Gender, Ethnicity, Ability, and Type of Item Response in Mathematics Proficiency of Fifth-Graders

Dimiter M. Dimitrov
Kent State University

The purpose of this study was to examine the interactive role of gender, ethnicity, ability, and type of item response in the performance of fifth-graders on the Ohio Off-Grade Proficiency Tests-Mathematics. While no gender differences were found, the performance profiles across multiple-choice, short response, and extended response items varied by ethnicity and ability level of the students.

Previous studies have reported gender and ethnic differences in mathematics achievement in elementary school (e.g., Doolittle & Cleary, 1987; Cooper & Dorr, 1995). However, very little is known about interactions of gender and ethnicity with student ability and type of item response. Knowledge about these interactions has become more important as a result of the increased use of both multiple-choice and open-ended items in many national and statewide assessment programs. This study examines the impact of gender and ethnicity on mathematics achievement of fifth-graders taking into account student ability and type of item response factors.

Method

Instrument and Subjects

The Ohio Off-Grade Proficiency Test-Mathematics (OOPT-M) for grade five was used (Riverside Publishing, 1995). It includes 30 multiple-choice items on a dichotomous scale (0, 1), 8 short-response items on a partial credit scale (0, 1, 2), and 2 extended-response items on a partial credit scale (0, 1, 2, 3, 4). The subjects were 4830 fifth-graders from a large urban area in North-East Ohio. The sample included 994 Caucasian (477 females and 517 males), 3242 African-American (1684 females and 1558 males), 348 Hispanic (189 females and 159 males), and 246 other students with no gender and/or ethnic group information. For this sample, the reliability of the test was adequate ($\alpha = .83$) to be used for group comparisons. The correlation between the multiple-choice and open-ended sections of the OOPT-M was .70. When corrected for unreliability, this correlation was .95 and remained stable across all gender and ethnic groups.

Procedures

In item response theory, the term "ability" connotes a latent trait that underlies the student's performance on a test (see, e.g., Hambleton et al., 1991, p. 77). The ability score of a student determines his or her probability of giving a correct response to any test item. The units of the ability scale, "logits", typically range from -4 to +4. They represent natural logarithms of odds ratios for success on the test items. If a student succeeds on twice as many items as he or she fails, the odds ratio for the test being $2/1 = 2$, the ability estimate of this student, in logits, is the natural logarithm of 2, which is 0.69. The ability scores of the students were calculated using the program PARSCALE which takes into account the partial credit scores of the students on the short-response and extended-response items (Muraki & Bock, 1996; Muraki, 1992).

Student ability scores were normally distributed within the range from -3.75 to 3.75 on the "logit" scale, with a mean of -0.20, and a standard deviation of 0.92. Students with ability scores in the lower 27% were assigned to the *low ability* group, those with ability scores in the upper 27%, to the *high ability* group, and the rest of the students, to the *medium ability* group. Multivariate analysis of variance (MANOVA) was conducted with three between-subjects factors, Gender, Ethnicity, and Ability, and one within-subjects factor, Type of item response (multiple-choice, short-response, extended-response) using SPSS (SPSS Inc., 1997).

Results

The MANOVA results showed no significant main effect for Gender ($F(3, 4564) = 1.71$, $p > .05$) and no significant interactions between Gender and the other factors, Gender x Ethnicity ($F(6, 9128) = 1.33$, $p > .05$), Gender x Ability ($F(6, 9128) = 0.80$, $p > .05$), and Gender x Type of item response ($F(2, 4565) = 2.40$, $p > .05$). There was a significant main effect for Ethnicity ($F(6, 9128) = 6.17$, $p < .01$). The interaction Ethnicity x Ability was significant on extended-response items ($F(4, 4575) = 6.19$, $p < .01$), but it was not significant on multiple-choice ($F(4,$

Table 1
Means and Standard Deviations of OOPT-M Scores by Ability Level and Ethnicity

	Type of Item Response					
	multiple-choice		short-response		extended-response	
Ethnicity	*M*	*SD*	*M*	*SD*	*M*	*SD*
			Low Ability Level			
Caucasian	11.19	2.67	0.54	0.67	0.53	0.66
African-American	11.18	2.44	0.54	0.70	0.63	0.79
Hispanic	10.70	2.21	0.63	0.73	0.77	0.79
			Medium Ability Level			
Caucasian	15.92	3.05	2.38	1.25	1.94	1.39
African-American	15.38	2.74	2.22	1.38	2.04	1.29
Hispanic	15.17	2.55	2.15	1.38	2.18	1.38
			High Ability Level			
Caucasian	21.09	3.02	6.78	2.60	4.25	1.68
African-American	20.21	3.27	6.49	2.44	3.85	1.62
Hispanic	19.79	3.46	6.75	2.72	4.13	1.65

Note. The maximum possible score is 31 for the multiple-choice, 16 for the short-response, and 8 for the extended-response items.

4575) = 2.34, $p > .05$) or short-response ($F(4, 4575) = 1.13$, $p > .05$) items. The results in Table 1 show that the differences between the average raw scores of any two ethnic groups were less than one point at all ability levels. Because of the significant interactions Ethnicity x Type of item response ($F(4, 9128) = 4.86$, $p < .05$) and Ability x Type of item response ($F(4, 9128) = 50.91$, $p < .05$), the achievement profiles of the ethnic groups across the types of item response were analyzed by ability levels. The post-hoc comparisons were conducted using the Dunett's T3 pairwise comparisons test (SPSS Inc., 1997, p. 37).

For the low ability students, the results in Table 2 show no significant differences between the ethnic groups on the multiple-choice and short-response items. However, significant differences were found between these groups ($F(2, 1231) = 3.06$, $p < .05$) on the extended-response items. The post-hoc test showed that Hispanic students performed higher than Caucasian students ($p < .05$) on the extended response items. For all ethnic groups, the average z-scores on the short-response and extended response items were equal ($F(1, 1234) = 0.06$, $p > .05$) and higher than the average z-score on the multiple-choice items ($F(1, 1234) = 45.03$, $p < .01$).

For the medium ability students, the results in Table 2 show no significant differences between the ethnic groups on the extended-response and short-response items. However, significant differences were found between these groups ($F(2, 1648) = 7.33$, $p < .01$) on the multiple-choice items. Caucasian students performed higher than African-American ($p < .01$) and Hispanic ($p < .05$) students. The average z-score on the short-response items was lower than this on multiple-choice ($F(1,2118) = 73.80$, $p < .01$) and extended-response ($F(1, 2118) = 35.54$, $p < .01$) items.

For the high ability students, the results in Table 2 show no significant differences between the ethnic groups on the short-response items. However, significant differences were found between these groups ($F(2, 1231) = 11.91$, $p < .01$) on the multiple choice items. Caucasian students performed higher than African-American ($p < .01$) and Hispanic students ($p < .01$) on the multiple-choice items. Also, significant differences were found between the ethnic groups ($F(2, 1231) = 7.94$, $p < .01$) on the extended-response items. Caucasians did better than African-Americans ($p < .01$) on the extended-response items. The average z-score on short-response items was higher than this on the multiple-choice ($F(1, 1234) = 50.92$, $p < .01$) and extended-response ($F(1, 1234) = 30.71$, $p < .01$) items.

Discussion

The results of this study showed that gender did not play a significant role in the performance of fifth-graders on the Ohio Off-Grade Proficiency Test-Mathematics. This finding is consistent with the conclusions of previous

Table 2
Multivariate Analysis for Ethnic Group Differences onOOPT-M Scores by Student Ability and Type of Item Response

Source	*df*	*F*		
		multiple-choice	short-response	extended-response
		Low Ability Level		
Ethnicity	2	1.86	0.79	3.06*
Error	1231	(0.30)	(0.06)	(0.18)
		Medium Ability Level		
Ethnicity	2	7.33**	2.76	1.99
Error	1648	(0.38)	(0.23)	(0.53)
		High Ability Level		
Ethnicity	2	11.91**	1.84	7.94**
Error	1231	(0.51)	(0.78)	(0.82)

Note. Values enclosed in parentheses represent mean square errors.
*$p < .05$. **$p < .01$

studies with fifth-graders (e.g., Lewis & Hoover, 1986) and high-school students (e.g., DeMars, 1998; O'Neil & Brown, 1998).

Significant differences were found between the mathematics performance of Caucasian, African-American, and Hispanic students. However, comparisons between ethnic groups across different ability levels and types of item response did not consistently favor one group over another. The difference between any two ethnic groups was less than one point in raw scores. Caucasian and Hispanic students did not differ except that high ability Caucasians did better on the multiple-choice items. This finding is consistent with the results reported by O'Neil and Brown (1998) indicating that, for eight graders from California schools, Caucasians performed better than Hispanics on multiple-choice items, but there was no difference between these two ethnic groups on open-ended items. At high ability level, Caucasians did better than African-Americans on multiple-choice and extended-response items. At medium ability level, Caucasians did better than African-Americans on multiple-choice items. There was no significant difference between Hispanic and African-American students regardless of ability and item response format.

Although the OOPT-M multiple-choice and open-ended items measure positively correlated characteristics of student ability, the results show that types of item response have differential effect on student performance. Regardless of ethnicity, student performance varied across multiple-choice, short-response, and extended-response items. One plausible explanation comes from previous research indicating that students employ different lines of reasoning in dealing with multiple-choice and open-ended items (Frederiksen, 1994). Also, open-ended items induce more cognitive strategy and worry than multiple-choice items (O'Neil & Brown, 1998). In addition, this study finds that types of item response operate differentially among students from different ethnic groups and ability levels. For example, Hispanic students performed as well or better on the extended-response items and lower on the multiple-choice items. This suggests that the mathematics performance of Hispanic students was not negatively influenced by their level of proficiency with the English language. Also, Hispanic students may need help improving their response strategies on multiple-choice items. Thus, if mathematics teachers want their instruction to be effective and responsive to proficiency testing, they should not trust common perceptions such as (a) students perform lower on open-ended items, or (b) multiple-choice items capture lower level skills. Research is needed to identify other factors (e.g., learning style, motivation, curriculum, environment, culture) to explain mathematics proficiency profiles across lines of ethnicity, ability, and type of item response.

References

Cooper, H., & Dorr, N. (1995). Race comparison on need for achievement: a meta-analytic alternative to Graham's narrative review. *Review of Educational Research, 65,* 483-508.

DeMars, C. E. (1998). Gender differences in

mathematics and science on a high school proficiency exam: the role of response format. *Applied Measurement in Education, 11*, 279-299.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Lewis, J. C., & Hoover, H. D. (1983, April). *Sex differences on standardized academic achievement tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. & Bock, R. D. (1996). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks [Computer program]. Chicago, IL: Scientific

O'Neil, H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*, 331-351.

Riverside Publishing (1997). *Ohio Off-Grade Proficiency Tests: specifically designed to measure Ohio's model course of study*. Chicago, IL: Author.

SPSS Inc. (1997). *SPSS (Windows version 7.5): User's guide*. Chicago, IL: Author.

Dimiter M. Dimitrov is an Assistant Professor of Evaluation and Measurement at the College of Education, Kent State University, Kent, Ohio.

mathematics and science on a high school proficiency exam: the role of response format. *Applied Measurement in Education, 11*, 279-299.

Doolittle, A. E., & Cleary, T. A. (1987). Gender-based differential item performance in mathematics achievement items. *Journal of Educational Measurement, 24*, 157-166.

Frederiksen, N. (1984). The real test bias: influences of testing on teaching and learning. *American Psychologist, 39*, 193-202.

Lewis, J. C., & Hoover, H. D. (1983, April). *Sex differences on standardized academic achievement tests.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Muraki, E. (1992). A generalized partial credit model: Applications of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muraki, E. & Bock, R. D. (1996). PARSCALE: IRT based test scoring and item analysis for graded open-ended exercises and performance tasks [Computer program]. Chicago, IL: Scientific

O'Neil, H. F., & Brown, R. S. (1998). Differential effects of question formats in math assessment on metacognition and affect. *Applied Measurement in Education, 11*, 331-351.

Riverside Publishing (1997). *Ohio Off-Grade Proficiency Tests: specifically designed to measure Ohio's model course of study.* Chicago, IL: Author.

SPSS Inc. (1997). *SPSS (Windows version 7.5): User's guide.* Chicago, IL: Author.

Dimiter M. Dimitrov is an Assistant Professor of Evaluation and Measurement at the College of Education, Kent State University, Kent, Ohio.