How Unusual is the Gender Specificity of Mathematical Test Item Types Reported for Dutch Primary School Students?

Lorraine Davis
Fintona Girls School
<ldavis@fintona.vic.edu.au>

David Clarke
University of Melbourne
<d.clarke@unimelb.edu.au>

Marja van den Heuvel-Panhuizen
Freudenthal Institute – University of Utrecht
< m.vandenheuvel@fi.uu.nl>

Gender differences in performance in Mathematics have been reported for grade 6 Dutch students both in overall performance, with boys outperforming girls, and for particular item types. Similar gender differences have not been reported for Australian students. Findings reported in this paper suggest that there are types of items which function differently for girls and boys in Australia as well as in the Netherlands.

Gender Specificity of Mathematical Item Types in the Netherlands

The impetus for this paper comes from van den Heuvel-Panhuizen's analyses of the performance of Dutch students in mathematical tests held by the National Institute of educational Measurement (CITO) at the end of primary schools in 1993, 1994, and 1995. In those tests and international testing, the performance of Dutch boys of a primary school age has been significantly better than that of their female counterparts. In fact, the analysis of results for the Third International Mathematics and Science Study (TIMSS) showed that this overall difference in favour of boys was the second greatest of all participating countries for Population 1, that is, the primary students taking part in the study (Mullis, Martin, Fierros, Goldberg & Stemmler, 2000).

In order to investigate the nature of the gender differences found in the national testing, further analysis was undertaken by van den Heuvel-Panhuizen (1996; 1997), the results of which are summarised as follows:

- Girls do not score lower than boys in all mathematical domains
- Test items have gender-specific characteristics

As a consequence of her research, van den Heuvel-Panhuizen (1997) found that she could clearly distinguish between what she described as "boys" problems and "girls" problems. Her definition of a "boys" problem was one which was answered correctly more often by boys than by girls, whilst "girls" problems were items for which boys and girls were correct approximately equally successfully or on which the girls did slightly better.

The most interesting findings from van den Heuvel-Panhuizen's study concerned the identification of particular regularities in the types of problems that were gender-specific. Most importantly, the characteristics of gender-specific problems were not related to the gender of a problem's protagonist nor (in every case except one) to the context in which the problem was figuratively situated (Clarke & Helme, 1998). The issues of gendered protagonists and gendered contexts are territory that has already been thoroughly worked over by Leder and her co-workers, amongst others (Leder, 1992). Van den Heuvel-Panhuizen's findings relate to the type of mathematics required by a task.

Her categorisation of "boys" problems consisted of:

- problems which ask for daily-life knowledge on numbers and measures,
- problems in which large numbers with many zeros are used,
- problems in which different numbers or different units of measurement are used,
- problems which have possibilities for "tinkering" with numbers, and
- problems which ask for reasoning backwards.

Her categorisation of "girls" problems consisted of:

- problems which ask for accuracy,
- problems for which the text is complex,
- problems which ask for (reflection on) strategies and not for calculations.
- well-known problems which refer to standard procedures,
- straight-forward problems, and
- problems which refer to shopping situations.

(van den Heuvel-Panhuizen, 1997, p. 70)

Mathematics learning and teaching in the Netherlands, especially at primary school level, is characterized by an approach described as Realistic Mathematics Education (RME). The philosophical basis for this approach is that "mathematics is a human activity and focuses on meaningful applications" (van den Heuvel-Panhuizen, 1996, p. 14). Ainley (1997), in describing the Australian situation, included at primary school level many characteristics highly similar to the approach advocated in the Netherlands. These include less emphasis on algorithms and a greater use of problem solving, modelling and investigative tasks, as well as showing a commitment to the relevance of mathematics to every-day life.

Although there may be similar approaches to teaching in both countries, studies involving Australian primary school students have not shown any significant difference in the standard of the performance of boys and girls in mathematical tests overall, and any difference on particular item-types has appeared to be minimal (Barnes, 1997; Queensland School Curriculum Council 1998; van Wyke 1999). This is true of the results of large-scale testing on a state-wide basis as well as from the results of Australian primary-aged students in TIMSS (Lokan, Ford & Greenwood, 1996).

The purpose of this paper is to report on a study designed to ascertain whether the performance of a relatively small group of year 6 Australian students on types of mathematical test items showed the same gender specificity as was reported by van den Heuvel-Panhuizen for similar aged students in the Netherlands. Gender specificity is defined as the measurable existence of statistically significant differences in responses to particular types of mathematical test items on the basis of gender.

The Test Items and Their Validation

The test used in the study consisted of nineteen multiple-choice items. The multiple-choice form was chosen to replicate the questions used in the Netherlands as accurately as possible. It is important to note that the range of items was not intended to represent the entire year 6 mathematics curriculum from either an Australian or Dutch perspective. Rather, the items were chosen to match the characteristics of items identified by van den Heuvel-Panhuizen as "extreme" (as previously described). As only two of the multiple-choice items used in the original testing of grade 6 students in the Netherlands were freely

available, most test items had to be obtained from other sources. Fourteen of the nineteen items had been previously used in TIMSS, or were adapted from TIMSS items. It is worth noting that most of these fourteen items were selected from those used for students in years 7 and 8 in TIMSS testing. This was to minimize any possible the ceiling effect should the items prove too simple for the Australian students in the study.

Of the other items, two had been identified as "extreme" items when they were used in the Netherlands and three were items written for the test in an attempt to have items which matched the characteristics of each of the categories described by van den Heuvel-Panhuizen as "extreme". Nine of the items were designed to be "boys" items and nine were designed to be "girls" items. The nineteenth item was a TIMSS measurement item. This item was included because the relative performances of boys and girls on it from both the nine-year old and the thirteen-year old groups from Australia were readily available. Thus it provided a means of comparing the performance of students in the study with the performance of a larger, more representative group of Australian students. Further, the item type was one that van den Heuvel-Panhuizen's results identified as gender-specific (in this case, a "boys" item).

In order to confirm that the items chosen actually matched the descriptions of the characteristics as identified by van den Heuvel-Panhuizen, the items were sent, without including the researcher's categorisation, to van den Heuvel Panhuizen for her "blind" comment. Of the nineteen items, there were only three for which there was any disagreement as to their gender specificity. For example, her description of the item intended to be a "tinkering" item and thus "boy-friendly", was "girls item, precise algorithmic calculation; boys only if they realized 25×99 must end in 75, in that case tinkering"[1]. This illustrates the possibility that an item could be classified differently depending on the researcher's ability to anticipate the way students would attempt to solve that problem.

Results

One hundred and seventy-one students, 85 boys and 86 girls, from 5 schools participated in the study. The manner in which their answers were analysed replicated as far as possible the approach taken in the Netherlands.

The Individual Student Level

The percentage of items correct was calculated for each student and then the average and standard deviations of these percentages were calculated for the boys and girls separately. Table 1 shows these results.

Although the relative performance of boys and girls was not a major focus of the study, it is of note that the two distributions have a similar central tendency and spread. Although there was a small difference in the mean scores, in favour of the boys, this was not statistically significant (t=-0.49, p>0.05). From this it can be concluded that for the test items taken as a group, there was no meaningful difference in the performance of girls and boys. It must be remembered that the selection of items for the Australian study did not constitute a complete assessment of student performance on the entire syllabus. This is

[1] van den Heuvel-Panhuizen, M. <m.vandenheuvel@fi.uu.nl> (2004). 24 March 2004. Re: An easy favour - I hope [Email to: David Clarke <d.clarke@unimelb.edu.au>]

different from the Dutch situation, where gender specificity was identified for item types embedded in whole-curriculum assessment. This comparison of performance on the 19 items used in the Australian study, therefore, demonstrated that the boys and girls in the Australian study were similarly successful on the item set. Any gender specificity of a particular item type is therefore all the more significant.

Table 1
Comparison of Boys' and Girls' Scores in the Test as a Whole

Gender	Total number of students	Mean number correct out of nineteen	Standard deviation	Mean percentage correct
Girls	86	10.7	3.3	55.1
Boys	85	10.9	3.2	57.4

Calibrating against TIMSS

Forty-one percent of the students in the study answered the TIMSS item included for calibration purposes correctly (the 19th item, see above), whilst when it was used in TIMSS, 23% of the grade 4 and 42% of the grade 7 Australian students were correct (Schmidt et al, 1991). As the sample of students in the current study was from grade 6, it would be expected that the percentage of students answering this question correctly would lie between the percentages for grade 4 and 7 students. The fact that this happened and that this percentage was closer to the percentage correct for grade 7 than for grade 4, suggests that there is some argument for the representativeness of the student sample in this study in relation to the national grade 6 population.

Analysis of Responses to Individual Items and Item Types

For this paper, the criteria for classifying items were similar to those used by van den Heuvel-Panhuizen. Thus "girls" items were defined as ones answered equally well by girls and boys, as well as ones on which girls outperformed boys. For purposes of empirical classification, "boys" items were defined as ones answered correctly by at least 5% more boys than girls. Table 2 shows the success rates for males and females on each of the 19 items.

Of the sixteen items where the classification of the gender specificity by the researcher agreed with van den Heuvel-Panhuizen's, five of the ten "boys" items and seven of the nine "girls" items, showed empirically the same gender specificity that might have been expected from the pre-test classification. However, the only item for which the difference was statistically significant was item 17, a "boys" one, that had been used in the Dutch testing and involved different units of measurement. This is not surprising, given the comparatively small Australian sample. The significance of the identified gender specificity of particular item-types derives from the empirical confirmation of a predicted pattern of response.

Table 2
The Success Rates for Males and Females on each of the 19 items

Item number	Intended Gender Specificity	Intended Characteristics	Boys (% correct)	Girls (% correct)	Overall Success Rate (% correct)
1	G	Well known problem using standard procedure	77.7	76.7	77.2
2	G	Straight forward	87.1	88.4	87.7
3	G	Accuracy	74.1	80.2	77.2
4	G	Complex text	89.4	90.7	90.2
5	B	Every day problem	56.5	54.7	55.6
6	B	Large numbers, estimation	52.9	51.2	52.1
7	G	Shopping	64.7	72.1	68.4
8	G	Reflection on strategies, not calculation	90.6	96.5	93.6
9	B	Daily life knowledge of measures	52.9	41.9	47.4
10	B	Different units	31.8	25.6	28.7
11	B	Reasoning backwards	32.9	31.4	32.2
12	B	Different units	51.8	41.9	46.8
13	B	Working backwards	28.2	23.3	25.7
14	G	Standard problem, use of algorithm	45.9	38.4	42.1
15	G	Complex text	69.4	64.0	66.7
16	G	Shopping	50.6	57.0	53.8
17	B	Different units of measurement	47.1	18.6	32.8
18	B	Tinkering	47.1	52.3	49.7
19	B	Measurement	40.0	41.9	40.9

Similarities to Dutch Gendered Performance

The relative performance of boys and girls in this study showed marked similarities to the performance of Dutch students for many of the item types found to have gender-specific properties in the Netherlands. This was evident for four of the item types considered to be "girls" ones. These were described as problems set in shopping situations, problems requiring accuracy, straight–forward problems and those requiring standard procedures. There was one item that had been initially considered to be a "standard problem, using an algorithm" and thus a "girls" item, which was answered correctly by a larger proportion of boys than girls. This item involved a calculation using fractions and it suggested that not all students may have been taught this skill and thus many may not have used an algorithmic approach. This confirms the sensitivity of the categorization to curricular variation and student history identified earlier in this paper. Although fewer of the item types identified in the Netherlands as "boy-friendly" showed a gender specificity in favour of boys in this study, the performance of the students on the types of items involving different units of measurement and large numbers showed the similar male-oriented gender specificity to that reported for Dutch students.

In summary, the following gender-specific item characteristics, first identified in the Dutch analysis, recurred in this analysis of Australian students' responses:

Boys Problems

- problems in which large numbers with many zeros are used,
- problems in which different numbers or different units of measurement are used,

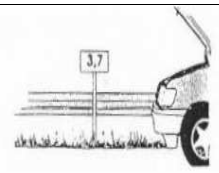
Girls Problems

- problems which ask for accuracy,
- well-known problems which refer to standard procedures,
- straight-forward problems, and
- problems which refer to shopping situations.

The significance of these similarities must be given serious consideration.

Questions Identified as extreme in the Netherlands

The most significant results in the entire study come from the performance of students on the two items shown in Figure 1 that had also been used in testing in the Netherlands.



The telephone to call for help is at the 3.4 km mark. How far is that from the 3.7 km mark?

A. 0.3 m B. 3 m
C. 30 m D. 300 m.

"Boys" item

Susan would like to buy this camera. She saved $40.75 in January, $39.20 in February, $75.15 in March and $80.95 in April. Her father is paying the rest. How much does he have to pay?

A. $173.45 B. $173.55
C. $233.55 D. $273.45

"Girls" item

Figure 1. Items used in testing in the Netherlands.

In this study 6.4% more girls than boys, compared with 4% in the Netherlands, answered the "girl-friendly" item correctly and 28.4% more boys than girls, compared with

26% in the Netherlands, answered the "boy-friendly" item correctly. Considering that the overall achievement of the Australian girls and boys in the current study was equivalent, in contrast to the gendered overall results from studies in the Netherlands, the similarity of the gender difference in performance on these two items, in direction and magnitude, in the two countries is surprising. It is the magnitude of the difference for both items that differs markedly from results reported from other studies involving Australian students of a similar age. This suggests that for Australian students, there are types of test items that have the potential to prompt markedly different responses from boys and girls.

The "girls" item, classified as a shopping one, showed the second largest difference in favour of the girls. Identification of the attributes that contribute to it being a "girls" item seems to be important in the understanding of gender specificity of items of this type. While the context is a "shopping" one, unless it is solved by a "smart" method, in order to obtain the correct answer, learnt procedures must be applied accurately with attention to detail. Yet the "boys" item shown also needs careful attention to detail, but in a different context. The response that was selected by the greatest number of students, who answered this item incorrectly (alternative A), ignored the detail that required students to change the units of measurement. As a higher percentage of girls made this error than boys, one can ask why do girls get this detail incorrect and yet are demonstrably better than boys at getting other details of an algorithmic nature correct.

Differences between Australian and Dutch Gendered Performance

In discussing differences in the performance on the basis of gender between the students in the study and students tested in the Netherlands, conjectures can be made concerning the contributions of the educational settings of both groups of students. In the study, Australian girls performed as well as Australian boys on some items that have been proved to be "boy-friendly" in the Netherlands. They also performed considerably better than boys on many of the standard, straight-forward questions, and those requiring accurate calculations, thus differing from Dutch girls, whose performance was comparable but not significantly superior to boys' performance. Reasons for these results can be sought from the educational contexts of the two groups of students, but must at this stage remain speculative (see below).

Concluding Remarks

This study provides some insight into the appropriateness of van den Heuvel-Panhuizen's classification, when used in a different, non-Dutch context. Certainly, in this study, the classification was appropriate for some questions on which girls tended to outperform boys, especially those requiring a straight-forward or algorithmic approach. However, the situation was not as clear-cut for other questions. The results of this study suggest that some questions, which on the surface seem to have the characteristics of 'boys' items, may, in some situations, because of students' learning experiences, be questions that can be solved by using learned procedures and thus answered better by girls than boys. If our goal is to develop in all students comparable expertise in solving all types of problems, then the question becomes, "What classroom experiences are most likely to promote this equity in capability?"

The implementation of RME approaches in Dutch primary schools, with its emphasis on applications to real life situations, may give Dutch boys an advantage in answering these

types of questions, while a lack of emphasis on the learning of set procedures may deny Dutch girls the opportunity to succeed in a way similar to that of students (particularly female students) from other educational contexts. The responses of the Australian students in this study suggest that girls' performance on a particular problem type can be improved if that problem type can be introduced and rehearsed in the classroom to the extent that its solution becomes a routine mathematical performance. It is less obvious how to develop in boys the proficiency in accurate, procedural performance that typifies girls' responses.

The importance of this study lies in the unexpected difference in performance of Australian boys and girls on particular item types, which had marked similarities with gender specificity of item types reported in the Netherlands. This suggests that the potential for gender differences in mathematical performance is still a concern for students in Australia and that from a wider international perspective, the gender differences in mathematical achievements reported for Dutch primary school students may not be unusual, but may reflect gender-specific tendencies to engage in particular types of mathematical thinking.

References

Ainley, J. G. (1997). Australia. In D. F. Robitaille (Ed.), *National contexts for mathematics and science education: An encyclopedia of the education systems participating in TIMSS* (pp. 39-49). Vancouver, Canada: Pacific Educational Press.

Barnes, M. (1997). Classroom views of gender differences. In B. Doig & J. Lokan (Eds.), *Learning from children: Mathematics from a classroom perspective* (pp. 41-61). Melbourne: ACER.

Clarke, D.J., & Helme, S. (1998). Context as construction. In O. Bjorkqvist (Ed.), *Mathematics Teaching from a constructivist point of view* (pp. 129-147). Vasa, Finland: Faculty of Education, Abo Akademi University.

Leder, G. (1992). Mathematics and gender: Changing perspectives. In D.A. Grouws (Ed.), *Handbook of research in mathematics education* (pp. 597-622). New York: Macmillan.

Lokan, J., Ford, P, & Greenwood, L. (1996). *Maths and science on the line: Australian junior secondary students' performance on the third international mathematics and science study* (TIMMS Australia Monograph No. 1) Melbourne: Australian Council for Educational Research.

Mullis, I., Martin, M., Fierros, E., Goldberg, A., & Stemmler, S. (2000). *Gender differences in achievement.* Chestnut Hill, MA: TIMSS International Study Center.

Queensland School Curriculum Council (1998). *Statewide performance of students in aspects of literacy and numeracy in Queensland 1995, 1996 and 1996.* Brisbane: Author.

Schmidt, W.H., McKnight, C. C., Cogan, L. S., Jakwerth, P. M., & Houang, R. T. (Eds.). (1999). *Facing the consequences : Using TIMSS for a closer look at U.S. mathematics and science education.* Dordrech; Boston: Kluwer.

van den Heuvel-Panhuizen, M. (1996). *Assessment and realistic mathematics education.* Utrecht, The Netherlands: CD- ß Press/Freudenthal Institute.

van den Heuvel-Panhuizen, M. (1997). How equally suited is realistic mathematics education for boys and girls? – A first exploration. In E. Pehkonen, (Ed.), *Proceedings of the 21st Conference of the International Group for the Psychology of Mathematics Education* (Vol. 3, pp. 65-72). PME.

van Wyke, J. (1999). Student achievement in mathematics in Western Australian government schools: Monitoring Standards in Education '98. Perth: Education Department of Western Australia.