Can Item Format (Multiple-Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement?

Michal Beller

The Open University of Israel

Naomi Gafni

The National Institute for Testing and Evaluation (NITE)

March 1996

The authors are listed alphabetically. This project was conducted under the auspices of the National Institute for Testing and Evaluation, Jerusalem, Israel and the Open University of Israel. The authors extend their appreciation to the Educational Testing Service for making the data bases of the International Assessment of Educational Progress (IAEP) available for this study. Special thanks are due to ETSers Thomas Jirele and David Freund. The authors would also like to thank Maya Bar-Hillel, Gershon Ben-Shakhar, Yoav Cohen, Chanan Goldschmidt, Sonia Rocas, and Jerry Levinson for their helpful comments.

Abstract

The purpose of this study was to investigate differential performance of boys and girls on open-ended (OE) and multiple-choice (MC) items on the 1988 and 1991 International Assessment of Educational Progress (IAEP) mathematics test. In the 1988 mathematics assesment a representative sample of approximately 1,000 thirteen-year-olds in each of the six participating countries was assessed. In the 1991 mathematics assessment a representative sample of nine- and thirteen-year-olds (approximately 1,650 from each age group) in some twenty participating countries was assessed. Analyses of both assessments yielded results which indicated that boys generally, performed better than girls in mathematics. In the 1988 assessment, gender effects were larger on MC items than on OE items, corresponding to results of earlier studies. However, the 1991 IAEP assessment produced contrary results: gender effects tended to be larger for OE items than for MC items. These inconsistent results challenge the assertion that girls perform relatively better on OE test items, and suggest that item format alone cannot account for gender differences in mathematics performance. Further investigation of the data revealed that the inconsistent patterns of gender effects with regard to item format were related to the difficulty level of the items, regardless of item format. Correlations between item difficulty and item gender effect size were computed for age 13 in the 1988 assessment and for age 9 and age 13 in the 1991 assessment. The correlations obtained were 0.26, 0.47, and 0.53, respectively, suggesting that the more difficult the items, the better boys perform relative to girls.

Can Item Format (Multiple-Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement?

Results from various national and international large-scale assessments of school children indicate that boys perform better than girls in several areas of mathematics (e.g., Beller & Gafni, 1996; Hedges & Nowell, 1995; Hyde, Fenemma & Lamon, 1990; Linn, 1991; Lumis & Stevenson, 1990; Steinkamp, Harnisch, Walberg & Tsai, 1985). One of the most consistent findings has been the generally similar achievements of girls and boys in mathematics in the early grades, and that boys tend to improve their math scores relative to girls as they move through school (Benbow, 1988; Maccoby & Jacklin, 1974, Willingham & Cole, 1997).

One possible explanation for gender differences on objective tests is based on the hipothesis that males perform better than females due to the fact that these tests are heavily based on multiple-choice items. Bennet (1993) noted that a key claim in the constracted-response/muliple-choice debate is that format affects the meaning of test scores because it restricts the nature of the content and the processes that can be measured. Even minor format differences, such as between multiple-choice and completion items, can change the nature of the constract assessed.

Several studies have analyzed scores on multiple-choice and open-ended (constructed-response) tests by gender of the examinees. In a review article Traub and MacRury (1990) discussed three studies (Bolger, 1984; Murphy, 1980, 1982) that reported a relationship between gender differences and item format. In all three studies, despite differences in test content, the performance of females relative to that of males was better on open-ended tests than on multiple-choice tests. Considering that the term "open-ended" encompasses a wide range of formats, from simple fill-in items to complex performance assessments, the consistency of the results across these studies is noteworthy (Mazzeo, Schmitt & Bleistein, 1993).

One possible reason for this format-gender difference is that the superiority exhibited by girls in verbal ability might find greater expression in open-ended questions. Bolger and Kellaghan (1990) examined gender differences in scholastic achievement as a function of test format by comparing the performance of 15-year-old boys and girls in Irish schools on multiple-choice tests and free-response tests (requiring short written answers) in mathematics, Irish language, and English language achievement. The results indicated that boys performed significantly better than girls on multiple-choice tests as compared to free-response tests. This held true for languages as well as for mathematics, but with a larger format-based gender effect for mathematics.

Another reason for this format-gender interaction might be different response strategies and risk-taking tendencies of girls and boys (as reflected by guessing behavior on multiple-choice items). Ben-Shakhar and Sinai (1991) examined gender differences in the tendency to omit items and to guess on multiple-choice tests. They found a consistent pattern of greater omission rates among females than males (i.e., more guessing by males). An attempt to correct raw scores for guessing (using formula scoring) reduced the male advantage, but only to a slight degree. Gafni and Melamed (1994) found that male and female examinees, as well as examinees with different linguistic and cultural backgrounds, differed in their tendency to guess. There was a small but significant interaction effect of gender and the language in which the examination was taken, on the tendency to guess on unreached items, indicating interaction of gender and culture on response strategies. It was concluded that although gender differences in guessing tendencies were robust, they accounted for only a small fraction of the observed gender differences in multiple-choice tests.

There is evidence suggesting that when offered a choice, females overwhelmingly prefer essay items, and males show a slight preference for multiple-choice items (Gellman & Berkowitz ,1993). Neither ability level nor level of education were related to this preference. Mazzeo et al. (1993) examined male and female performance on the multiple-choice and constructed-response sections of four Advanced Placement (AP) examinations: American History, Biology, Chemistry, and English Language and Composition. A fairly small number of items were found to exhibit substantial gender-related differential item functioning, but removing these items resulted in almost no reduction in the magnitude of gender-related differences on the multiple-choice sections. The researchers also attempted to determine whether a similar pattern of gender-related differences would be observed for individual constructed-response questions or question types. The researchers found some consistent patterns across ethnic and racial groups regarding the types of questions on which females perform relatively better. Taken as a whole, the results suggested that item type variability might have a greater effect than the variability associated with particular question types or broadly defined content areas. Mazzeo et al. (1993) concluded that the major factor accounting for the relatively better performance of females on constructed-response tests might be related to the different construct measured by this type of item. Constructed-response tests probably require different sets of competencies than their multiple-choice counterparts, and gender-related differences in performance profiles across the two assessment formats most likely reflect real disparities in the average level of achievement of males and females with regard to these different competencies.

Breland, Danos, Kahn, Kubota, and Bonner (1994) explored gender differences in Advanced Placement examinations in American history. Examinations in American history were chosen because they consistently showed significant gender differences in favor of males on multiple-choice items but no gender differences on open-ended items. Even after controlling for possible irrelevant factors affecting the scoring of open-ended items in favor of females (i.e., handwriting quality, neatness, and English composition skills), no gender differences in the open-ended questions were found. Breland et al. (1994) concluded that the differential performance pattern was due to the fact that the two types of tests measured different skills, both of which are important to the study of history.

Similar conclusions were reached by Bennet (1993) who noted that a key claim in the constructed-response/multiple-choice debate is that format affects the meaning of test scores because it restricts the nature of the content and the processes that can be measured. Even minor format differences, such as those between multiple-choice and completion items, can change the nature of the construct assessed. In a more recent paper Bridgeman and Morgan (1996) also suggested that multiple-choice and essay examinations measure at least somewhat separable constructs.

Willingham and Cole (1997), in their monumental comprehensive review of gender and fair assessment, concluded that females were more likely than males to do better on free-response formats compared to multiple-choice formats, but this was not a consistent effect. They also concluded that format effects tended to vary across participants (seldom found in mathematics, language and literature, and occuring more frequently in science and geopolitics). When there was a free-response format effect favoring women, writing often appeared to play a role. Their review suggested that the multiple-format in itself does not have a consequential effect on performance and is not likely to present a significant fairness problem.

The relatively better performance of girls on open-ended tests as compared with multiple-choice tests has important implications for standardized testing. A high percentage of standardized testing is administered today in multiple-choice format, and crucial educational decisions are often made on the basis of the results of such tests. Therefore, it is of utmost importance to better understand the nature of the competencies measured by each of these item-formats, and the degree to which they are relevant and valid with regard to what is being assessed.

This study investigates the differential performance of 9- and 13-year-old boys and girls on open-ended and multiple-choice items in two International Assessment of Educational Progress (IAEP) mathematics tests conducted in 1988 (Lapointe, Mead, & Phillips, 1989), and 1991 (Lapointe, Mead, & Askew, 1992), respectively. The results of the two international studies are compared in order to determine whether a consistent pattern emerges, and whether item format, in and of itself, could account for gender differences in performance.

Method

IAEP - General

IAEP is an international comparative study of performance on tests in various school subjects, conducted by the Educational Testing Service at Princeton (ETS). More specifically, IAEP was designed to collect and report comparative data on student achievement, attitudes, backgrounds, and classroom experiences. In non-English-speaking countries, each question was translated into the appropriate language and then checked for accuracy by language experts at ETS. All countries made minor adaptations in the questions due to cultural differences (e.g., changing names, units of measurement, etc.), but these adaptations did not alter the psychometric nature and the content of the assessment questions. An important advantage of studying gender differences using the IAEP data is that the student samples used in the different countries are designed to be representative and do not suffer from selection biases of any known sort.

Approximately one fifth of the mathematics items included in both IAEP studies were open-ended questions (scored dichotomously), which required students to generate and write their own answers, while the remaining questions were of the multiple-choice type. The OE items included in the IAEP assessments were simple fill-in items. The advantage of comparing performance on such items to performance on MC items is that item format per se constitutes the main difference between them, and variables such as language styles, levels of cognitive skills, and handwriting play a lesser role and, to a certain extent, are controlled for in the present study. Moreover, whereas in many studies the two item formats are scored on different grading scales, in this study item difficulty levels of both MC and OE items are directly comparable, because both item types are scored dichotomously. This allowed for an investigation of the hypothesis, based on the results obtained from other studies, that girls perform relatively better on open-ended mathematics questions than on multiple-choice questions (i.e., multiple-choice items would produce larger gender effects in favor of boys).

IAEP 1988

Participants

The first IAEP study, conducted in 1988, provided data on the mathematics and science achievements of 13-year-olds in six countries: Canada, Ireland, Korea, Spain, the United Kingdom, and the United States (for a detailed description of the 1988 math and science assessments see Lapointe, Mead & Phillips, 1989). A representative sample of 13-year-olds in each country was assessed in both mathematics and science. In each country (with the exception of USA and Canada) samples were drawn at random from about 100 different schools selected with a probability proportional to their size, and included about 2,000 students from each country divided equally among girls and boys; half were assessed in mathematics and half in science. In the United States the samples size was only about 1,000 students from 200 schools. In Canada the data was aggregated across seven sub-samples (N = approximately 15,000). The samples were equally divided among boys and girls.

Description of the mathematics assessment

The assessment consisted of 63 questions (one item was eventually dropped because of differential item functioning [DIF] leaving a total of 62 items), and lasted 45 minutes. Twenty-two percent of the items were open-ended (OE) and the rest were multiple-choice (MC) items. Open-ended items were hand-scored, using standardized scoring guides, and responses were keyed or scanned. The items were selected from a total pool of 281 mathematics questions used in the 1986 National Assessment of Educational Progress (NAEP). The IAEP assessment consisted of items from the following six content areas typically taught in mathematics: Numbers and Operations (NUM); Measurement (MEA); Geometry (GEO); Data Organization and Interpretation (DAT); Relations, Functions and Algebraic Expressions (ALG1); Logic and Problem Solving (ALG2). The last two categories seem to include the same type of problems as those included in the category referred to as Algebra and Functions (ALG) in the 1991 IAEP study described below. A detailed description of the IAEP mathematics objectives (mathematical processes and content areas) appears in Lapointe, Mead and Phillips (1989). Table 1 presents the distribution of mathematics items by content area and format.

IAEP 1991

Participants

The second IAEP study, conducted in 1991, assessed the mathematics, science and geography skills of 9- and 13-year-olds in some twenty countries; all twenty countries participated in the mathematics and science assessment of 13-year-olds, while participation in the other assessment component was optional. The participating countries were: Brazil (Fortaleza and San-Paulo), Canada, China, England, France, Hungary, Ireland, Israel, Italy, Jordan, Korea, Mozambique, Portugal, Scotland, Slovenia, the Soviet Union, Spain, Switzerland, Taiwan, and the Unites States (for a detailed description of the 1991 math and science assessments see Lapointe, Mead & Askew, 1992). The following countries participated only in the assessment of 13-year-olds and not in the assessment of 9-year-olds: Taiwan, China, France, Jordan, Mozambique, Switzerland, and Brazil (Fortaleza and San-Paulo). Typically, a random sample[2] of 3,300 students from about 110 different schools was selected from each country at each age level; half were assessed in mathematics and half in science. In Canada the data was aggregated across seven sub-samples (in mathematics, 9,365 and 19,691 studentws aged 9 and 13, respectively were tested). The samples were divided equally among boys and girls.

The analyses in this study were first carried out for 13-year-olds in the six countries that participated in the 1988 assessment: Canada, Ireland, Korea, Spain, UK/Scotland[3], and the USA. Additional analyses were carried out for the age 13 cohort in the remaining 14 countries participating in the 1991 assessment, and for the age 9 cohort in all 14 countries participating in the 1991 assessment.

Description of the mathematics assessment

The assessment for each age level was developed through a consensus-building process involving curriculum and measurement experts from each participating country. The assessment was not aligned with the curriculum of any one country.

The mathematics assessment of 13-year-olds consisted of 75 questions (one item was excluded after a DIF analysis by IAEP) and lasted one hour. Approximately 21 percent of the items were open-ended. The study measured overall mathematics performance as well as performance in five content areas: Numbers and Operations (NUM), Measurement (MEA), Geometry (GEO), Data Analysis, Statistics and

Probability (DAT), and Algebra and Functions (ALG). Table 1 presents the distribution of mathematics items by content area.

Careful inspection of Table 1 points to similar distributions of items across content areas in the 1988 and 1991 assessments of 13-year-olds, although the various sub-domains were not equally represented by both item formats (e.g., GEO in 1988 contained only MC items).

The mathematics assessment of 9-year-olds consisted of 62 questions (one item was excluded following a DIF analysis by IAEP) and lasted one hour. Approximately 28 percent of the items were open-ended. Table 1 presents the distribution of mathematics items by content area.

A detailed description of the IAEP mathematics objectives (mathematical content areas) appears in *The 1991 IAEP Assessment: Objectives for Mathematics, Science, and Geography* (Princeton, NJ: Educational Testing Service, 1991).

Gender Effect Size

Gender differences in performance were measured by computing gender effect sizes (ES). ES is defined as the mean performance (as measured by mean item percent correct) for boys minus the mean for girls, divided by the standard deviation computed across the two groups. An ES of zero indicates that no differences were found between boys and girls; a positive ES indicates that boys have higher scores; and a negative ES reflects an advantage for girls. Effect sizes are comparable from test to test (due to the standardization procedure). The total group standard deviation contains both within- and between-group variances. Therefore, when there are large between-group variances, computed effect sizes that are based on this variance would be smaller than computed effect sizes that are based on a pooled-within variance. In this study, recomputing effect sizes using pooled variance estimates resulted in highly similar effect sizes (the largest difference between these two estimates was 0.01).

Mean item percent-correct scores were computed for each gender group across all items (Total Score); as well as for all items within each item format (MC and OE). Contrary to the practice reported for the IAEP assessment (e.g., Lapointe et al., 1992), omitted questions at the ends of sections were not excluded from the calculations for those questions, (i.e., an unreached item was counted as an error). This may result in slight discrepancies between our results and those reported in the above-mentioned IAEP reports. Gender effects were computed for each type of score. For each mean percent correct, an estimate of the standard error was calculated for each subgroup within each sample by using a jackknife replication procedure. The jackknife procedure (Wolter, 1985) provides good quality estimates of the sampling variability of most statistics. The statistical significance of the differences between boys and girls was tested using jackknifed standard errors.

Results

Gender Effects for the Total Score

Figure 1 presents the gender effects for the Total Score within each of the six countries participating in both the 1988 and 1991 assessments of 13-year-olds. In general, gender effects were relatively small (ranging from -0.12 to 0.26 of a standard deviation). Girls performed better than boys (-0.12) only in the 1988 UK sample. In Ireland, the UK, the USA and Canada larger effects were found in the 1991 study than in the 1988 one, while for Korea and Spain the reverse was true.

Gender Effect and Item Format

To examine the hypothesis that gender effect is related to item format, gender effects were calculated separately for multiple-choice (MC) and open-ended (OE) items. Figures 2 and 3 present the gender effects for all OE and all MC items within the various countries, for the 1988 and 1991 assessments of 13-year-olds, respectively.

Figure 2 clearly indicates that in 1988 the gender effect in favor of 13-year-old boys was larger for MC items (as compared with OE items). Even in countries such as the UK, where the gender effect was in favor of girls, the difference was smaller on MC items. It is of interest to note that while gender effects for the Total Score as well as for MC items were positive for all countries except the UK (see Figure 1), gender effects for scores based on OE items were positive only for Korea and Spain.

Contrary to the 1988 IAEP results and the findings reported in the literature (e.g., Bolger & Kellaghan, 1990; Traub & MacRury, 1990), Figure 3 indicates that gender effects in favor of 13-year-old boys in the 1991 IAEP assessment were not larger for MC items than for OE items. In fact, in Scotland, Spain, the USA, and Canada, the reverse pattern was found -- gender effects in favor of boys were larger for OE items than for MC items.

It should be noted that the reliability of the scores based on OE items was somewhat lower than that based on MC items due to the smaller number of OE items (see Tables 1,2,3 in the Appendix). Correcting gender effects for differences in reliability (Beller & Gafni, 1996) would have resulted in larger effects. The magnitude of the correction is a function of the reliability; the lower the reliability, the larger the magnitude of the correction. Careful inspection of Figures 2 and 3 reveals that the results obtained after correcting gender effects for differences in reliability would in general strengthen the phenomenon reported above (i.e., that in the 1988 assessment, larger gender effects were found for MC items than for OE items, while for the 1991 assessment the reverse is true).

Inspection of Table 1 reveals that the 1988 and 1991 distributions of item x content x format for age 13 were fairly similar (except for GEO), suggesting that the above-mentioned inconsistent gender effect patterns cannot be explained by differences in the internal structure of the two assessments.

Further inspection of the data raised the possibility that the inconsistent patterns of gender effects with regard to item format in the 1988 and 1991 assessments might be associated with the level of difficulty of the items in each format within each assessment. Table 2 presents the average overall difficulty level (measured by percent correct) of OE items vs. MC items within each country for the 1988 and 1991 assessments of 13-year-olds, respectively.

In 1988, the average difficulty level of MC items was similar to that of OE items, with OE items slightly easier in Ireland, Korea, Spain and Canada, and more difficult in the UK and the USA (a similar pattern was also found within each gender group). It should be remembered that in this assessment gender effects were larger for MC items, which were the more difficult items (with the exception of the UK). The overall difficulty level of the 1991 assessment was somewhat greater than that of the 1988 assessment, and interestingly enough, in contrast with the 1988 assessment, OE items were found to be more difficult than MC items for all countries (a similar pattern was also found within each gender group). It should be noted that in the 1991 assessment, gender effects were larger for OE items, and, as in 1988, they were larger for the more difficult items. In summary, gender effect in both assessments was associated with item difficulty level.

Additional Analyses of the 1991 IAEP Assessment

In the above analyses, results were compared for just the six countries participating in both the 1988 and the 1991 assessments of 13-year-olds. It was therefore of interest to examine the relationship between item format and gender effect size among 13-year-olds in all 20 countries participating in the 1991 mathematics assessment, as well as among 9-year-olds in all 14 countries that participated in the 1991 assessment.

Tables 3 and 4 present gender effects for MC and OE items, as well as their respective average item difficulty and jackknifed standard errors, for ages 13 and 9, respectively, in all countries participating in the 1991 assessment. As in the findings for the six countries in the 1991 assessment (see Figure 3), the results indicate that at age 13, girls performed relatively better on the MC items than on the OE items in the majority of countries, and at age 9 they performed relatively better in all of the countries. Furthermore, in all cases, OE items were, on the average, more difficult than MC items. These results further support the hypothesis that format-based gender effects are mediated by item difficulty level.

Analysis of the Relationship Between Gender Effect Size and Item Difficulty

The hypothesis of a possible relationship between gender effect size and item difficulty was further examined by correlating item difficulty and item effect size (within each assessment and age group). Item effects were first computed by subtracting the mean percent correct for girls from that of boys and dividing the difference by the standard deviation of the item scores across the two gender groups.

There is an inherent curvilinear relationship between gender effect size and item difficulty as measured by percent correct (i.e., for a given constant gender difference -- as item difficulty approaches both extremes, gender effect size increases). The observed relationship between gender effect and item difficulty depends on the range of the difficulty levels of the items. When the difficulty level of most of the items is above 0.5, the observed relationship between item difficulty and gender effect would be linear even for a constant gender difference[4]. To control for such an artifact, item variance must be stabilized. This was done by normalizing the mean percent correct of the items and recomputing the gender effects by subtracting the mean normalized z score for girls from that of boys.

Correlations were computed between the normalized percent correct and the normalized gender effect and are presented in Table 5. Normalized percent correct was computed in such a way that the more difficult the item, the higher the respective normalized value. Correlations were computed separately within each age grup for each of the participating countries as well as across all countries participating in the 1988 and the 1991 assessments. A clear pattern emerges: The correlations between item percent correct and item effect size were consistently positive (Korea in 1988 and Mozambique in 1991, were the only exceptions). These positive correlations reflect the fact that larger effect sizes were found on the more difficult items.

An identical analysis, carried out separately within MC items and OE items, yielded similar results. Table 6 presents the correlations across all countries for all items, as well as separately for MC and OE items, within each age group. These results strongly support the hypothesis that effect size increases with item difficulty regardless of item format. This held true across all items as well as within each item format for both assessments and age groups.

Discussion

This study examines item format-based gender differences on two international mathematics assessments consisting of representative samples of boys and girls. Specifically, gender effects on MC and OE items were investigated for the 1988 and 1991 IAEP mathematics assessments. Generally, the term “open-ended” encompasses a wide range of formats, from simple fill-in items to complex performance assessments. As mentioned previously, the OE items included in the IAEP assessments were simple fill-in items, not extended free responses. The advantage is that such items allow for a more accurate examination of the question of item-format per-se, because the two item types differ only in format and not in the construct measured and the scale used. In other types of open-ended items the format and the construct measured by the items seem to be confounded.

Gender effects on MC and OE items were compared for the six countries that participated in both the 1988 and 1991 IAEP mathematics assessments of 13-year-olds. As in other studies (e.g., Bell & Hay, 1987; Bolger & Kellaghan, 1990; Bridgeman & Morgan, 1996; Mazzeo et al., 1993; Murphy, 1982; Traub & MacRury, 1990), in the 1988 IAEP mathematics assessment the gender gap in favor of boys was larger on MC items than on OE items. While the gender effects for the Total Score in this assessment were positive (in favor of boys) for all countries except the UK, the effects for scores based on OE items were positive only for Korea and Spain.

Interestingly, the 1991 IAEP assessment yielded the opposite results: gender effects were not found to be larger on MC than on OE items. In fact, in the majority of the 20 countries participating in the assessment of 13-year-olds, and in all 14 countries participating in the assessment of 9-year-olds, gender effects in favor of boys were greater on OE items than on MC items. These inconsistent results challenge the simplistic assertion that girls perform relatively better on OE test items; they suggest that item format alone cannot account for gender differences in mathematics performance, and that in order to understand gender differences in mathematics, a more complex explanation than merely the mode of assessment is required.

It should be noted that the above analysis was carried out across all items included in each assessment. However, there is the possibility that item format and item content/cognitive operation were confounded (i.e., certain content and cognitive operations are better measured by a particular item format and therefore appear more often in that format). To allow for a better understanding of the relationship between test format and gender differences, item content should be held constant, and gender effects should be estimated for MC and OE items within each content category. Due to the small number of items in each category, it was impossible to perform this analysis for the IAEP assessments. However, since the distributions of item format were found to be similar in the 1988 and 1991 assessments, such confounding cannot explain the format-based differences between the two assessments.

Other confounding variables might explain the relationship frequently found between item format and gender effect. The analyses performed in this study investigated the hypothesis that one possible confounding variable is item difficulty. Indeed, gender effects in favor of boys were larger on the more difficult items for both the 1988 and 1991 data sets (a positive correlation between item difficulty and gender effect was obtained in 39 out of the 41 data sets). In the1988 assessment, the average difficulty level of MC items was similar to that of OE items, with OE items slightly easier than MC items in most countries, and correspondingly, gender effects were relatively larger on MC items. By comparison, the 1991 OE items were more difficult (compared with MC items) for all countries in both age groups, and correspondingly, gender effects were relatively larger on OE items.

To correct for the possibility of an artifactual relationship between item difficulty level and gender item effect (caused by the fact that the difficulty of all items was above 50 percent correct), correlations between normalized item difficulties and gender effects were computed. It should be noted that in most countries, a considerable proportion of the items were below the 50 percent difficulty level. Furthermore, even if the positive correlation between item difficulty level and item gender effect size were artifactual, it could not be argued that item format rather than item difficulty accounted for gender differences in performance.

The results of this study support Willingham and Cole's assertion (1997) that in mathematics, gender differences formats have not been generally observed between MC and OE. This means that item format per se cannot account for gender differences in test performance. The results indicate that a relationship exists between item difficulty and gender effect regardless of item format, suggesting that boys do relatively better than girls as items increase in difficulty. This result corresponds to previous results showing that the average ratio of girls to boys among the top 10% math scores is 0.63 (Hedges & Nowell, 1995) to 0.70 (Wellingham & Cole, 1997). Benbow (1988) and Stanley (1993) reported similar results. These findings suggest that among those answering the most difficult questions correctly there are more boys then girls. Duffy, Gunther, and Walters (1997) found that 12-year-old girls performed better than boys on questions of low difficulty, while both genders performed equivalently on questions of high difficulty. These somewhat different findings might be attributed to a different definition of item difficulty (based on rater judgements) and on small and non-representative samples. However, it is not clear what it is about the item that makes it difficult. Items that measure problem solving and reasoning are likely to be among the more difficult questions, whereas computation items are likely to be easier (e.g., Wellingham & Cole, 1997). Detecting the cognitive variables that make items more difficult and that are confounded with item format would require further analyses based on a well-designed experiment that carefully controls for variables such as level of cognitive operation, handwriting, content, item format, and item difficulty. It would be of great interest to design and conduct an experiment that manipulates item difficulty while keeping the content specifications of the items constant. This future study should be carried out also among older boys and girls, because in general, observations made at the ages of the participants in the current study show smaller gender differences then might be observed at later ages (e.g., Willingham & Cole, 1997).

It is believed that the first priority should be given to what is measured rather than how it is measured. Revealing those cognitive variables that affect item difficulty is, in and of itself, of great theoretical and practical importance, and may also provide a better understanding of gender differences in test performance.

References

Bell, R. C., & Hay, J. A. (1987). Differences and biases in English language examination formats. British Journal of Educational Psychology, 57, 212-220.

Beller, M., & Gafni, N. (1996). The 1991 International Assessment of Educational Progress in Mathematics and Sciences: The gender differences perspective. Journal of Educational Psychology, 88, 365-377.

Benbow, C. P. (1988). Sex differences in mathematical reasoning ability in intellectually talented preadolescents: Their nature, effects, and possible causes. Behavioral and Brain Sciences, 11, 169-183.

Ben-Shakhar, G., & Sinai, Y. (1991). Gender differences in multiple-choice tests: The role of differential guessing tendencies. Journal of Educational Measurement, 28, 23-35.

Bennet, R. E. (1993). On the meanings of constructed-response. In R. E. Bennet & W. C. Ward (Eds), Construction versus choice in cognitive measurement: Issues in constructed-response, performance testing and portfolio assessment (pp. 1-27). Hillsdale, NJ.

Bolger, N. (1984). Gender differences in academic achievement according to method of measurement. Paper presented at the annual meeting of the American Psychological Association, Toronto.

Bolger, N., & Kellaghan, T. (1990). Method of measurement and gender differences in scholastic achievement. Journal of Educational Measurement, 27, 165-174.

Breland, H. M., Danos, D. O., Kahn, H. D., Kubota, M. Y., & Bonner, M. W. (1994). Performance versus objective testing and gender: An exploratory study of an Advanced Placement History Examination. Journal of Educational Measurement, 31, 275-293.

Bridgeman, B., & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. Journal of Educational Psychology, 88, 333-340.

Duffy, J., & Gunther, G. (1997). Gender and mathematical problem solving. Sex Role, 37, 477-494.

Feingold, A. (1992). Sex differences in variability: A new look at an old controversy. Review of Educational Research, 62, 61-84.

Gafni, N., & Melamed, E. N. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. Studies in Educational Evaluation, 20, 309-319.

Gelleman, E. S., & Berkowitz, M. (1993). Test item type: What students prefer and why? College Student Journal, 27, 17-26.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. Science, 269, 41-45.

Hyde, J. S., Fennema, I., & Lamon, S. J. (1990). Gender differences in mathematics performance attitudes/affect: A meta-analysis. Psychological Bulletin, 107, 139-155.

Lapointe, A. E., Mead, N. A., & Askew, J., M. (1992). Learning Mathematics. The International Assessment of Educational Progress. Princeton, NJ: Educational Testing Service. Report No. 22-CAEP-01.

Lapointe, A. E., Mead, N. A., & Phillips (1989). A world of differences. An international assessment of mathematics and science. The International Assessment of Educational Progress. Princeton, NJ: Educational Testing Service.

Linn, M. C. (1991). Gender differences in educational achievement. In Sex Equity in Educational Opportunity, Achievement, and Testing. (Proceedings of the 1991 ETS Invitational Conference). ETS, Princeton, NJ 08541.

Lumis, M., & Stevenson, H. W. (1990). Gender differences in beliefs and achievement: A cross-cultural study. Developmental Psychology, 26, 254-263.

Maccoby, E. E., & Jacklin, C. N. (1974). The psychology of sex differences. Stanford, CA: Stanford University Press.

Mazzeo, J., Schmitt, A., & Bleistein, C. (1993). Sex-related differences on constructed response and multiple-choice sections of Advanced Placement Examinations (CB Report No. 92-7, ETS RR 93-5). New York: College Entrance Examination Board.

Murphy, R. J. L. (1980). Sex differences in GCE Examination entry statistics and success rates. Educational Studies, 6, 169-178.

Murphy, R. J. L. (1982). Sex differences in objective test performance. British Journal of Educational Psychology, 52, 213-219.

Stanley, J. C. (1993). Boys and girls who reason well mathematically. In G. R. Bock & K.Ackrill (Eds.), Ciba Foundation Symposium 178, The origins and development of high ability (pp. 119-138). Chichester, England: Wiley.

Steinkamp, M. W., Harnisch, D. L., Walberg, H. J., & Tsai, S. N. (1985). Cross-national gender differences in mathematics attitude and achievement among 13-year-olds. The Journal of Mathematical Behavior, 4, 259-277.

The 1991 IAEP Assessment. Objectives for mathematics, science, and geography (1991). Educational Testing Service, Princeton NJ.

Traub, R. E., & MacRury, K. (1990). Antwort-Ausswahl vs. Freie-Antwort-Aufgaben Bei Lernerfolgstestes. In Test und Trends 8: Jarbuch der Paedgogischen Diagnostik, edited by K. Ingekamp and R. S. Jager. Weinheim. Germany: Beltz-Verlag Publishing Co. (English-language version, entitled Multiple-Choice vs. Free-Response in the Testing of Scholastic Achievement, is available from the authors.)

Willingham, W. W., & Cole, N. S. (1997). Gender and fair assessment. Mahwah New-Jersy: Larwrence Erlabuam Associates.

Wolter, K. M. (1985). Introduction to variance estimation. New York: John Wiley & Sons.

Appendix

Table 1

Internal Consistency Reliabilities (KR-20) for the 1988 Assessment - Age 13

	MC - 48 Items			OE - 14 Items		
Country	Total	Males	Females	Total	Males	Females
Ireland	0.94	0.94	0.94	0.89	0.91	0.87
Korea	0.93	0.93	0.93	0.91	0.91	0.91
UK	0.95	0.96	0.95	0.93	0.92	0.93
Spain	0.93	0.93	0.93	0.85	0.85	0.84
USA	0.95	0.95	0.95	0.92	0.93	0.92
Canada	0.92	0.92	0.92	0.86	0.86	0.85

Table 2

Internal Consistency Reliabilities (KR-20) for the 1991 Assessment - Age 13

	MC - 59 Items			OE - 16 Items		
Country	Total	Males	Females	Total	Males	Females
Ireland	0.92	0.93	0.91	0.88	0.88	0.87
Korea	0.94	0.95	0.94	0.87	0.88	0.86
Scotland	0.92	0.92	0.91	0.84	0.84	0.84
Spain	0.89	0.90	0.87	0.81	0.82	0.79
USA	0.92	0.93	0.92	0.87	0.87	0.86
Canada	0.91	0.91	0.90	0.82	0.82	0.81
Hungary	0.94	0.94	0.93	0.87	0.88	0.87
Israel	0.92	0.92	0.91	0.84	0.85	0.83
England	0.92	0.93	0.91	0.86	0.88	0.84
Slovenia	0.91	0.91	0.90	0.84	0.85	0.83
Italy	0.92	0.93	0.90	0.84	0.85	0.83
Portugal	0.89	0.90	0.88	0.81	0.82	0.79
Soviet Union	0.93	0.93	0.93	0.87	0.87	0.86
Taiwan	0.96	0.96	0.96	0.91	0.91	0.91
China	0.91	0.90	0.92	0.79	0.78	0.80
France	0.92	0.92	0.92	0.84	0.83	0.84
Jordan	0.89	0.88	0.89	0.85	0.84	0.85
Mozambique	0.67	0.67	0.68	0.63	0.61	0.65
Switzerland	0.91	0.91	0.90	0.81	0.81	0.80
Fortaleza	0.86	0.86	0.85	0.80	0.81	0.80
Sao-Paulo	0.89	0.90	0.88	0.83	0.85	0.82

Table 3

Internal Consistency Reliabilities (KR-20) for the 1991 Assessment - Age 9

	MC - 44 Items			OE - 17 Items		
Country	Total	Males	Females	Total	Males	Females
Ireland	0.90	0.91	0.89	0.86	0.87	0.85
Korea	0.88	0.88	0.88	0.81	0.78	0.82
Scotland	0.89	0.89	0.89	0.83	0.84	0.82
Spain	0.90	0.91	0.90	0.83	0.85	0.82
USA	0.90	0.90	0.89	0.85	0.86	0.85
Canada	0.87	0.88	0.86	0.82	0.83	0.81
Hungary	0.91	0.91	0.91	0.84	0.85	0.84
Israel	0.89	0.89	0.90	0.81	0.81	0.81
England	0.90	0.90	0.91	0.87	0.85	0.89
Slovenia	0.84	0.84	0.84	0.73	0.75	0.72
Italy	0.90	0.89	0.89	0.83	0.83	0.82
Portugal	0.87	0.89	0.86	0.83	0.84	0.82
Soviet Union	0.91	0.91	0.90	0.86	0.86	0.86
Taiwan	0.91	0.91	0.91	0.82	0.82	0.83

Author Notes

Table 1

Distribution of Mathematics Items by Content Area and Format for Age 13 in the 1988 Assessment and for age 13 and 9 in the 1991 Assessment

Content Format	NUM	MEA	GEO	DAT	ALG	Total
1988 - Age 13						
MC	17	7	8	5	11	48
OE	7	3	0	1	3	14
Total	24	10	8	6	14	62
1991- Age 13						
MC	24	11	7	6	11	59
OE	3	2	4	3	4	16
Total	27	13	11	9	15	75
1991 - Age 9						
MC	23	6	4	8	3	44
OE	9	3	2	0	3	17
Total	32	9	6	8	6	61

Table 2

Average Performance on MC Items vs. OE Items across Gender Groups for Age 13 in the 1988 and 1991 Assessments (jackknifed standard errors appear in parentheses)

	1988		1991	
Country	MC	OE	MC	OE
Ireland	61.3	63.6	59.9	53.7
	(0.8)	(1.0)	(0.9)	(1.7)
Korea	75.2	77.1	74.2	68.5
	(0.5)	(0.7)	(0.6)	(0.8)
UK/Scotland	66.5	48.6	60.0	58.2
	(1.2)	(1.2)	(0.9)	(1.0)
Spain	61.9	65.6	53.8	44.1
	(1.1)	(1.1)	(0.8)	(1.1)
USA	55.3	53.8	56.8	44.9
	(1.0)	(1.1)	(1.0)	(1.2)
Canada	67.2	67.7	62.3	57.9
	(0.4)	(0.5)	(0.5)	(0.7)

Table 3

Gender Effect Size (ES)*, Average Difficulty Level (P) and Jackknifed Standard Error (SE) of MC and OE Items within Each Participating Country for Age 13 in 1991.

	MC - 59 Items			OE - 16 Items		
Country	ES	P	SE	ES	P	SE
Ireland	0.18	59.9	0.9	0.17	53.7	1.7
Korea	0.10	74.2	0.6	0.10	68.5	0.8
Scotland	-0.03	60.3	0.9	0.04	58.2	1.0
Spain	0.14	53.8	0.8	0.24	44.1	1.1
USA	0.02	56.8	1.0	0.10	44.9	1.2
Canada	0.08	62.3	0.5	0.2	57.9	0.7
Hungary	-0.04	66.8	0.7	0.01	63.5	0.9
Israel	0.13	62.1	0.8	0.21	58.7	1.0
England	-0.01	59.9	2.0	0.01	57.2	2.8
Slovenia	0.09	56.8	0.7	0.16	53.8	1.1
Italy	0.15	63.2	0.8	0.31	56.5	1.1
Portugal	0.04	49.6	0.7	0.07	37.1	1.0
Soviet Union	-0.04	68.1	1.1	-0.04	67.6	1.6
Taiwan	0.01	72.8	0.7	0.06	70.5	0.9
China	0.19	80.9	1.0	0.22	75.1	1.2
France	0.12	63.8	0.7	0.19	60.9	1.0
Jordan	0.11	41.4	1.0	0.19	32.3	1.2
Mozambique	0.00	27.5	0.3	0.06	16.1	0.3
Switzerland	0.22	70.6	1.2	0.27	65.6	1.7
Fortaleza	0.25	32.2	0.6	0.23	19.6	0.7
Sao-Paulo	0.07	36.5	0.7	0.05	24.3	1.2

* Significant gender differences on OE items (α=0.05) were found for China, France, Slovenia, Israel, Italy, Spain, Fortaleza, Canada and Switzerland. Gender differences on MC items were significant for Ireland, Italy, Fortaleza and Switzerland.

Table 4

Gender Effect Size (ES)*, Average Difficulty Level (P) and Jackknifed Standard Error (SE) of MC and OE Items within Each Participating Country for Age 9 in 1991.

	MC - 44 Items			OE - 17 Items		
Country	ES	P	SE	ES	P	SE
Ireland	-0.09	60.9	0.8	0.0	50.9	1.2
Korea	0.23	74.8	0.6	0.35	74.1	0.7
Scotland	-0.01	67.4	0.8	0.02	58.3	1.1
Spain	0.01	62.5	1.0	0.03	52.9	1.1
USA	0.02	62.5	1.0	0.10	45.9	1.3
Canada	-0.04	61.9	0.5	0.04	52.1	0.7
Hungary	-0.04	64.6	0.7	0.0	64.1	0.9
Israel	0.14	65.1	0.7	0.19	55.0	0.9
England	-0.12	61.5	1.7	-0.03	46.7	2.6
Slovenia	-0.05	52.8	0.5	0.04	47.9	0.6
Italy	0.17	66.7	0.9	0.18	59.2	1.1
Portugal	0.13	56.1	0.9	0.17	48.2	1.2
Soviet Union	0.01	64.4	1.3	0.07	62.4	1.5
Taiwan	0.02	68.6	0.8	0.05	64.1	0.9

* Significant gender differences on OE items were found for Israel, Italy, Korea, and Portugal; on MC items: Israel, Italy, and Korea.

Table 5

Correlations between Item Percent Correct and Item Effect Size (Normalized) for the 1988 and 1991 Assessments within Each Participating Country

Country	1988 - Age 13	1991 - Age 9	1991 - Age 13
Ireland	0.35	0.55	0.63
Korea	0.00	0.34	0.41
UK/Scotland	0.09	0.26	0.30
Spain	0.19	0.50	0.28
USA	0.32	0.58	0.45
Canada	0.24	0.53	0.38
Hungary	*	0.32	0.27
Israel	*	0.32	0.40
England	*	0.35	0.41
Slovenia	*	0.29	0.38
Italy	*	0.35	0.52
Portugal	*	0.32	0.32
Soviet Union	*	0.10	0.30
Taiwan	*	0.16	0.34
China	*	*	0.39
France	*	*	0.28
Jordan	*	*	0.22
Mozambique	*	*	-0.18
Switzerland	*	*	0.21
Fortaleza	*	*	0.24
Sao-Paulo	*		0.37

* did not participate in the assessment

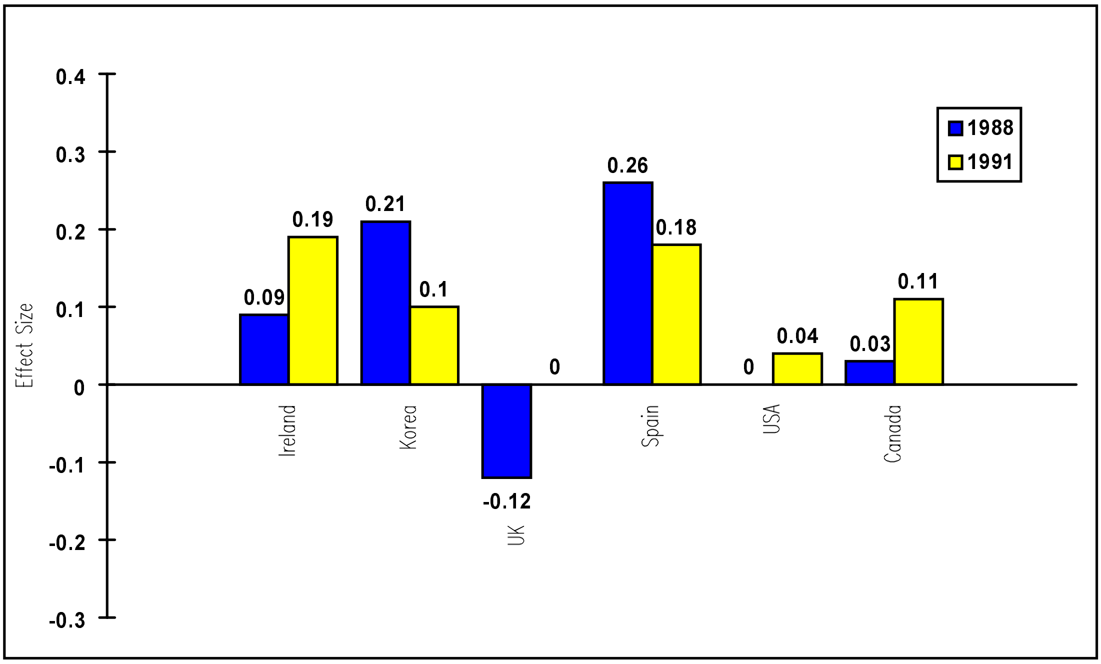
Table 6

Correlations between Item Percent Correct and Item Effect Size (Normalized) Across All Countries within Each Age Group and Assessment

	1988 - Age 13	1991 - Age 9	1991 - Age 13
All Items	0.26	0.47	0.53
MC	0.28	0.41	0.54
OE	0.32	0.51	0.51

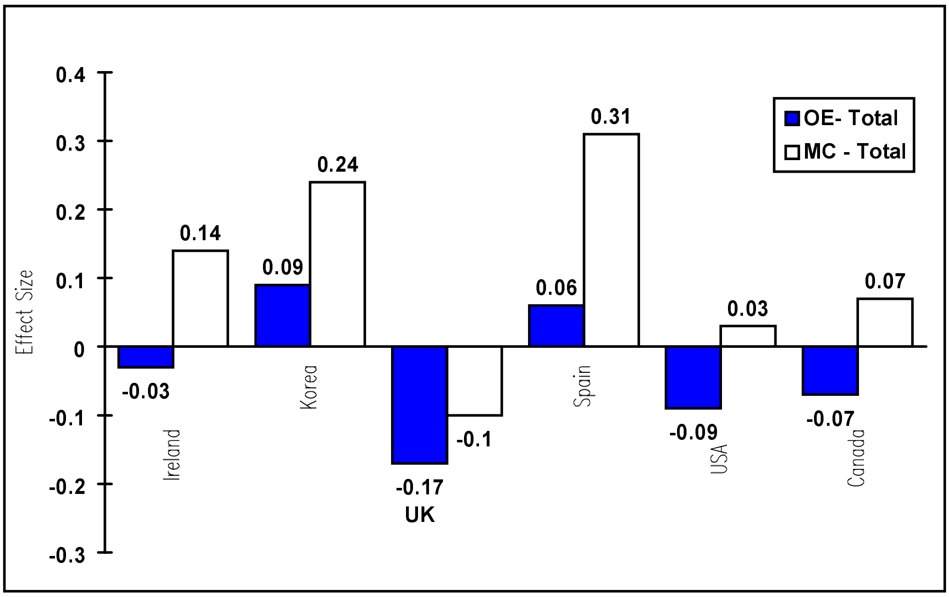
Figures

Figure 1: Gender Effect for the Total Score in the 1988 and 1991 Assessments for Age 13



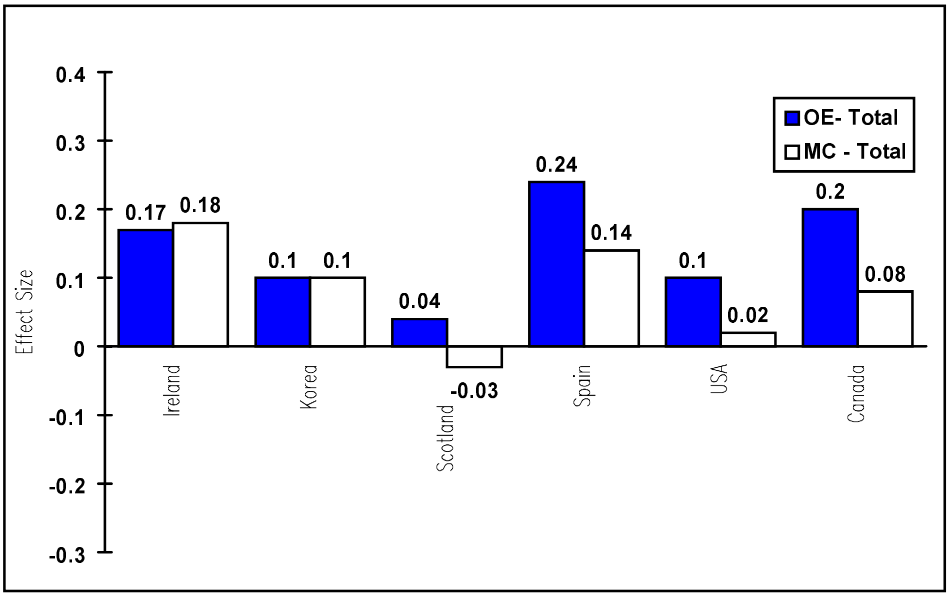
In 1988 gender differences were significant (α=0.05) only for Korea and Spain. In 1991 gender differences were significant for Ireland, Spain and Canada.

Figure 2: Gender Effects Found for OE and MC Items in the 1988 Assessment for Age 13



On OE items the only significant (negative) gender difference (α=0.05) was found for the UK. Gender differences on MC items were significant (positive effects) for Ireland, Korea and Spain.

Figure 3: Gender Effects Found for OE and MC Items in the 1991 Assessment for Age 13



On OE items the only significant gender difference (α=0.05) was found for Spain and Canada. Gender differences on MC items were significant (positive effects) for Ireland.

Notes

[1] This data set served the authors in analyzing gender differences in mathematics and sciences across all participating countries (see, Beller and Gafni, 1996).

[2] In several countries (i.e., China, England, Portugal, Brazil, and Mozambique for age 13; and Italy, Scotland, England, and Portugal for age 9) participation rates were relatively low, or certain sub-populations were excluded.

[3] In 1988 the UK sample was drawn from England, Scotland, and Wales combined. In 1991 separate samples were drawn from Scotland and England, and results are presented here only for Scotland due to some restrictions in the English sample.

[4] The distribution of the difficulty level of the items in the various countries was examined, and in the majority of the cases it contained substantial number of items with a difficulty level below 50 percent.