriculum subscales. Figure 1 shows results of overall estimated ability (logits) by gender and grade for the overall scale using all 80 items.

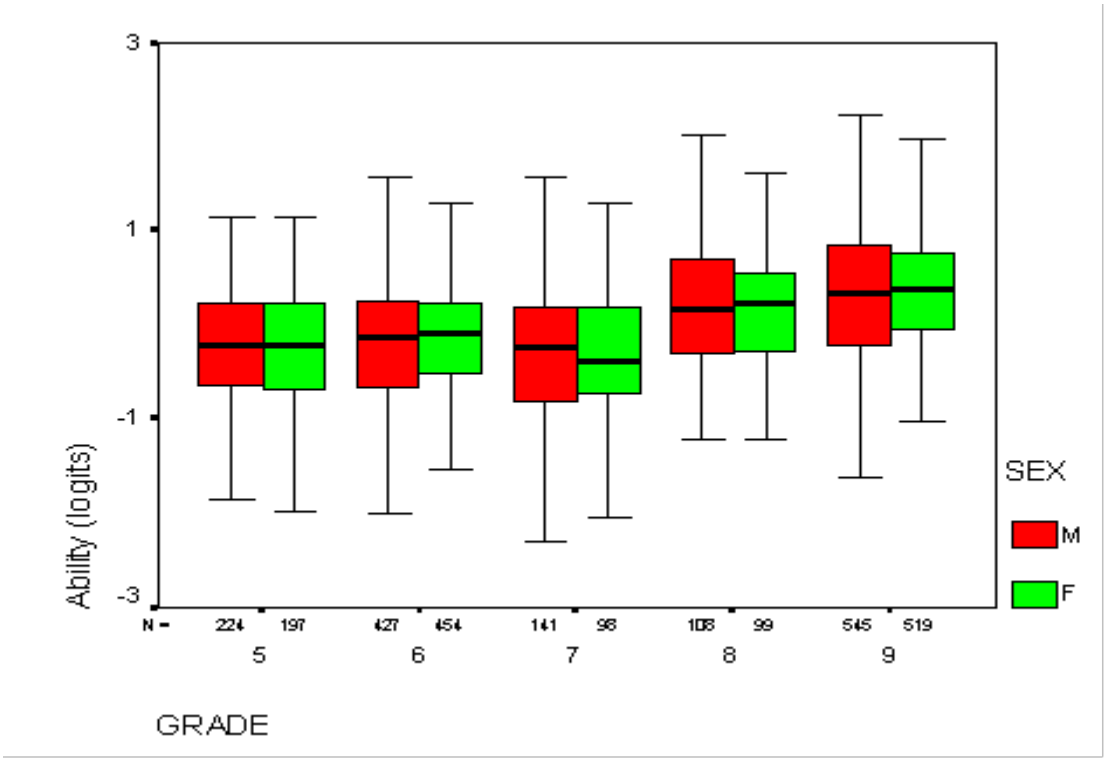

Figure 1: Estimated ability on all items by grade and sex
(anchored to common items)

There appeared to be a slight drop in performance in grade 7, and there was considerable overlap among the grades. Between male and female students, however, there was little visible difference other than a slightly greater spread of ability in male students. Similar patterns were seen for the two anchored subscales.

To explore the gender aspects further, case estimates for the overall scale, and for the two subscales, were written into SPSS files. Summary statistics were calculated and an independent samples t-test was undertaken to test the difference between males and females. This was repeated for the two subscales, curriculum and context. Summary results are shown in Table 2. A critical t-value was applied as a basis on which to retain or reject the null hypothesis that the means were equal. On this basis, it is not possible to reject the null hypothesis for any scale.

Table 2: Summary statistics and independent samples t-test (estimated ability in logits)

Scale	n M	Mean M	SD M	n F	Mean F	SD F	Mean Difference	t	df
Overall	1444	0.023	0.85	1367	0.066	0.74	-0.043	-1.42	2809
Curriculum	1420	0.036	0.88	1343	0.071	0.79	-0.034	-1.068	2761
Context	1389	0.061	0.99	1336	0.064	0.91	-0.003	-0.086	2723

In order to establish the effect of gender on ability measures on each scale, effect sizes were calculated (Effect size = Mean female - Mean male / S_p where S_p is the pooled standard deviation). For each scale these were small (<0.05), suggesting that gender had little effect on achievement on any of the three scales.

Fit to the model was determined using the commonly accepted values of item infit (weighted mean square) lying within the range 0.77 to 1.30 (Adams & Khoo, 1996; Keeves & Alagumalai, 1999). Model fit was acceptable for all scales, as shown in Table 3.

Table 3: Fit statistics for all scales

Scale	Infit	Infit t
Overall	1.01	0.34
Curriculum	0.99	0.48
Context	0.90	-2.28

These findings suggest that all scales fitted the model and that the achievement of males and females on all scales seemed to be the same.

Differential Item Functioning

Having established that the responses from the two subgroups, male and female, appeared to be behaving the same way with respect to the model, Du's criteria were applied to establish whether the initial DIF detected was genuine. Using the Quest COMPARE command (Adams & Khoo, 1996) and a grouping variable of grade and gender, separate

comparison of items was undertaken for males and females in each grade against the Context and Curriculum scales. Results were obtained as a DIFFMAP, showing the different behaviour of each item with respect to males and females, and which also produced an indication of significance at the 0.05 level. The map for grade 9 students against the Context scale is shown in Figure 2 as an example. Items showing statistical significance at the 0.05 level are shown in bold. The stars for these items lie outside the dashed vertical lines that indicate the significance level. Items with stars to the right of the zero position indicate items that are easier for year 9 females, and those with stars to the left indicate items that are easier for year 9 males.

Plot of Standardised Differences

Easier for year 9 males Easier for year 9 females

-6 -5 -4 -3 -2 -1 0 1 2 3 4

-------+------+------+-----+------+------+------+------+-----+------+-----+

CH11 . *| .

Q17 . * | .

SM19 . | * .

Q20 . * | .

M1CH . * | .

M2PI * . | .

M3OD * . | .

M4DR . * | .

M5AV . * | .

M7CH . * | .

M8GR * . | .

M8QU . | * .

M9C * . | .

M9D * . | .

M10A . | .

M10B . * | .

M6AB . | * .

M6C . | * .

M6D . | *

MVE1 . | *

MVE2 . | * .

MVE3 . | . *

MVE4 . | . *

MVE7 . | *

MVE5 . | . *

MVE6 . | . *

BT1A . * | .

BT1B . | * .

DRG1 . * | .

MVE8 . * .

Figure 2: DIFFMAP showing comparison of item estimates for groups year 9 males and year 9 females on the Context scale

A table of results was also obtained, showing standardised differences and a chi-square statistic for each item, and for each subscale. The overall results provided a measure of DIF for each subscale, and can be considered as a measure of test bias. Summary results for each of the subscales, Context and Curriculum, against grade are shown in Table 3. The varying degrees of freedom shown in each grade reflect the number of items presented to each grade level. For each analysis, the chi square value is sufficiently large as to suggest that the null hypothesis of equal means should be rejected.

Table 3: Overall DIF between male and female across grades

Grade	Mean Male	Mean Female	ChiSq	df
CONTEXT SCALE				
5	-0.62	0.04	17.86	7
6	0.00	0.00	83.25	14
7	0.00	0.00	30.94	13
8	0.01	0.17	57.12	16
9	0.00	0.00	160.89	29
CURRICULUM SCALE				

5	0.27	0.20	55.21	33
6	-0.09	0.02	91.59	15
7	0.10	0.26	129.76	34
8	0.33	0.08	32.33	14
9	0.03	0.04	106.21	42

To further explore the nature of the DIF, models allowing for different interactions were fitted to the data using the Conquest software (Wu, Adams & Wilson, 1998). The main effects modelled were item and gender. The output provided estimates of mean item difficulty and mean student ability by gender. These were then combined to provide an estimate allowing for item*gender. In order to determine the model that best fitted the data, estimates of a model that considered the step levels invariant with respect to gender - item*step, were compared with a model where the step levels varied with respect to gender - gender*item*step. These analyses were undertaken for both Context and Curriculum scales. Summary results are shown in Table 4.

Table 4: Summary table for polytomous DIF analysis

Scale	Model	Number estimated parameters	Deviance
Curriculum	Item*step	91	48808.132
	Gender*item*step	126	48357.391
Context	Item*step	103	57602.336
	Gender*item*step	146	57450.489

For both scales the more complex model that included gender in the item*step interaction, showed lower deviance with a larger number of estimated parameters. Thus the model that included gender with item and step fitted the data better than the model that was invariant with respect to gender (Wu, Adams & Wilson, 1998). This suggests that although there are overall differences in the behaviour of items with respect to gender, these are particularly evident in the threshold structures of the items for males and females.

Further examination of the threshold differences is needed to explain this finding and this leads to Du's third criterion for interpreting DIF - the substantive nature of any DIF exhibited and the effects this has on construing the test results. Both subscales contained items that required numerical answers or calculations as well as items that demanded interpretation and written explanation of this. A content and skills analysis of these items was undertaken to determine how the DIF detected could be interpreted.

Items showing significant DIF

Table 5 provides a summary of those items in each scale that showed significant DIF. Of the 49 items included in the Curriculum scale, 16 showed DIF significant at the 95% confidence level. Of these, 10 were easier for males. In contrast, of the 31 items included in the Context scale, nine showed DIF at the 95% confidence level but only three were easier for males. This suggests that the overall DIF shown for each of the two scales would favour males in the Curriculum scale and females in the Context scale.

Table 5: Summary of items showing significant DIF on each scale

Scale	Easier for males	Easier for females
Curriculum	ME13, CF15, CP16, Q10D, DIE2, SP1, SP2B, SP4A, SP5B, TBL5	SMP4, HAT8, SP3B, TRV5, TWN3, SP9
Context	M9C, M9D, M10A	M8QU, M6AB, M6D, MVE4, MVE6

Of the 13 items in total shown in Table 5 that were easier for males, nine required either a numerical response or a series of numbers. In contrast, none of the items shown that were easier for females required this kind of response. Rather these items relied on interpretation of a situation, and usually some kind of written explanation. The single item of this nature that was easier for males was TBL5, which asked for fair methods of choosing four students to lead a parade, and in which the highest level response expected chance methods. Lower levels of response took into account various behavioural aspects, such as the best performer in each sport. Figure 3 shows the estimated difficulties of each threshold for male and female students on TBL5.

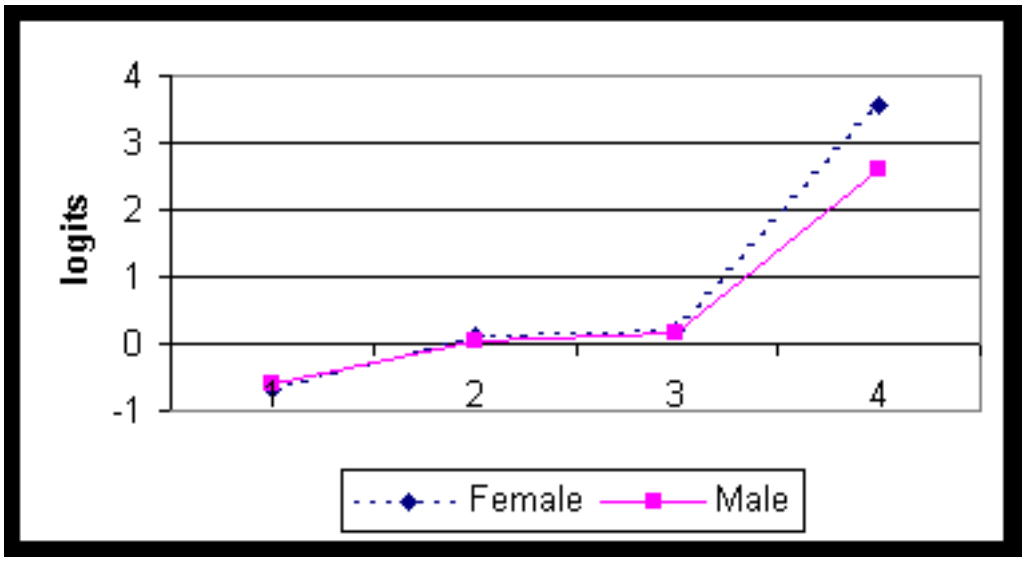

Figure 3: Threshold difficulties for males and females on TBL5

A consideration of the threshold levels on TBL5 for males and females reveals some interesting differences. The estimated difficulties of the four threshold levels for males and females suggest that there is little difference between them with the exception of level 4. This is borne out when the jump in difficulty between the thresholds is considered. There is very little difference between the steps needed to reach the next threshold except for the step from level 3 to 4. This shows an increase in difficulty of 3.35 logits for girls as opposed to 2.45 logits for boys, suggesting that it is more difficult for girls to reach the highest level of response, although the steps along the way suggest that it is no more difficult for girls than boys to reach a level 3 response. These findings are summarised in Table 6.

Table 6: Threshold levels for males and females on TBL5

Threshold Level	1	2	3	4
Est. difficulty F	-0.66	0.10	0.21	3.56
Est. difficulty M	-0.61	0.05	0.16	2.61
Threshold difference		1 to 2	2 to 3	3 to 4
F		0.76	0.11	3.35
M		0.66	0.11	2.45

In contrast, Figure 4 shows the results for MVE6 on the Context scale, which was easier for females. This item required written critical analysis of subtle aspects of survey methodology. The pattern of response on this item is somewhat different. The gap between males and females appears fairly constant across the thresholds. This suggests that, unlike TBL5, this item was more difficult for males at each threshold, but that the jump from one threshold to the next was very similar for boys and girls.

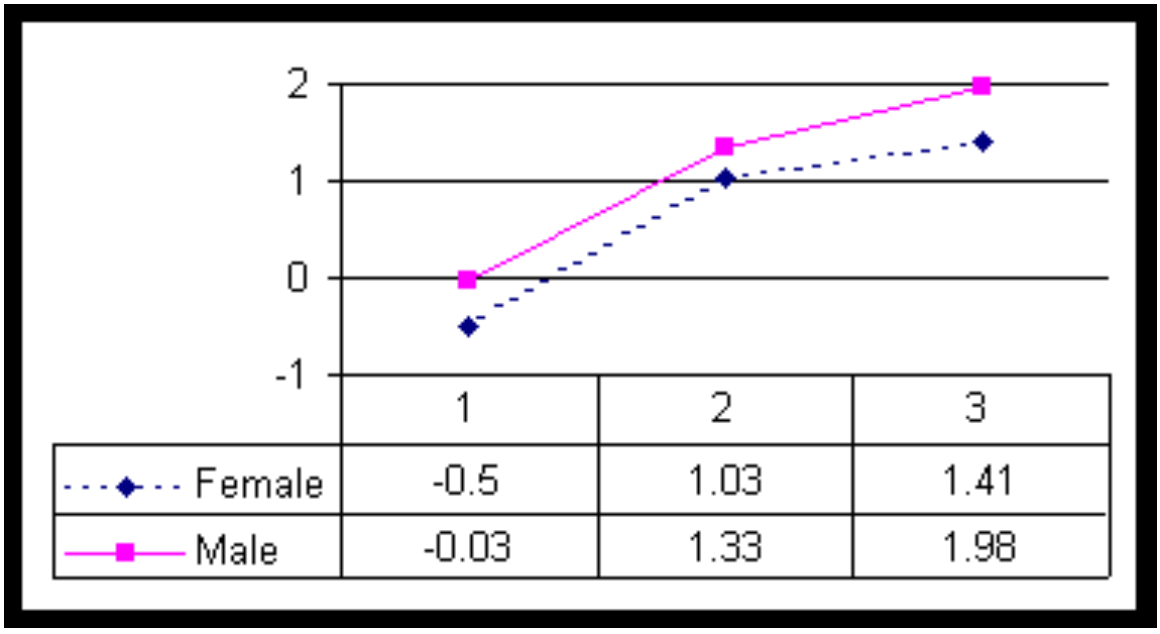

Figure 4: Threshold difficulties for males and females on MVE6

Items that showed DIF in favour of females followed similar patterns, whereas the items that showed DIF in favour of males had very similar male and female threshold levels apart from the jump to the highest level of response, similar to the pattern shown in Figure 3. Examples of the kinds of items that showed DIF for males and females are provided in Appendix B.

One explanation could be the overall difficulty of the items - items having a lower average difficulty could be easier for females. Figure 5 shows fifteen items that showed significant DIF ordered from left to right by their average difficulty. Lines indicating the average difficulty estimates of these items obtained from the whole data set, and estimates of the average difficulty levels for male and female students separately on these items are included for illustration.

Data points appearing below the overall average line show items that are easier for the particular group, and those above the average difficulty line show items that are at a higher difficulty for the group. The vertical line shows the point where the lines for male and female students cross the average difficulty line. This is the point where the items become, on average, easier for females.

The differences between male and female students appear, on average, greater for easier items than more difficult ones. In general boys seem to find items at a lower average difficulty easier than girls do but this is not a consistent pattern across all items; MVE4 and M6AB, for example, are more difficult for boys. The overall difficulty levels of the items cannot apparently explain the patterns of differences observed between males and females.

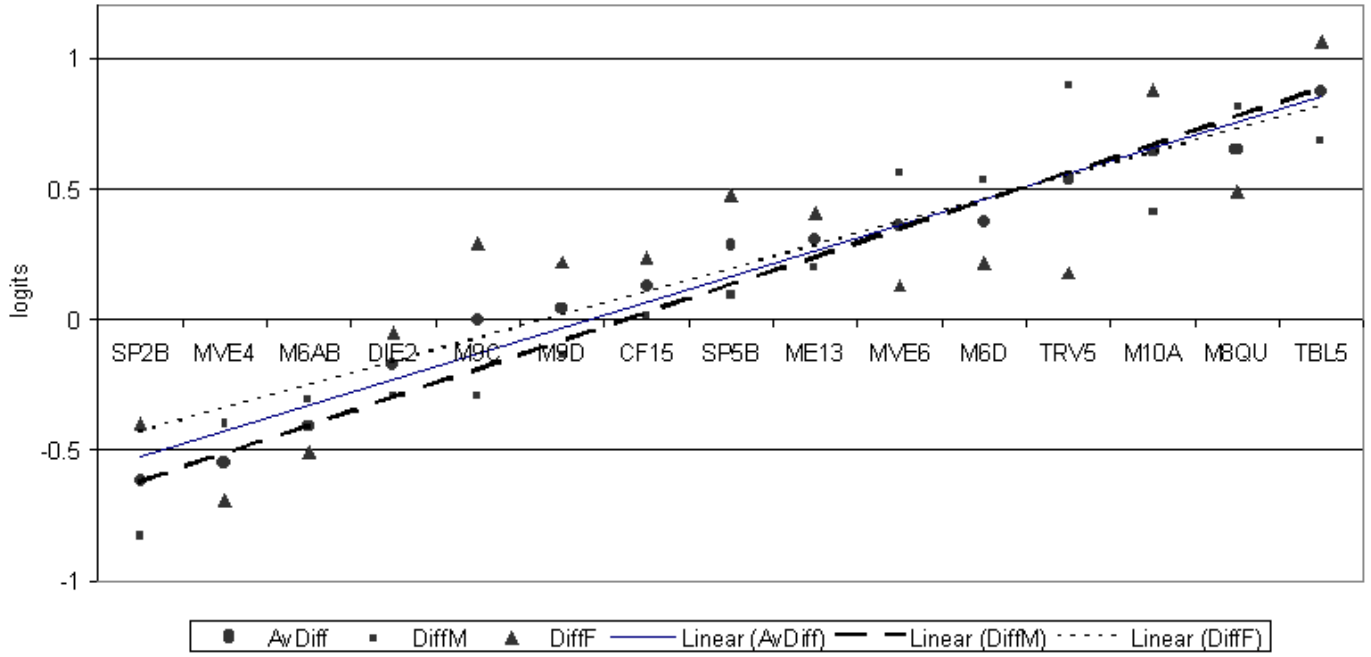

Figure 5: Average item difficulties for 15 items showing DIF

Discussion

Using Rasch measurement techniques, 80 previously archived items were combined to form a single scale of statistical literacy. These showed good fit to the model, indicating that they all measured the same construct, and exhibited no apparent difference in ability between male and female students. Subscales of items that matched the chance and data curriculum, the Curriculum scale, and items which were context based, the Context scale, showed similar properties. These subscales, however, exhibited DIF at significant levels between males and females across all grades. Further analysis using multi-faceted Rasch models showed test level DIF for models that took account of item*gender interactions and those that included threshold levels, gender*item*step. These latter models better fitted the data, suggesting that the differences occurred within the structure of the threshold levels. Consideration of items that showed significant DIF revealed differences between the subscales, and in the nature of the items. Boys appeared to find items that required numerical responses easier than those requiring written explanations or extended reading. Examination of the threshold levels in these items suggested that in general differences between the threshold difficulties were very similar for males and females, except the step to the highest level. In contrast, for items that females found easier the differences were apparent at each threshold, and the jump between thresholds was generally very similar for both groups. Overall, boys tended to find the lower difficulty items easier, in contrast to the girls where the opposite was true. The DIF observed was significant, replicable and explicable, meeting the three criteria suggested for testing DIF (Du, 1995).

Does this suggest that the items themselves, or the Curriculum and Context scales, should be modified to eliminate DIF? The original purpose of these items was as a research tool to explore children's understanding of chance and data ideas. In this situation the DIF can be safely ignored, since it is cognitive functioning rather than achievement that is under consideration. Both scales measured the same construct for males and females, and effect sizes for gender on each scale were small. The significant DIF shown at test level, however, suggests that the subscales, Curriculum and Context, appear to be measuring the construct in subtly different ways for boys and girls in this sample.

At issue is the educational significance of these findings: what this means for teaching. The items were chosen particularly because they mirrored approaches to teaching advocated for the middle years of schooling. The clear differences shown in the nature of the items that boys or girls found easier suggest that there are some implications for the classroom.

Both curriculum content and contextual applications need to be addressed across all grades. In both situations attention should be paid to providing numerical responses as well as written explanations. By providing a range of approaches, both in the stimulus material and in the nature of the response expected, any bias in favour of boys or girls should be reduced. Further, it seems that specific intervention may be needed to encourage girls to address calculations and numerical responses, and boys to produce quality written responses.

The different patterns of response in items that males found easier than females, and vice versa, also need consideration. Items that males found easier showed very similar threshold structures for males and females at the lower thresholds, but a difference at the step to the highest response level. This may be an artifact of the item construction, and the coding of the responses. Coding of the responses was based on the complexity of the responses provided, exemplified by Biggs and Collis's cognitive developmental model (Biggs & Collis, 1982; 1991). At the highest level, students justified their answers by calculation or by providing numerical support for their argument, demonstrating understanding that connected their mathematical knowledge with the situation presented in the item. The numerical

argument was often concise, requiring little writing. This kind of numerical justification was less expected in the contextually based items - rather many of these required complex written responses that drew on understanding of the situation presented. The implications for teaching and assessment remain, however: boys need to be encouraged to write clear explanations whereas girls need to use mathematical language and text effectively. It is somewhat disappointing that after years of awareness of gender differences with respect to mathematics that the differences observed here are still being seen.

In the middle years of schooling there is also increasing emphasis on teaching through an integrated curriculum by generalist teachers (Hill & Russell, 2000). These results have some implications also for teacher training and professional development. Many teachers lack confidence in teaching mathematics, and particularly so in the area of chance and data (Callingham, Watson, Collis, & Moritz, 1995). If teachers are to make connections between relevant questions of interest to middle school students and the mathematical knowledge needed to underpin the development of statistical literacy, well-developed and extensive programs of professional development will be needed (Watson & Callingham, 2001).

The results presented here suggest that gender issues are not yet fully resolved, but have become more subtle. A statistically literate populace needs connected skills in both mathematics and literacy - the ability to express and interpret ideas drawing on underpinning mathematics and writing skills. Unless both aspects, mathematics and literacy, are addressed explicitly it is likely that boys and girls will both be disadvantaged, but in different ways.

Acknowledgement

This research was funded by the Institutional Research Grant Scheme at the University of Tasmania. At the time of writing Rosemary Callingham was employed by the University of Tasmania.

References

Adams, R.J., & Khoo, S. (1996). *Quest: The interactive test analysis system, Version 2.1*. Melbourne: ACER.

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

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). Multi-modal learning and the quality of intelligent behaviour. In H. Rowe (Ed.) *Intelligence: Reconceptualisation and measurement.* Hillsdale, NJ: Lawrence Erlbaum.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika 23*, 67-95.

Callingham, R.A., Watson, J.M., Collis, K.F. & Moritz, J.B. (1995). Teacher attitudes towards chance and data. In B. Atweh & S. Flavel (Eds.) *Galtha.*(Proceedings of 18th Annual Conference of Mathematics Education Research Group of Australasia. pp 143-150). Mathematics Education Research Group of Australasia: Darwin.

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

Department of Education, Tasmania (2002). E*ssential learnings*. *Framework 1.* Hobart: Author.

Du, Y. (1995). When to adjust for differential item functioning. *Rasch Measurement Transactions, 9*(1),414. Retrieved 20 November 2001 fromhttp://www.rasch.org/rmt/rmt91.htm

Earl, L.M. (2000). *Reinventing education in the middle years*. Paper presented at Middle Years of Schooling Conference: Collaborating For Success.Melbourne, Victoria. August, 2000.

Education Queensland, (2000). *The new basics project technical paper*. Retrieved 11 January 2002 from http://education.qld.gov.au/corporate/newbasics/html/library.html

Eyers, V., Cormack, P., & Barratt, R. (1993). *The education of young adolescents in South Australian government schools: Report of the Junior Secondary Review. Summary*. Adelaide: Education Department of South Australia.

Freebody, P., Ludwig, C., & Gunn, S. (1995). *Everyday literacy practices in and out of schools in low socio-economic urban communities.* Griffith University: DEETYA.

Gal, I. (1995). Big picture: What does numeracy mean? *GED Items, 12,* 4/5.

Griffin, P. (1997). *An introduction to the Rasch model*. Melbourne: Assessment Research Centre.

Griffin, P. (2000). *Competency based assessment of higher order competencies.* Keynote address delivered at the NSW ACEA State Conference, Mudgee, April 28 2000. Retrieved 13 September 2001 from http://www.edfac.unimelb.edu.au/ARC/recentpubs.html

Griffin, P. (2001). *Performance assessment of higher order thinking.* Paper presented at the annual conference of the American Education Research Association, April 10th, Seattle. Retrieved 13 September 2001 from http://www.edfac.unimelb.edu.au/ARC/recentpubs.html

Griffin, P., Callingham, R.A., Smith, A., & Kays, M. (1998, December). *A twenty year equating study of mathematics achievement*. Paper presented at the Australian Association for Research in Education Conference, Adelaide.

Grimby G., Andrén E., Daving Y., & Wright B.D. (1998). Dependence and perceived difficulty in daily activities in stroke survivors. *Stroke* 29, 1843-1849.

Grimby, G. (1999). Useful reporting of DIF. *Rasch Measurement Transactions, 12*(3), 651. Retrieved 1 January 2002 from http://www.rasch.org/rmt/rmt123d.htm

Hill, P.W., & Russell, V.J. (2000). *Systematic, whole-school reform of the middle years of schooling.* Retrieved 18 April 2001 from www.sofweb.vic.edu.au/mys/pdf/phill.pdf

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.

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 49*, 359-381.

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

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

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

Moritz, J.B. & Watson, J.M. (1997). Graphs: Communication lines to students? In F. Biddulph & K. Carr (Eds.), *People in mathematics education, Vol. 2*. Waikato: MERGA, pp. 344-351.

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

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

National Council on Education and the Disciplines (NCED). (2001). *Mathematics and democracy: The case for quantitative literacy.* Princeton, NJ: Woodrow Wilson Foundation. Retrieved 14 September 2001 from http://www.woodrow.org/mellon/nced/mathematics_democracy.html

Organisation for Economic Co-operation and Development (OECD). (2001). *PISA in brief from Australia's perspective*. Melbourne: ACER.

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

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

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

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

Watson, J. M., Collis, K. F., & Moritz, J. B. (1994). Assessing statistical understanding in Grades 3, 6 and 9 using a short answer questionnaire. In G. Bell, B. Wright, N. Leeson, & G. Geake (Eds.), *Challenges in Mathematics Education: Constraints on*

Construction (Proceedings of the 17th Annual Conference of the Mathematics Education Research Group of Australasia, pp. 675-682). Lismore, NSW: MERGA.

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

Watson, J.M. & Callingham, R. (2001, May). *Preparing teachers for a middle school mathematics Classroom: Creating Connections*. Paper presented at the Middle Years of Schooling Conference, Brisbane.

Watson, J.M. & Callingham, R. (2002). *Statistical literacy: A complex developmental construct*. Manuscript submitted for publication.

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

Watson, J.M. (1995). Statistical literacy: A link between mathematics and society. In A. Richards, G. Gillman, K. Milton, & J. Oliver (Eds.), *Flair: Forging links and integrating resources* (pp. 12-28). Adelaide, SA: Australian Association of Mathematics Teachers.

Watson, J.M. (1997). Assessing statistical thinking using the media. In I. Gal & J.B. Garfield (Eds.). *The assessment challenge in statistics education*. Amsterdam: IOS Press.

Watson, J.M. (1998a). Statistical literacy: What's the chance? *Reflections, 23*(1), 6-14.

Watson, J.M. (1998b). Numeracy benchmarks for years 3 and 5: What about chance and data? In C. Kanes, M. Goos, & E. Warren (Eds.), *Teaching mathematics in new times. Volume 2* (pp. 669-676). Brisbane: Mathematics Education Research Group of Australasia.

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

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

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

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

Watson, J.M., & Moritz, J.B. (2002b). *The development of comprehension of chance language in context: evaluation and interpretation*. Manuscript submitted for publication.

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

Wilson, M. (1990). Investigation of structured problem-solving items. In G. Kulm (Ed.) *Assessing Higher Order Thinking in Mathematics*. 187-203. Washington, DC: American Association for the Advancement of Science.

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

Wu, M.J., Adams, R.J. & Wilson, M.R. (1998). *ACER Conquest. Generalised item response modelling software*. Melbourne: ACER Press.

Appendix A

Item names, descriptions, coding values, content, and source (in reference list).

Name	Description of context	Code	Content	Source
AVG2	Meaning of average	0-3	Average	W&M 1999
SMP4	Meaning of sample	0-3	Sampling	W&M 2000
DIE7	1 or 6 more likely outcome	0-4	Chance	WC&M 1997
HAT8	Draw names from hat	0-4	Chance	WC&M 1997
BOX9	Marbles red:blue, 60:40, 6:4	0-3	Chance	WC&M 1997
CH11	Medicine - 15% chance of rash	0-2	Chance	WC&M 1994
AV12	2.2 children/family	0-2	Average	W&M 1999
ME13	Median of science data	0-3	Average	W&M 1999
CP14	Left-handed men, conditional	0-1	Chance	W&M 2002a
CF15	Colds and school, conjunction	0-1	Chance	W&M 2002a
CP16	Female school teacher, conditional	0-1	Chance	W&M 2002a
Q17	Fish-tagging	0-1	Chance	WC&M 1994
CF18	Male heart attacks, conjunction	0-1	Chance	W&M 2002a
SM19	Toyota/Honda, data for purchase	0-3	Sampling	W&M 2000
Q20	Actors' performance, regression to mean	0-1	Inference	WC&M 1994

M1CH	Chance newspaper headlines	0-2	Chance	W&M 2002b
M2PI	128.5% pie chart	0-2	Graph/table	W 1997
M3OD	7:2 odds, North:South game	0-4	Chance	MW&C 1996
M4DR	Phone-in marijuana survey	0-2	Sampling	W&M 2000
M5AV	Median house price	0-3	Average	W&M 1999
M7CH	Chicago/US, non-representative sample	0-4	Sampling	W&M 2000
M8GR	Graph heart deaths, car usage claim	0-3	Graph/table	W 2000
M8QU	Question heart deaths, car usage claim	0-2	Inference	W 2000
M9C	Call rates next 10 mins, picto-bar graph	0-2	Graph/table	M&W 1997
M9D	Call rates first 30 mins, picto-bar graph	0-2	Graph/table	M&W 1997
M10A	Cricket 4 tails in 4 tosses	0-4	Chance	M&W 2000
M10B	Cricket choice for next toss	0-2	Chance	M&W 2000
RAN3	Meaning of random	0-3	Chance	MWP-M 1996
Q10A	Sports table: number of girls tennis	0-2	Graph/table	W 1998b
Q10B	Sports table: number of boys netball	0-2	Graph/table	W 1998b
Q10C	Sports table: children choose swimming	0-2	Graph/table	W 1998b
Q10D	Sports table: evenly divided sport	0-2	Graph/table	W 1998b
Q10E	Sports table: more girls or boys?	0-4	Graph/table	W 1998b
M6AB	Wrinkles/smoking conditional statements	0-2	Chance	W 1998c
M6C	Wrinkles/smoking conditional statement	0-1	Chance	W 1998c
M6D	Wrinkles/smoking conditional	0-2	Chance	W 1998c

	statement			
DIE2	Number times 1 to 6 in 60 tosses of die	0-4	Variation	WKC&S in press
SP1	50:50 spinner chance	0-2	Chance	WKC&S in press
SP2A	50:50 spinner 10 spins outcome	0-3	Variation	WKC&S in press
SP2B	50:50 spinner 50 spins outcome	0-3	Variation	WKC&S in press
SP3A	50:50 spinner 10 spins again, same?	0-3	Variation	WKC&S in press
SP3B	50:50 spinner 50 spins again, same?	0-3	Variation	WKC&S in press
SP4A	Spinner surprise /10?	0-1	Variation	WKC&S in press
SP4B	Spinner surprise /50?	0-1	Variation	WKC&S in press
SP5A	6 sets of 10 spins	0-2	Variation	WKC&S in press
SP5B	6 sets of 50 spins	0-2	Variation	WKC&S in press
SP6	Spinner outcomes graph - lowest value	0-1	Graph/table	WKC&S in press
SP7	Spinner outcomes graph - highest value	0-1	Graph/table	WKC&S in press
SP8	Spinner outcomes graph - range	0-1	Graph/table	WKC&S in press
SP9	Spinner outcomes graph - mode	0-1	Graph/table	WKC&S in press
SP10	Spinner outcomes graph - describe shape	0-1	Graph/table	WKC&S in press
SP11	Which graphs are made up?	0-2	Variation	WKC&S in press

TRV1	Pictograph: how many children walk?	0-1	Graph/table	WKC&S in press
TRV2	Pictograph: how many more bus than car?	0-1	Graph/table	WKC&S in press
TRV3	Pictograph: graph the same everyday?	0-1	Variation	WKC&S in press
TRV4	Pictograph: new student by car, girl/boy?	0-3	Inference	WKC&S in press
TRV5	Pictograph: row with train, explain void	0-2	Inference	WKC&S in press
TRV6	Pictograph: Tom away, how will he travel?	0-5	Inference	WKC&S in press
MVE1	How to survey 600 school children?	0-3	Sampling	WKC&S in press
MVE2	Assess method: all students, choose 60	0-3	Sampling	WKC&S in press
MVE3	Assess method: 10 members computer club	0-3	Sampling	WKC&S in press
MVE4	Assess method: all 100 Grade 1	0-3	Sampling	WKC&S in press
MVE5	Assess method: ask 60 friends	0-3	Sampling	WKC&S in press
MVE6	Assess method: volunteer at tuck shop	0-3	Sampling	WKC&S in press
MVE7	Best survey method?	0-2	Sampling	WKC&S in press
MVE8	Predicted % students buying a ticket	0-3	Inference	WKC&S in press
SMP3	Meaning of sample; give example	0-3	Sampling	WKC&S in press
TBL1	Table: how many girls chose tennis?	0-1	Graph/table	WKC&S in press
TBL2	Table: most popular sport for girls?	0-1	Graph/table	WKC&S in

				press
TBL3	Table: most popular sport for boys?	0-2	Graph/table	WKC&S in press
TBL4	Table: how many children?	0-1	Graph/table	WKC&S in press
TBL5	Parade - fair choice of 4 leaders	0-4	Sampling	WKC&S in press
TWN1	Stacked dot plot, no scale, families	0-3	Graph/table	WKC&S in press
TWN2	Stacked dot plot, scale, families	0-3	Graph/table	WKC&S in press
TWN3	Stacked dot plots: which tells story better?	0-3	Inference	WKC&S in press
BT1A	Bar graph, boat deaths: unusual features	0-2	Graph/table	WKC&S in press
BT1B	Bar graph, boat deaths: variation coding	0-3	Variation	WKC&S in press
VAR	Meaning of variation	0-3	Variation	WKC&S in press
AVG1	Average value of science data	0-3	Average	WKC&S in press
DRG1	Marijuana survey - sample size	0-1	Sampling	WKC&S in press

W (Watson) and M (Moritz) are the first authors of these references in the reference list.

Appendix B

Examples of items that males or females found easier.

DIE2 Easier for males

Imagine you threw the die 60 times. Fill in the table below to show how many times each number might come up.

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

Why do you think these numbers are reasonable?

TBL5 Easier for males

A primary school had a sports day where every child could chose 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

1) How many girls chose Tennis?

2) What was the most popular sport for girls?

3) What was the most popular sport for boys?

4) How many children were at the sports day?

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

NB Only the last of these five questions (TBL5) showed significant DIF.

MVE6 Easier for girls

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

Claire set up a booth outside of the tuck shop. Anyone who wanted to stop and fill out a survey could. She stopped collecting surveys when she got 60 kids to complete them. What do you think of Claire's survey?

c GOOD c BAD c NOT SURE

Why?

TWN3 Easier for females

A class of students recorded the number of years their families had lived in their town.

Here are two graphs that students drew to tell the story.

Graph 1

			X												
			X						X						
		X	X						X			X			
X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	
I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
0	1	2	3	4	5	6	10	11	12	13	14	17	25	37	

YEARS IN TOWN

1. What can you tell by looking at Graph 1? [Space provided]

Graph 2

			X																																		
			X									X																									
		X	X									X					X																				
X	X	X	X	X	X	X				X	X	X	X	X			X								X												X
I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
0					5					10					15					20					25					30					35		

YEARS IN TOWN

2. What can you tell by looking at Graph 2? [Space provided]

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

NB Only the last of these three questions, TWN3, showed significant DIF.