PREDICTIONS OF GENDER DIFFERENCES IN PERFORMANCES OF YEARS 5 AND 6 CHILDREN ON PENCIL-AND-PAPER MATHEMATICS ITEMS

FRED BISHOP
The University of Newcastle

M. A. (KEN) CLEMENTS
The University of Newcastle

Sixteen items were chosen from tests used in a primary school mathematics competition taken, in 1992 and 1993, by over 25000 students in Years 5 and 6. For 5 of the items, males did better than females; for 6, the proportions of males and females who gave correct answers were virtually identical; and for 5, females did better than males. The 16 items were randomly sequenced and shown to primary teachers, trainee primary teachers, trainee secondary teachers, and mathematics education researchers, who were asked to select items on which (a) girls did noticeably better than boys; (b) boys did noticeably better than girls; or (c) girls and boys performed equally well. Analyses showed that (a) female respondents were more likely to give correct responses than male respondents; and (b) the mathematics education researchers and the practising primary teachers were more likely to give correct responses than trainee teachers.

Gender Differences in Performances on Pencil-and-Paper Mathematics Competitions

Gender-related Differences on the Hunter Region Primary School Mathematics Competition

The present paper will make use of data generated by the Hunter Region Primary Mathematics Competition *(HRPMC)*. This Competition has been operating annually since 1981, and in 1992 and 1993 there were 12352 and 13067 competitors, respectively, from over 220 State, Catholic and Independent schools (all of which are located in the Hunter Region of New South Wales). Since its beginnings, one of the authors (Bishop) has been director of *HRPMC*, and in this capacity he has received large support from mathematicians, mathematics educators, mathematics teachers, school systems, business and industry.

In both 1992 and 1993 about half of the *HRPMC* competitors were in Year 5, and half in Year 6. At both year levels, about half were girls and half boys. For both years, results were totally determined by performance on timed (45-minute) pencil-and-paper tests, each comprising 35 items. On the papers, items were grouped into three sections, with those in the first section tending to be the easiest, and those in the third section the hardest. Students at both Year levels took the same test

and, as would be expected, Year 6 students tended to obtain higher scores than Year 5 students. However, equal number of awards were given to Year 5 as to Year 6 competitors. It would be fair to say that the competitors probably comprised students whom teachers regarded as average or above average in mathematics.

For both 1992 and 1993, just over twice as many boys as girls won *HRPMC* awards, at both the Years 5 and 6 levels. This same pattern, of boys outperforming girls at the extreme top end of the distribution has, in fact, occurred consistently since the *HRPMC* began.

The Main Research Question

The research described in this paper was stimulated by a comment made by Edwards (1984) who, on looking closely at statistics derived from the 1983 Australian Mathematics Competition (AMC) test, noted that overall, boys responded correctly to more questions than did girls, the difference being one to one-and-a-half questions per 30-question paper. Edwards (1984) then stated:

> On looking more closely at the statistics it is found that this difference is not spread evenly over all the questions. On just a handful of questions boys are far more successful than girls—and the questions involved are certainly not identifiable in advance. (p. 11)

This statement is a direct challenge to all who construct pencil-and-paper tests of mathematics. Is it possible for test constructors to predict whether any particular item will be such that an educationally significant higher proportion of boys than girls (or vice-versa) will give the correct answer?

The main issue addressed in this paper is concerned with a different but related question—are teachers and other educators able to predict, reasonably accurately, whether items on the pencil-and-paper, multiple-choice *HRPMC* tests are likely to generate gender-related differences in performance? Edwards' question raised the issue of whether *test constructors* could predict differences. It is a matter of interest whether persons engaged in the day-by-day activities of school mathematics can accurately predict the kinds of pencil-and-paper questions likely to be associated with gender-related differences in mathematical performance.

Having posed these questions, we would wish to emphasise, at the outset, that we do not believe those who develop mathematics competition instruments should have their *primary* focus on gender issues. Their major concern must be the mathematics. All members of committees responsible for developing AMC tests and *HRPMC* test do not believe that they should deliberately exclude items—for example, items involving time/distance/speed relationships—for which it is likely that gender-related differences will occur.

Related Literature from the Australian Mathematics Competition

Since Edwards' (1984) paper, there have been a number of papers that have shed light on AMC items which have been associated with gender-related differences. Annice, Atkins, Pollard and Taylor (1990) divided the questions which appeared on the AMC tests for the period 1983–1987 into eight mutually exclusive categories:

1. Basic Manipulations—Arithmetic
2. Basic Manipulations—Algebra
3. Basic Manipulations—Geometry
4. Routine Problems—Arithmetic
5. Routine Problems—Algebra
6. Routine Problems—Geometry
7. Problem Solving
8. Non-routine Problem Solving

Annice et al's (1990) analysis indicated that although the secondary school boys taking the AMC tests did slightly better than girls overall, (a) the trend was not uniform, (b) results seemed to be affected by a confidence factor associated with guessing on multiple-choice items—with boys being more willing to guess, and (c) the most marked difference in favour of boys occurred with items involving time-speed-distance relationships. Edwards (1984, 1985) also noted the discrepancy, in favour of boys, for items involving time-speed-distance relationships.

Boys have continued to do better on AMC tests than girls. Atkins, Taylor, Leder and Pollard (1994), in a recent analysis of AMC data for the period 1983–1992, showed that boys had performed slightly better than girls on all the categories defined by Annice et al. (1990). However gender-related differences in performance had narrowed in six of the eight categories, the exceptions being *Basic Algebra* and *Routine Algebra.* Also, differences between girls and boys in Years 7–10 were less—but had remained fairly constant at the Year 12 level.

The ratio of male to female AMC prizewinners has remained over 2:1 throughout the period of operation of the Competition. This is still the case in the 1990s (see Pederson, 1992, 1993), despite the fact that just over half of the 400 000 students (approximately) who enter the AMC each year are female.

The 1993 AMC tests were such that students in Years 7 and 8 were asked to answer the same sets of items. However, summary results which have been presented separately for Years 7 and 8 students (O'Halloran, 1993), and these reveal that, despite Forgasz's (1994) suggestion to the contrary, if Year 7 boys did better than Year 7 girls on an item, then this pattern was also likely to occur for Year 8 students. Similarly, if Year 7 girls did better than Year 7 boys on an item, then this pattern was likely to be repeated for Year 8 students. And, if there were no significant gender differences in mean performance on an item at Year 7, then the same was likely to be true at Year 8. The same stability of gender performance relativity was evident for 1993 AMC items which were common on the Years 9 and 10 tests, and for items which were common on the Years 11 and 12 tests.

The Present Study

Rather than provide a post-hoc classification of the types of questions on which gender-related differences occurred for the 1992 and 1992 *HRPMC* tests—which could be done at any time— it was decided to use the *HRPMC* data to investigate whether educators who did not have access to the *HRPMC* data for 1992 and 1993 could identify items on the tests for which large gender-related differences had occurred?

Method

Developing the Research Instrument

Each item on the 1992 and 1993 *HRPMC* tests was of the pencil-and-paper multiple-choice variety. The proportions of boys and girls giving correct answers for each item were considered, and a z-score, that could be used to test whether the difference between the proportions was statistically significant, calculated (see Ferguson, 1971, pp. 160–162, for the relevant formula).

Sixteen items were selected according to the following criteria:

1. Five items with the highest z-score, *and for which boys did better than girls,* were chosen.
2. Five items with the highest z-score, *and for which girls did better than boys,* were chosen.
3. Six items for which the absolute value of the z-score was almost zero (and therefore, items *for which boys and girls performed at about the same level)* were chosen.

Items in Category 1 will, for the remainder of this paper be denoted as "B>G" items. Items in the other two categories will be denoted "G>B" and "G=B" items, respectively. The B>G items had z-values ranging from 7.5 to 14.0 and the G>B items had z-values ranging from -5.3 to -8.3.

The 16 items were randomly sequenced and shown to (a) 26 primary teachers (in the Hunter Region), (b) 81 trainee primary teachers (at the University of Newcastle), (c) 23 trainee secondary teachers (also at the University of Newcastle), and (d) 25 mathematics education researchers (MERGA members, who responded to a written request—altogether 35 MERGA members were invited to respond). These "respondents" were told that almost identical numbers of boys and girls entered the *HRPMC*, and they were asked to indicate, for each item, whether they thought (a) girls' performance on the item would be noticeably better than that of boys; (b) boys' performance on the item would be noticeably better than that of girls; or (c) girls and boys would have performed at approximately the same level.

The respondents were *not* told that five of the items were such that G>B, five were such that B>G, and six were such that B=G.

Illustration .Three of the 16 items are shown below, and readers of this paper are invited to place ticks in what they consider to be the appropriate boxes beside the items (one tick per item).

	Girls Better	Girls = Boys	Boys Better
Question 5: If a 3/4 hour test started at 1:58 p.m. then it finished at A. 1:13 p.m. B. 2:03 p.m. C. 2:33 p.m. D. 2:43 p.m.			
Question 10: What unit of measure would be used to measure the length of the diagonal of the cover page of this examination booklet? A. mL B. cm C. ha D. m^2			
Question 12: If 2.8 metres of ribbon is shared equally among 4 girls then each girl would get A. 7 m B. 7 cm C. 70 cm D. 70 mm			

Of the five questions for which G>B, three required only elementary arithmetic calculations (for example, "502 plus 379 minus 497 equals ..."), one asked respondents to calculate the number of days in the first three months of 1992, and the fifth, Question 10 above, asked for the most appropriate unit to measure the length of the diagonal of the cover page of the examination booklet.

Of the five questions for which B>G, one required students to find ❑, if ❑ ÷ 8 = 48, one involved time/distance/speed relationships, two involved non-trivial time calculations (Question 5, above, was one of these), and one was concerned with sharing a piece of ribbon, 2.8 metres long, equally among 4 girls (Question 12 above).

The six items for which G=B included three items concerned with elementary spatial concepts (like, for example, stating the number of faces, edges and vertices on a tetrahedron). Of the other three G=B items, one asked students to state the next two multiples of 10 after 890, one showed a picture of a thermometer registering about 37°C, and asked students to select an appropriate month—for Hunter Region weather; the other involved time/distance/speed relationships.

Scoring. A method of scoring was devised which meant that if respondents placed ticks randomly in the right-hand columns for the 16 items they would be expected to score 0. Also, if someone had a firm but misguided educational ideology that there should be no gender-related differences in the responses, "because girls and boys are equal," then this person should also score 0. (Thus, for instance, a respondent who placed 16 ticks in the G=B column would score 0.)

The following method of scoring responses was devised. Only 10 of the 16 items would be scored—the five G>B items and the five B>G items. For scoring purposes, responses to the G=B items would be disregarded.

So far as the 10 items to be scored:

1. For a G>B item, a tick in the G>B column would score 1; a tick in the B>G column would score -1; a tick in the G=B column would score 0.

2. For a B>G item, a tick in the B>G column would score 1; a tick in the G>B column would score -1; a tick in the G=B column would score 0.

By this method, the maximum possible score would be 10, and the minimum -10. Theoretically, any integer score from -10 to 10 would be possible.

Hypotheses

Since strict sampling procedures were not employed in the selection of respondents, it could be argued that it would be inappropriate to apply inferential statistical techniques in any analyses of the data. However, it could be argued that the 25 MERGA respondents were reasonably representative of Australian mathematics education researchers, the 26 practising primary teachers represented primary teachers in the Hunter Region, and the two groups of trainee teachers represented the views of trainee teachers.

It might be expected that the MERGA and the practising primary teacher respondents would obtain higher mean response scores than the trainee teachers.

Thus, recognising the inadequacy of the sampling, it was decided to test the following null hypotheses:

Hypotheses 1–4: Each of the groups would have a mean response score of zero (four hypotheses, one for each group)

Hypothesis 5: The mean response scores for the four groups would be equal.

Hypothesis 6: The mean scores for the group of female respondents (taken from the 4 main groups) and the group of male respondents (also taken from the 4 main groups) would be equal.

t-tests would be employed for Hypotheses 1–4, and a two-way analysis of variance for Hypotheses 5 and 6. The critical value for statistical significance would be .05, and two-tailed tests would be employed.

Results

Two respondents scored 7, which was the highest score for any of the 155 respondents. The lowest score was -8. In fact, 97% (or 150 of 155) respondents gained scores between (or including) -3 and 4. The mean score for the total group of 155 respondents was 1.10, and standard deviation 2.14. Table 1 presents a summary of means and standard deviations for the four groups, and Table 2 shows the means and standard deviations of scores of the male and female respondents.

Table 1
Means and Standard Deviations of Scores for the Four Groups

Group	Number in Group	Mean	Standard Deviation
MERGA members	25	1.56	2.10
Primary Teachers	26	1.50	2.32
Trainee (Secondary)	23	0.39	2.35
Trainee (Primary)	81	1.03	2.01

Table 2
Means and Standard Deviations of Scores of Female and Male Respondents

Gender of Respondent	Number in Group	Mean	Standard Deviation
Female	107	1.32	2.00
Male	48	0.60	2.38

When 2-tailed t-tests were applied to the data in Table 1, testing the hypotheses that the means were equal to zero, three of the four means were found to be statistically significantly different from zero ($\alpha = .05$), the exception being for the group of trainee secondary teachers.

A two-way analysis of variance was conducted for Score, using "Group" and "Gender" as Factors. This analysis indicated that with $\alpha = .05$, both "Group" and "Gender" had statistically significant influences on Score. There was a difference in the predictive ability of female and male respondents, with females tending to be the better predictors. Also there was a difference in the predictive power of "Groups." By inspecting Table 1 it can be seen that of the four Groups, the mathematics education researchers and the practising primary teachers were the best predictors, and the trainee secondary teachers the worst.

Discussion

Despite the statistically significant results, indicating that mathematics education researchers, practising primary teachers, and trainee primary teachers had a better than random chance of placing ticks in the correct positions, means of 1.56, 1.50, and 1.03 can hardly be regarded as *educationally* significant. The result suggests that despite at least a decade of great interest in gender differences performance and participation in school mathematics, mathematics education researchers and practising teachers have not made much headway in being able to predict whether given pencil-and-paper, multiple-choice items will yield gender-related performance differences.

Also, although it is interesting that females were better predictors than males, it would be difficult to sustain any argument that the mean of 1.32 for the females represented a result which

had much educational significance. Females were the better predictors, but not by much, and both males and females predicted very little better than would have been expected of someone who merely allocated ticks randomly in response to the specified task.

Respondents were allowed space to comment on the criteria they used to place ticks. Although about one-third of the respondents offered comments, it was clear that most had little idea of gender-based differences for categories of questions like those given in Annice et al. (1990) and Atkins et al. (1994). However, some mathematics education researchers were aware that girls are likely to do better than boys on standard elementary computational tasks, and boys are likely to do better than girls on complex measurement tasks. It was often assumed, incorrectly, that any kind of spatial task was likely to favour boys.

The most common incorrect classification was for Question 12 (which is reproduced earlier in this paper). Many respondents thought that because its context was concerned with girls and ribbons, it should favour girls. In fact, of the 70 items on the 1992 and 1993 HRPMC tests, it was second on the list of items most favouring boys.

This present study suggests the need for similar but larger studies, with tighter sampling procedures being adopted. It would be especially interesting if a team of mathematics competition test developers (for, say, the AMC or the *HRPMC*) were asked to predict, *before* the test was actually taken by students,which items would be associated with gender-related differences. Possibly the most important contribution of the present paper has been to develop a promising form of instrumentation and method of scoring for the research which needs to be carried out.

References

Annice, C., Atkins, W. J., Pollard, G. H., & Taylor, P. J. (1990). Gender differences in the Australian Mathematics Competition. *Mathematics Competitions, 3*(1), 34–41.

Atkins, W., Taylor, P., Leder, G., & Pollard, G. (1994). Where are Raelene, Marjorie and Betty now? *Australian Mathematics Teacher, 50*(1), 40–42.

Edwards, J. (1984). Raelene, Marjorie and Betty: Success of girls and boys in the Australian Mathematics Competition. *Australian Mathematics Teacher, 40*(2), 11–13.

Edwards, J. (1985). Boys and girls in the Australian Mathematics Competition. *Mathematics in School, 14*(5), 5–7.

Ferguson, G. A. (1971). *Statistical analysis in psychology and education.* New York: McGraw Hill.

Forgasz, H. (1994). *Society and gender equity in mathematics education.* Geelong: Deakin University.

O'Halloran, P. (Ed.). (1993). *A selection of Australian gender statistics by question.* Canberra: Australian Mathematics Trust.

Pederson, D. G. (Ed.) (1992). *Australian Mathematics Competition: 1992 solutions and statistics.* Canberra: Australian Mathematics Trust.

Pederson, D. G. (Ed.) (1993). *Australian Mathematics Competition: 1993 solutions and statistics.* Canberra: Australian Mathematics Trust.