Journal of Educational Measurement
Summer 1990, Vol. 27, No. 2, pp. 165–174

Method of Measurement and Gender Differences in Scholastic Achievement

Niall Bolger
University of Denver
and
Thomas Kellaghan
Educational Research Centre, St. Patrick's College, Dublin

Gender differences in scholastic achievement as a function of method of measurement were examined by comparing the performance of 15-year-old boys (N = 739) and girls (N = 758) in Irish schools on multiple-choice tests and free-response tests (requiring short written answers) of mathematics, Irish, and English achievement. Males performed significantly better than females on multiple-choice tests compared to their performance on free-response examinations. An expectation that the gender difference would be larger for the languages and smaller for mathematics because of the superior verbal skills attributed to females was not fulfilled.

The influence of method of measurement on research findings has long been recognized (e.g., Madaus, 1967; Smith, 1933). In drawing a distinction between method and trait, Campbell and Fiske (1959) suggested that a considerable portion of variance in test scores may be due to the form of the test used (method) as well as to the individual characteristics (traits) that the test is designed to measure. The fact that measurement method may affect test performance has implications beyond the domain of research because it may also affect students' school-examination performance, on the basis of which important educational and vocational decisions may be made.

An important distinction in the method of measuring scholastic achievement is that between objective multiple-choice and free-response modes, the equivalence of which has received some attention (e.g., Ackerman & Smith, 1988; Traub & Fisher, 1977). American experience with free-response modes has been limited, though in recent years methods other than multiple choice have been proposed as alternatives or supplements to objective tests (Dwyer, 1979). In Europe, the reverse is happening. Countries with a long tradition of free-response, essay-type public examinations have been adding multiple-choice questions to those examinations.

One might expect a variety of possibly interrelated student characteristics (e.g., gender, race, cognitive style, testwiseness, risk taking, guessing behavior) to interact with measurement method (see Frary, 1985; Rowley, 1974). In this study, our concern is with gender. A limited amount of evidence of an association

The authors are grateful to Owen Egan and Michael Martin for their assistance in carrying out analyses.

between gender and measurement method already exists in the British literature, although the topic has received little attention in America. In a wide range of subjects in public examinations in Britain, males have been found to perform relatively better than females on multiple-choice tests compared to their performance on essay-type tests (Harding, 1980; Murphy, 1982; Wood, 1976, 1978). Furthermore, gender-related differences in performance on public examinations in Britain have been related to a change from the use of only essay-type questions to a combination of essay-type and multiple-choice items (Mathews, 1985). For example, Murphy (1980) provided time-series evidence that, following the introduction of a multiple-choice paper into a 1977 examination in geography, on which the performance of male and female candidates had always been similar, the percentage of male candidates who obtained A, B, or C grades became approximately 10% higher than the equivalent figure for female candidates.

One explanation offered for gender differences in achievement associated with method invokes the influence of a trait other than the one ostensibly being measured. According to this explanation, females possess a relative advantage on essay-type questions because of their superior verbal ability. This explanation is consistent with Murphy's (1982) finding that the male advantage in achievement, which was associated with multiple-choice methods for a wide range of subjects, was not consistent in the case of mathematics; because verbal content is relatively low in mathematics examinations, whether these involve multiple-choice or free-response formats, there is little opportunity for females to exhibit their verbal superiority over males.

The present study has two aims. The first is to test the generalizability of the British public-examination findings that gender differences in achievement are related to measurement method. We do this by using measurement instruments and subjects that differ from those employed in the British studies. Our multiple-choice tests are different in that they were designed to cover the whole syllabus prescribed for the essay-type examinations, not just part of it. Furthermore, our sample involved students from another country—Irish boys and girls, aged about 15 years, who had completed three years of postprimary (high) school. The second aim of the study is to test the adequacy of the verbal hypothesis as an explanation of gender differences (if found) using a statistical model (repeated measures ANOVA) that differs from the one used in Murphy's (1982) study (a sequence of t tests).

In our study, the performance of students is examined in three school subjects (mathematics, Irish, English) using two methods (standardized multiple-choice tests and free-response public examinations). It is obviously critical for the study design that both the multiple-choice and free-response tests cover the same content and skills; if multiple-choice tests other than ones based on the syllabi for the free-response tests had been used, it could be argued that the tests differed in the trait that was measured as well as in method of measurement. In fact, both types of test used in our study were based on the same syllabus, which had been designed by the Irish Department of Education and was followed by all schools preparing students for the Intermediate Certificate Examination (the free-response tests in our study). Thus, although our multiple-choice tests are similar

in format to the commercial multiple-choice tests used in the United States, they differ from them in the content and skills they cover. Although commercially available test batteries in the United States at the high-school level are oriented more to the basic skills of literacy and numeracy than to what is taught in specific subject fields such as mathematics and English literature (Madaus, 1988), the tests used in our study were designed to reflect curricula prescribed for schools.

Tables of specification were not available for the free-response examinations. On the basis of an analysis of the prescribed syllabus and of past examinations in each subject, such a table was drawn up for the multiple-choice tests (Educational Research Centre, 1978). In deriving its specifications from the syllabus for the free-response examinations, the authors made every effort to ensure that the same content and skills were covered in both tests. Thus, in the case of the language tests (Irish and English), both types of test covered reading comprehension of continuous everyday prose, comprehension of literary passages (prose and poetry), and knowledge of grammar. Both types of test also contained a relatively large number of items (over 100 on each test). However, the tests differed in the type of response that was required of the student. Although many questions were similar on both types of test, the free-response tests required the student to recall information and produce an answer (e.g., "Describe the underwater world shown to us in O'Flaherty's short story 'The Rockfish' ") rather than recognize a correct answer and check it. There were other differences between the two types of language test. First, the student was allowed to answer some items and not others on the free-response tests; no such choice existed on the multiple-choice tests. Second, in addition to short-answer questions, the student was required to write an essay on the free-response tests (e.g., "Write a short essay on 'Sunday in our home' ") but not on the multiple-choice tests. And third, although there were formal tests of vocabulary and of spelling on the multiple-choice tests, a student's ability to use vocabulary and to spell were assessed on the basis of her or his answering throughout the free-response tests.

There were fewer differences between multiple-choice and free-response tests in the case of mathematics. Both tests consisted of about 100 items that covered computational skills, mathematical concepts, and problem-solving skills. The main difference between the tests lay in the fact that in the multiple-choice test the student picked the correct answer from among four alternatives, whereas the free-response format required the student to supply the correct answer. Sometimes, however, the writing skills demanded in the free-response test went beyond the simple provision of a numerical value. For example, one item required a student to explain how to construct a tangent to a circle at a point on the circle.

Method

Sample

Data were drawn from a database established on a stratified random sample of Irish postprimary (high) schools. All schools in the Republic of Ireland were stratified by type (secondary, vocational, comprehensive), by gender of student served (male, female, or mixed), and by size—large (greater than 350 students) or small (equal to or less than 350 students). Within strata, schools were selected

randomly. Test scores and examination results were obtained for 739 male and 758 female students in 37 schools. The students who participated in the study were in their third year in postprimary school and were about 15 years old.

Measures

Multiple-choice tests. The Drumcondra Attainment Tests in mathematics, Irish, and English, Level VI, Form A, are norm-referenced tests that were standardized in Ireland (Educational Research Centre, 1978). The English and Irish tests comprise subtests designed to measure vocabulary, comprehension, language usage, and spelling. The mathematics test has subtests in computation, concepts, and problem solving. In the present study, a student's score on each subtest is simply the number of items correctly answered. Subtest scores were added for each subject area (Irish, English, and mathematics), and these total scores were used in analyses. There was no correction for guessing. Test-retest reliability coefficients for individual subtests vary between .71 and .90.

Free-response examinations. The Intermediate Certificate Examination of the Irish Department of Education may be taken by students who have followed approved courses of not less than three years' duration in a recognized postprimary school (Ireland: Department of Education, 1976). Courses leading to examinations are offered in over 20 subjects, but most students sit for examinations in between 6 and 8 subjects. The performance of students on three subjects—mathematics, Irish, and English—is examined in the present study. Practically all students taking the Intermediate Certificate Examination take an examination in these three subjects at either a "higher" (honors) level or a "lower" level.

Marking schemes for the examinations are given to examiners (usually teachers working under the direction of the Department of Education) but are not published. The schemes are quite detailed and indicate the number of marks to be assigned to each question as well as to such topics as accuracy of factual information, language usage, and spelling. Model answers for each question are also provided. To help ensure uniformity in marking, a conference of examiners is held to discuss the interpretation of the marking schemes. Examiners are also required to return scripts to a supervisor over regular intervals so that standards of marking can be checked. The distribution of each examiner's marks is checked against the distribution for examiners in general, and deviations from the general distribution are checked further (Greaney & Kellaghan, 1979). Estimates of the reliability of the examinations are not available.

Because scores on examinations are converted to letter grades and are reported in that form, some numerical transformation was necessary for statistical analysis. It was also necessary to equate performance on higher- and lower-level papers. Although both procedures are to some extent arbitrary, a scale developed by Martin & O'Rourke (1984) has been found to correlate well with other measures of achievement and was used for this purpose. Students' examination grades were mapped onto this scale in the following manner: Higher paper A = 11, B = 10, C = 8, D = 7, E = 4, F = 2, No grade = 0; Lower paper A = 9, B = 6, C = 5, D = 3, E = 1, F = 0, No grade = 0.

Procedure

The standardized multiple-choice tests of achievement were administered to students by their classroom teachers in the middle of the school year (all within a period of one month) as part of an experimental testing program in schools. Tests were returned to a central research agency for scoring; students' scores were available at the agency for the analyses described in this paper.

At the end of the school year, students sat for the free-response public examination (the Intermediate Certificate Examination). The examination is normally taken in a student's own school but is administered and scored under the direction of the Irish Department of Education. Information on the performance of students was obtained from the schools attended by students.

It should be pointed out that the Intermediate Certificate Examination is taken very seriously by most students, because performance on it can have important consequences for students' future educational careers (though this is less true now than it was in the past). The standardized tests, on the other hand, were presented to students as part of an experimental testing program. Thus, it is reasonable to assume that the tests were approached with less seriousness than the public examinations. However, it does not follow from this that the standardized tests were taken lightly by students. Most tests are treated with a degree of seriousness in Irish schools. Besides, as far as the present study is concerned, we have no reason to believe that girls were less serious than boys in their approach to the standardized tests.

Analysis

A mixed-model repeated measures analysis of variance (ANOVA), which treated each of the six achievement variables as repeated measures of a single variable (scholastic achievement), was used (Searle, 1971; Winer, 1971). The repeated measures, representing the factors Trait and Method, formed the within-subjects design of the ANOVA. The between-subjects design consisted of a single factor, Gender. In this procedure, the fixed components of the mixed model are Gender, Trait, and Method; the subjects are treated as a random factor.

In the language of the design above, our hypothesis that males perform relatively better versus females on multiple-choice tests than on free-response tests would require a significant Gender $\times$ Method interaction. Similarly, our hypothesis that the method-based gender difference is larger for the languages and smaller for mathematics would require a significant Gender $\times$ Method $\times$ Trait interaction. Although these interactions are necessary to support our hypotheses, they are not of course sufficient; the direction of mean differences might not support our predictions.

Results

Raw scores on the six measures of achievement were converted to z scores. In this way, the main effects of Trait and Method and the Trait $\times$ Method interaction effect were set to zero because they were not of interest in the study.

Table 1 presents means and standard deviations of the z transformation

Table 1
Means and Standard Deviations for Multiple-choice and Free-response Measures of Achievement by Gender in z-scores

	Multiple-choice			Free-response		
	Math	Irish	English	Math	Irish	English
Males						
Mean	.269	-.024	.007	.163	-.078	-.048
SD	.992	1.026	1.050	1.004	.994	1.000
Females						
Mean	-.262	.024	-.007	-.159	.076	.047
SD	.936	.974	.949	.970	1.000	.995
Gender Difference (Female mean - Male Mean)	-.531	.048	-.014	-.322	.154	.095

separately by gender. Two patterns are discernible. First, there is evidence of gender differences at the level of Trait, regardless of method of measurement. This is most clear for mathematics: Males score one third of a standard deviation higher than females on the free-response test and almost one half of a standard deviation higher on the multiple-choice test. Gender differences by Trait are less pronounced for the languages and are qualified by method. Girls do better than boys on all language measures except the multiple-choice English measure, on which the difference is negligible.

The second pattern concerns the consistent way in which the multiple-choice measures, compared to free-response measures, favor males, and conversely, the consistent female advantage on free-response measures. This effect can be seen by comparing for each trait the gender differences in achievement (in the final row of Table 1) obtained using the two measurement methods. For Irish, the male advantage associated with method of measurement is .106 standard deviation units; for English, it is .109 standard deviation units, and for mathematics, it rises to .209 standard deviation units. Thus, on average, there is an effect size of approximately .143 standard deviation units across all traits. That is, on average, females score .143 standard deviation units higher relative to males when traits are measured by free-response rather than by multiple-choice methods (and vice versa for males).

Table 2 provides results of the repeated measures ANOVA of the standard deviate scores. As is to be expected from the data in Table 1, the hypothesis of a substantial Gender $\times$ Method interaction is supported ($p < .001$), reflecting female superiority when the free-response method is used and male superiority when the multiple-choice method is used. A large Gender $\times$ Trait interaction is also evident ($p < .001$). This is also consistent with the data in Table 1 and

Table 2
Repeated measures ANOVA of Academic Achievement (z-scores), by Trait, Method (within subjects), and Gender (between-subjects)

SOURCE	SS	df	MS	F	p
Between subjects	6371.34	1496			
Gender	20.31	1	20.31	4.78	<.05
Error	6351.03	1495	4.25		
Within subjects	2604.66	7480			
Gender x Trait	124.70	2	62.35	137.97	<.001
Error	1351.20	2990	0.45		
Gender x Method	11.16	1	11.16	27.34	<.001
Error	610.23	1495	0.41		
Gender x Trait x Method	1.29	2	0.64	3.80	<.05
Error	506.08	2990	0.17		
Total	8976.00	8976			

Note: Because the achievement measures are in z-score form, the effects of Trait and Method and the Trait x Method interaction are zero by definition and are not included in the table.

reflects the finding that males, regardless of method, perform better than females on mathematics. The three-way interaction, whereby the Method × Gender effect varies by Trait, is also significant ($p < .05$). However, the method-based gender difference is not larger in the case of languages, which we would have expected if the superior verbal ability of girls were the reason for the difference.

Discussion

In this study we sought evidence of a gender difference in scholastic achievement related to the use of multiple-choice and free-response examination methods. We found only slight gender differences on the language measures, whether multiple choice or free response. Differences in favor of boys on both types of mathematics measure were greater, supporting the findings of earlier studies regarding the superiority of males in mathematics (Maccoby & Jacklin, 1974). Our findings also indicate that males perform relatively better than females on multiple-choice forms of assessment, compared with free-response examinations. Conversely, females do relatively better when free-response methods are used. The effect of measurement method is found across the traits (mathematics, Irish, English).

The presence of a method-based gender difference in our findings that holds for

mathematics as well as for languages, and in fact is greater for mathematics, does not support an explanation of the method-gender relationship by the superior verbal ability of girls.

To what then are we to attribute the poorer performance of girls on standardized tests? One possible explanation may be that girls respond less well to novel situations than do boys, a proposition for which there is some supporting evidence (Kimball, 1989). For the explanation to hold in our case, standardized tests should be used less frequently than other forms of testing in schools, and the material or problems on the standardized tests should be more unfamiliar than material and problems on free-response tests. The first condition regarding familiarity is met and so may help account for our findings. Whether or not the second condition is met is more problematic, because the tests were designed to reflect classroom content, and when standardized tests are designed as in the New York State Regents examinations, differences in favor of boys are less likely to occur, though they do not disappear (Smith & Walker, 1988). It may be, of course, that the standardized tests were not entirely successful in reflecting classroom teaching. If teachers teach to anticipated public examinations rather than to the syllabus—and there is every reason to believe that they do (see Madaus & Macnamara, 1970)—then it is possible that sections of the syllabus represented in the standardized tests were not usually included in examinations and so were not emphasized in classroom teaching. If this occurred, then some items on the standardized tests would have been unfamiliar to students, and girls, according to the familiarity hypothesis, would have been at a disadvantage.

Two other possible explanations of our findings warrant consideration. The first arises from the findings of a number of studies that variables unrelated to content may affect the score awarded by examiners to an essay test; such variables include quality of handwriting, expectations of readers, and students' race and gender (Briggs, 1980; Chase, 1986; Hughes, Keeling, & Tuck, 1983). It is of interest that in our study, although the identity of individual students was not known to markers of the free-response examinations, the gender of students was, because of the practice (now discontinued) of using different colored answer books for boys and girls.

A further factor that may have contributed to the gender differences in our findings may have been the greater tendency of males than of females to guess the answers to multiple-choice questions. In a study of guessing behavior on multiple-choice tests, Harris (1971) found that girls were at a disadvantage relative to boys because of their lower propensity to guess; when required to guess the answers to originally omitted items, the gain accruing to girls was greater than that accruing to boys. (In our study, students were not informed of the advantages of guessing in multiple-choice tests.) This interpretation is consistent with the findings of Hanna (1986) that girls, if given the option, make more omission errors than boys on standardized mathematics tests.

Whatever the explanation of our findings, they raise issues for educational policymakers regarding the choice of method of measurement in examinations. This is particularly important if the results of examinations are used, as they are in a number of European countries, including Britain and Ireland, to make

important decisions about a student's educational and vocational future (e.g., admission to third-level educational institutions and a variety of occupations). On the basis of our findings, girls seem to have an advantage over boys when free-response measures are used, whereas the introduction of multiple-choice test items would tend to improve the performance of males relative to females. Further, it appears likely that the introduction of multiple-choice test items would result in changes in the pattern of gender differences in that differences would increase in mathematical subjects (in which the advantage of males would increase) and decrease in verbal subjects (in which the advantage of females would decrease).

References

Ackerman, T. A., & Smith, P. L. (1988). A comparison of the information provided by essay, multiple-choice, and free-response writing tests. *Applied Psychological Measurement, 12,* 117–128.

Briggs, D. (1980). A study of the influence of handwriting upon grades using examination scripts. *Educational Review, 32,* 185–193.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Chase, C. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement, 23,* 33–41.

Dwyer, C. A. (1979). The role of tests and their construction in producing apparent sex-related differences. In M. A. Wittig & A. C. Petersen (Eds.), *Sex-related differences in cognitive functioning. Developmental issues.* New York: Academic Press.

Educational Research Centre. (1978). *Drumcondra Attainment Tests, Level VI, Form A.* Dublin: Educational Research Centre, St. Patrick's College.

Frary, R. B. (1985). Multiple-choice versus free-response: A simulation study. *Journal of Educational Measurement, 22,* 21–31.

Greaney, V., & Kellaghan, T. (1979). School leaving examinations in Ireland. In F. M. Ottobre (Ed.), *Criteria for awarding school leaving certificates. An international discussion.* New York: Pergamon.

Hanna, G. (1986). Sex differences in the mathematics achievement of eighth graders in Ontario. *Journal for Research in Mathematics Education, 17,* 231–237.

Harding, J. (1980). Sex differences in performance in science examinations. In R. Deem (Ed.), *Schooling for women's work.* London: Routledge & Kegan Paul.

Harris, J. W. (1971). *Aspects of the guessing behaviour of young Irish subjects on multiple-choice items.* Unpublished master's thesis, University College, Cork, Ireland.

Hughes, D. C., Keeling, B., & Tuck, B. F. (1983). Effects of achievement expectations and handwriting quality on scoring essays. *Journal of Educational Measurement, 20,* 65–70.

Ireland: Department of Education. (1976). *Rules and programme for secondary schools 1976–77.* Dublin: Stationery Office.

Kimball, M. M. (1989). A new perspective on women's math achievement. *Psychological Bulletin, 105,* 198–214.

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

Madaus, G. F. (1967). A cross-cultural comparison of the factor structure of selected tests of divergent thinking. *Journal of Social Psychology, 73,* 13–21.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.),

Critical issues in curriculum (Eighty-seventh Yearbook of the National Society for the Study of Education, Part 1). Chicago: NSSE.

Madaus, G. F., & Macnamara, J. (1970). *Public examinations. A study of the Irish leaving certificate.* Dublin: Educational Research Centre, St. Patrick's College.

Martin, M. O., & O'Rourke, B. (1984). The validity of the DAT as a measure of scholastic aptitude in Irish post-primary schools. *Irish Journal of Education, 18,* 5–24.

Mathews, J. C. (1985). *Examinations. A commentary.* Boston: Allen & Unwin.

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.

Rowley, G. L. (1974). Which examinees are most favoured by the use of multiple choice tests? *Journal of Educational Measurement, 11,* 15–23.

Searle, S. R. (1971). *Linear models.* New York: Wiley.

Smith, G. M. (1933). Group factors in mental tests similar in material or in structure. *Archives of Psychology, 156,* 1–56.

Smith, S. E., & Walker, W. J. (1988). Sex differences on New York State Regents examinations: Support for the differential course-taking hypothesis. *Journal for Research in Mathematics Education, 19,* 81–85.

Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355–369.

Winer, B. J. (1971). *Statistical principles in experimental design.* New York: McGraw-Hill.

Wood, R. (1976). Sex differences in mathematics attainment at GCE Ordinary level. *Educational Studies, 2,* 141–160.

Wood, R. (1978). Sex differences in answers to English language comprehension items. *Educational Studies, 4,* 157–165.

Authors

NIALL BOLGER is Assistant Professor, Department of Psychology, University of Denver, Denver, CO 80208. *Degrees:* BA, University of Dublin; MS, PhD, Cornell University. *Specializations:* social psychology of health and illness; research methods.

THOMAS KELLAGHAN is Director, Educational Research Centre, St. Patrick's College, Dublin 9, Ireland. *Degrees:* BA, PhD, Queen's University of Belfast. *Specialization:* educational measurement and evaluation.