Effects of Test Administrator Characteristics on Achievement Test Scores[1]

William D. Schafer
University of Maryland

Maria Papapolydorou
Ministry of Education and Culture, Cyprus

Taslima Rahman
U. S. Department of Education

Lori Parker
Towson University

Presented at the April, 2005 Conference of the National Council on Measurement in Education in Montreal, Canada.

[1]This research was carried out by the Maryland Assessment Research Center for Education Success with funding provided through the Maryland State Department of Education. The opinions are solely those of the authors.

Effects of Test Administrator Characteristics on Achievement Test Scores

Abstract

Possible relationships between five test examiner characteristics (gender, race, tenure, experience as a test administrator, and experience as a test developer or scorer) and six student achievement scores (reading, writing, language usage, mathematics, science, and social studies) were studied at the school level in a statewide assessment. The school-level results were aggregated using meta-analysis to explore the plausibility of examiner variables as threats to test validity. Very few of the average correlations across schools were statistically significant, and for all of them, even for those that were statistically significant, confidence intervals for the correlations were extremely small at both ends. Significant heterogeneity of effect sizes was found for virtually all of the 60 analyses, suggesting that further exploration is needed. Some directions for further research are discussed.

Effects of Test Administrator Characteristics on Achievement Test Scores

Departures from prescribed test administration procedures may result in bias and thus affect test validity. According to the 1999 Standards (AERA, APA, NCME, 1999), standardization of test administration involves "maintaining a constant testing environment and conducting the test according to detailed rules and specifications, so that testing conditions are the same for all test takers" (p. 182). In order to ensure valid and reliable scores, administrators must deliver the testing material uniformly (e.g. assemble and present materials as instructed in the test manual, avoid idiosyncratic verbal elaboration of written directions).

Several mechanisms have been cited for how unstandardized conditions may arise. Unintended cues can be given inadvertently, such as by facial expressions or words of encouragement (Franco & LeVine, 1985; Cronbach, 1970). The way a test administrator talks and gestures can encourage or discourage a student and the examiner may respond to a question with cues for a particular answer. Rereading directions or offering explanations not given in testing materials may assist students' understanding, but can create inconsistency because students in other testing groups may not receive the same assistance. Bishop and Frisbie (1999) found significant differences in both test scores and students' work rates when test administration strategies differed among administrators.

Information surrounding a test may also have an impact on validity. Unless examinees are given a personally relevant reason for taking a test, data collected can have uncertain meaning (Cronbach, 1970). Responses may be casual, or examinees may even fake results, such as by trying to miss items or trying to respond in an arbitrary direction. Cronbach (1970) suggested that when the examiner increases the examinee's motivation to do well, test scores improve but creating test anxiety or communicating an expectation of failure can result in lower test scores. Additionally, students' test scores improve when the stakes of the test results are raised (Kiplinger & Linn, 1993).

There appears to be reasonable evidence that test administrators can impact students' performance during standardized test administration, but what attributes of administrators are relevant? Research in the area of teacher characteristics tends to focus on similarities or differences between demographic characteristics of the test administrator and test taker (Argeton & Moran, 1995; Franco & LeVine, 1985). Fuchs and Fuchs (1986) found that differential performance favoring a familiar examiner could become greater depending on student's socioeconomic status, difficulty of the test or degree of familiarity between examiner and examinees.

It seems reasonable to hypothesize that test administrators who express positive feelings toward the test will project that attitude during test administration, particularly during pre-assessment activities that involve group participation. Conversely, a negative attitude towards specific aspects of the test, such as inadequate time for materials preparation or directions perceived as unclear could alter the administration. Some administrators may feel motivated to offer more time to students not able to complete tasks in the allotted time. Other variables that might affect manner of administration include attitudes towards standardized tests and school-level accountability, level of teaching experience, and familiarity of the administrator with the assessment.

However, the extent of relationships between even easily-identified administrator characteristics and test performance has not been studied as it occurs in actual practice in a large-scale test administration. This study was conducted to address this need in the context of a standardized performance assessment administered statewide. A performance assessment could be an ideal vehicle to show administrator effects if they exist since it entails significantly more interaction between administrators and examinees than do traditional standardized tests. Indeed, as noted later (see the Achievement Measures section, below), the possibility of influence of administrator demographics on the particular performance assessment used in this research has been raised in an independent review.

Method

Participants

The data were gathered as part of a regular statewide assessment program given in grades 3, 5, and 8 (only grades 3 and 5 were used here) in April, 2002. Students were assessed in six content areas: reading, writing, language usage, math, science, and social studies. The test administrators were teachers in the students' schools. Assessments were completed in test groups separately, each taking one of three unique forms. Students were assigned to test groups by a quasi-random process (using position in an alphabetic list) within schools; administrators and forms were randomly assigned to the test groups.

Achievement Measures: The Statewide Performance Assessment

The 2002 Maryland School Performance Assessment Program (MSPAP) was administered during the eleventh (and final) year of a statewide testing program designed to measure school effectiveness by assessing higher-order thinking processes in contexts that demanded integrated applications of students' content knowledge. It was a performance-based assessment that required students to produce individual, written constructed responses to items, each presented as smaller elements of larger tasks designed to elicit a variety of both brief and extended responses based on criteria set forth in the Maryland Learning

Outcomes (http://www.mdk12/mspap.org). There were six achievement scores for each examinee: reading (RD), writing (WT), language usage (LU), mathematics (MA), science (SC), and social studies (SS).

Compared with other standardized tests, the nature of MSPAP allowed considerable opportunity for variation among administrators. The tasks were complex and often required that materials be assembled prior to the test. Many of the tasks used these or other forms of administrator-dependent, pre-assessment activities, which were intended to acquaint students with information required for them to demonstrate their proficiency on the scored portions of the tasks. For example, the test administrator might have been required to pre-assemble some of the materials needed for a science task and then perform the experiment as a demonstration during the course of the administration. In some tasks, administrators led student discussions or other activities that were intended to convey understandings that students would then be expected to apply to the items. Perhaps inhibited by the time constraints, background knowledge, or motivation, teachers may not have become equivalently expert with the test manual and materials. Other variation in administrations may have naturally resulted from stylistic differences among administrations during pre-assessment or actual assessment activities.

Since MSPAP was a statewide performance assessment, standardization was crucial to its validity. Several steps were taken each year to ensure standardization. Administration manuals were field-tested and revised. Teacher-administrators had two weeks with the actual test materials to prepare for the testing. Officials from the Maryland State Department of Education held training workshops for district-level Local Accountability Coordinators who, in turn, provided training and materials to the School Test Coordinators. School Test Coordinators then trained the classroom teachers who administered the MSPAP.

As in earlier years, three non-overlapping forms of MSPAP were used in 2002. Since there were extensive pre-assessment activities that were unique to forms, each form was administered to students in a given test group in a self-contained room. Before MSPAP administration, students were randomly assigned to test groups and teachers (test administrators) were randomly assigned to the room in which they would administer the test, such that a student's regular classroom teacher may or may not have been his or her test administrator. Test forms were also assigned to rooms randomly (in larger schools, there were often more than three rooms; care was taken to make sure there were at least three forms in even very small schools, whenever possible). The contracted test-development company provided a linking between the three forms of MSPAP each year to put the scores on equivalent scales. Procedures for test construction, administration, and analysis are described in the technical manual (available at mdk12.org).

MSPAP administration took place over five days, with a 90-105 minute test block each day. Students worked on two or three tasks per day; some tasks were completed in one day, while others stretched across test blocks on multiple days.

The results of MSPAP had a direct effect on the school. Data resulting from the program were published yearly; rewards and sanctions, including possible state takeover, existed. Schools were rated on their achievement, both statewide and relative to other schools in their districts by the media.

In their psychometric review of MSPAP, Hambleton, Impara, Mehrens, & Plake (2000) raised a concern that is directly related to the motivation for this study. They questioned the validity of MSPAP for school-level results.

> Literature has shown that teacher familiarity with the tasks of an assessment is an important factor in student performance, and is likely to be even more critical for MSPAP because it has tasks that are novel and complex. Thus, the impact on school performance of higher teacher turnover rates in poorer schools will likely be greater for MSPAP than it would be for assessments composed of multiple-choice questions. (p. 26)

Hambleton et al. (2000) were clearly concerned that an assessment such as MSPAP may be seriously impacted by test examiner characteristics, more so than assessments using other formats.

Administrator Characteristics: The Survey

For the 2002 administration, School Test Coordinators (not the test administrators) in schools with third and/or fifth graders were asked to complete a survey about the teachers who were assigned to each testing group. They were asked to report each teacher-examiner's gender (male or female), ethnicity (white or non-white), tenure status (yes or no), experience with MSPAP as a writer or scorer (yes or no), and experience as an administrator of MSPAP (number of times: 0, 1, 2, or 3 or more). Additionally, they noted whether the administrator was present for all five days of testing. Survey forms were returned to the State Department of Education rather than the testing or scoring contractors. School Test Coordinators knew that the questionnaire was part of a special study and that return of the form was optional.

Procedures

The study was conceived as a replicated field study (Schafer, 2001) and analyzed using meta-analysis for each grade-content combination, which provided independent (across schools) correlations. Schools with test administrators not present for all five test days were removed from the analysis. Some administrator characteristics were constant across all test groups in some schools and those school-level correlations could not be computed. In all, there were 4,669 useable correlations for the meta-analyses.

Results

There were 60 sets of correlations (six content domains by five administrator characteristics by two grade levels). The number of schools that contributed useable data ranged from a low of 51 (for correlations between examiner ethnicity and LU) to a high of 160 (for correlations between prior examiner experience and four of the achievement variables). Tables 1 through 10 include the number of schools for each of the 60 sets of correlations along with the minimum, average, and maximum numbers of students providing useable data across the schools in each set.

Each of the 60 sets of correlations was analyzed separately in the same way. Within each set, the meta-analytic study unit was the school and thus the correlations in each set were independent. Following the meta-analysis procedures described by Hedges & Olkin (1985), each school's correlation (its effect size) was transformed using Fisher's r-to-z transformation $\{Z_r=0.5*\log_e[(1+r)/(1-r)]\}$and each transformed correlation was weighted by the inverse of its sampling variance; i.e., each was weighted by (n-3) where n was the number of examinees in that school. The weighted mean of the transformed effect sizes is the overall transformed effect size estimate for the set of correlations and its significance from zero may be tested with a one-degree-of-freedom chi-square that is equal to the square of the weighted sum of Z_r divided by the sum of the weights:

$$\overline{Z_r} = \frac{\sum_{i=1}^{k} w_i Z_{r_i}}{\sum_{i=1}^{k} w_i}$$

$$x_1^2 = \frac{(\sum_{i=1}^{k} w_i Z_{r_i})^2}{\sum_{i=1}^{k} w_i}$$

where Z_{r_i} is the Fisher transform of the correlation in the i^th^ school

w_i is the weight of the transformed correlation in the i^th^ school [$w_i = (n_i - 3]$

k is the number of schools in the set of correlations being analyzed.

The standard error of the overall effect size is the square root of the reciprocal of the sum of the weights:

$$S_{\overline{Z}_r} = \sqrt{1 / \sum_{i=1}^{k} w_i}$$

An advantageous feature of meta-analysis is that the homogeneity of the effect sizes may be tested for significance using a chi-square with degrees-of-freedom equal to one minus the number of schools in the set. The chi-square is the weighted sum of the squared effect sizes minus the square of the sum of the weighted effect sizes divided by the sum of the weights:

$$Q_E = \sum_{i=1}^{k} w_i Z_i^2 - \frac{(\sum_{i=1}^{k} w_i Z_i)^2}{\sum_{i=1}^{k} w_i}$$

The Q_E statistic is distributed as a chi-square with *df* = k-1. If the chi-square is statistically significant, the interpretation is that the effect sizes are not homogeneous and therefore there exists variation to be explained (i.e., there are characteristics of the schools that affect the correlations). In all, 49 of the 60 sets resulted in statistically significant heterogeneity. However, we had no more information about the schools.

If all explanatory variables have been exhausted, Hedges & Vevea (1998) recommend treating the between-study effects as random variables, where each is a sample from its own distribution. This analysis can be accomplished by adding a term to the sampling variance for each effect size and re-running the analysis. The term to be added is the larger of either zero or a fraction whose numerator is the heterogeneity chi-square minus one less than the number of schools and whose denominator is the sum of the original weights minus the ratio of the sum of the squared original weights divided by the sum of the original weights:

$$\hat{t}^2 = \frac{Q_E - (k-1)}{\sum_{i=1}^{k} w_i - \left[\sum_{i=1}^{k} w_i^2 \Big/ \sum_{i=1}^{k} w_i \right]}$$

The sampling variance of each Z_r was then augmented by $\hat{\tau}^2$ so that the new weights became

$$w_i^* = \frac{1}{\frac{1}{n_i - 3} + \hat{\tau}^2}$$

and the new weights were substituted for the old in the fixed-effects analyses and the analyses re-run.

For consistency, we applied the random-effects approach to all 60 sets of correlations. All the results that we report are from the analyses that treated the between-school effects as random.

Confidence intervals were obtained by adding and subtracting 1.96 standard errors about the average effect size. All effect sizes were then converted back to the correlation metric using the Fisher z-to-r transformation to obtain the results that are presented.

Table 1 and Figure 1 present the results of examinee gender for third grade students. Gender was coded such that a positive correlation indicates greater scores for female examiners. The order of the achievement variable presentation is from larger to smaller average effect sizes. That none of the average effect sizes reached statistical significance is evident from noting that all of the confidence intervals include zero. But this finding is nevertheless interesting since the overall effect of examiner gender is estimated very closely; the first decimal place is zero at both ends of the confidence interval.

Table 1. Third Grade Effect Sizes for Examiner Gender*.

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Reading	52	23	73.08	153	0.037	-0.005	0.079
Mathematics	52	25	80.75	159	0.028	-0.015	0.071
Social Studies	52	25	80.35	160	0.013	-0.024	0.050
Science	52	25	79.69	158	0.011	-0.037	0.059
Language Usage	52	24	75.23	153	0.003	-0.030	0.037
Writing	52	25	79.67	157	0.002	-0.041	0.044

*Gender is coded 0 for male and 1 for female.

<u>Insert Figure 1 about Here</u>

The results for examiner gender at the fifth grade are presented in Table 2 and Figure 2. As before, none of the average effect sizes was statistically significant and both ends of the confidence interval indicate that even if the correlation is non-zero, it is of trivial magnitude.

Table 2. Fifth Grade Effect Sizes for Examiner Gender*.

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Science	98	16	78.62	197	0.011	-0.020	0.042
Mathematics	98	16	80.98	204	0.003	-0.028	0.034
Language Usage	98	16	77.22	198	-0.008	-0.037	0.021
Reading	98	16	74.14	197	-0.009	-0.037	0.020
Social Studies	98	16	80.85	200	-0.009	-0.037	0.018
Writing	98	16	80.04	203	-0.013	-0.044	0.019

*Gender is coded 0 for male and 1 for female.

Insert Figure 2 about Here

Taken together, the results for examiner gender suggest that any effect on student test scores at either the third or fifth grade levels is close to zero, if it exists at all. It appears that examiners may be of either gender without much impact on student achievement results.

The results for examiner race are similar to those for examiner gender. Table 3 and Figure 3 present the third-grade data and Table 4 and Figure 4 the fifth-grade. In all cases the confidence intervals included zero and all first decimal places were zero at both ends of the intervals. Again, examiner race seems to have trivial effects if there are any at all at both grade levels.

Table 3.Third Grade Effect Sizes for Examiner Ethnicity*

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Reading	63	22	73.29	171	0.003	-0.026	0.033
Writing	64	26	79.84	178	-0.003	-0.037	0.030
Social Studies	64	26	81.00	179	-0.009	-0.041	0.022
Science	64	24	79.89	176	-0.011	-0.042	0.020
Language Usage	62	24	75.48	173	-0.018	-0.049	0.013
Mathematics	64	26	81.28	179	-0.023	-0.052	0.007

*Ethnicity is coded 0 for White and 1 for Non-White.

Insert Figure 3 about Here

Table 4. Fifth Grade Effect Sizes for Examiner Ethnicity*

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Language Usage	51	29	72.20	198	0.025	-0.011	0.061
Science	52	30	76.46	197	0.021	-0.025	0.067
Writing	52	29	77.88	203	0.015	-0.025	0.056
Reading	53	29	70.09	197	0.010	-0.032	0.053
Mathematics	52	31	78.88	204	0.004	-0.036	0.044
Social Studies	52	31	78.69	200	-0.008	-0.047	0.031

*Ethnicity is coded 0 for White and 1 for Non-White.

Insert Figure 4 about Here

Table 5 and Figure 5 present the third-grade results for examiner tenure. The data are coded so that a positive correlation indicates that tenured examiners were associated with higher test scores. At the third grade, there were four of the six achievement variables for which a statistically significant correlation was observed. Only for writing and language usage did the confidence intervals span zero. However, even where the average correlation was significantly different from zero, the first decimal place of both ends of the confidence interval was zero.

Table 5. Third Grade Effect Sizes for Examiner Tenure*

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Reading	99	10	75.45	171	0.040	0.009	0.070
Social Studies	100	11	82.25	179	0.039	0.009	0.068
Science	100	11	81.29	176	0.038	0.007	0.068
Mathematics	100	11	82.50	179	0.037	0.015	0.059
Writing	100	11	81.35	178	0.028	-0.006	0.062
Language Usage	99	11	77.19	173	0.022	-0.005	0.049

*Tenure is coded 0 for not tenured and 1 for tenured.

<u>Insert Figure 5 about Here</u>

Table 6 and Figure 6 display the results for examiner tenure for the fifth grade. A positive correlation would indicate that higher student scores are associated with examiners who were tenured. As for examiner gender and race, all average effect sizes were non-significant and the first decimal place was zero at both ends of all confidence intervals.

Table 6. Fifth Grade Effect Sizes for Examiner Tenure*

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Writing	94	16	80.45	203.00	0.020	-0.014	0.054
Science	94	16	79.09	197.00	0.015	-0.020	0.050
Social Studies	94	16	81.37	200.00	0.007	-0.024	0.038
Mathematics	94	16	81.54	204.00	0.005	-0.033	0.042
Reading	91	16	74.47	197.00	-0.004	-0.033	0.025
Language Usage	92	16	75.65	198.00	-0.005	-0.035	0.026

*Tenure is coded 0 for not tenured and 1 for tenured.

Insert Figure 6 about Here

The results for examiner experience at the third grade are displayed in Table 7 and Figure 7. The data are coded so that the correlation is positive when examiners with more experience are associated with larger student scores. Although two of the confidence intervals appear to include zero, in fact the lower end of each was negative, but rounded to zero.

Table 7. Third Grade Effect Sizes for Prior Examiner Experience*.

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Language Usage	159	11	71.03	173	0.025	-0.003	0.052
Science	160	11	75.24	176	0.023	0.000	0.047
Mathematics	160	11	76.31	179	0.023	-0.001	0.048
Reading	159	10	69.43	173	0.023	0.000	0.045
Writing	160	11	75.33	178	0.023	-0.004	0.049
Social Studies	160	11	76.04	179	0.018	-0.004	0.040

*Prior Examiner Experience is coded 0 for none, 1 for once, 2 for twice, 3 for three or more times.

Insert Figure 7 about Here

Table 8 and Figure 8 display the fifth grade results for examiner experience. There was one achievement variable for which a statistically significant average correlation existed but again the first decimal place of both ends of all confidence intervals was zero.

Table 8. Fifth Grade Effect Sizes for Prior Examiner Experience*.

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Writing	132	16	78.06	203	0.027	0.001	0.053
Science	132	16	76.67	197	0.023	-0.004	0.051
Social Studies	132	16	78.86	200	0.013	-0.014	0.039
Reading	134	16	71.35	197	0.007	-0.017	0.030
Mathematics	132	16	79.05	204	0.003	-0.021	0.027
Language Usage	132	16	72.79	198	0.002	-0.021	0.026

*Prior Examiner Experience is coded 0 for none, 1 for once, 2 for twice, 3 for three or more times.

<u>Insert Figure 8 about Here</u>

Tables 9 and 10 and Figures 9 and 10 present the results for examiner involvement in MSPAP. The third grade results in Table 9 and Figure 9 and the fifth grade results in Table 10 and Figure 10 both show non-significant average effect sizes except for third-grade reading and a first decimal place of zero at both ends of the confidence intervals.

Table 9. Third Grade Effect Sizes for Examiner Involvement*.

	number of schools	minimum n	averag e n	maximu m n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Reading	49	10	71.98	171	0.042	0.003	0.081
Social Studies	49	14	78.55	179	0.030	-0.014	0.073
Writing	49	12	77.90	178	0.025	-0.015	0.065
Science	49	13	77.82	176	0.021	-0.021	0.064
Mathematics	49	14	78.82	179	0.016	-0.030	0.062
Language Usage	48	24	75.08	173	-0.020	-0.060	0.020

*Involvement is coded 0 for no involvement and 1 for involvement.

Insert Figure 9 about Here

Table 10. Fifth Grade Effect Sizes for Examiner Involvement*.

	number of schools	minimum n	average n	maximum n	mean effect size (r)	r conf. int. lower limit	r conf. int. upper limit
Language Usage	50	18	77.68	198	0.019	-0.015	0.054
Writing	50	18	83.04	203	0.019	-0.019	0.058
Mathematics	50	18	83.94	204	0.012	-0.032	0.056
Social Studies	50	18	83.64	200	0.001	-0.036	0.038
Science	50	17	81.36	197	-0.008	-0.044	0.028
Reading	50	17	76.60	197	-0.028	-0.065	0.010

*Involvement is coded 0 for no involvement and 1 for involvement.

Insert Figure 10 about Here

There were two ways that examiners may have been involved. One is in the development of MSPAP tasks and items and the other is in the operational scoring of MSPAP. In either case, examiners who were more involved may be expected to have a better understanding of the assessment than those who were not. But this did not seem to translate into higher student scores.

Discussion

The results of this study should be reassuring to test specialists. No trend was found that threatens validity for either demographic teacher characteristics (gender, race), a surrogate for teaching experience (tenure) or test familiarity (experience as an administrator or involvement as either a test developer or scorer). With the exception of tenure at the fifth grade on four achievement variables (reading, social studies, science, mathematics), involvement at the third grade for reading, and prior experience on two achievement variables at the third grade (science, reading) and one at the fifth grade (writing), all average effect sizes were not significantly different from zero. The finding of eight out of 60 results represents a 13% rate of statistical significance at the $\alpha = 0.05$ level.

These results appear to rebut the concerns raised about MSPAP by Hambleton et al. (2000). Normally, one would not be able to conclude much from statistically non-significant findings. In this case, though, the narrow range of the 95% confidence intervals for the average effect sizes suggests that only trivial, if any, overall effects exist for all these administrator characteristics. Even those 13% that were statistically significant were estimated to be so small as to be virtually meaningless as threats to test validity.

This study examined the effects of examiner characteristics at a global level. It should be noted that significant heterogeneity of effects were noted in almost all of the 60 data series studied. This implies that more remains to be learned about the effects of examiner characteristics. One approach to addressing that result would be to examine the effects of examiner characteristics on examinees separated by their characteristics. For example, it may be that the effects of gender or race are greater for students of like gender or race. It may also be that the effects of tenure, experience, or involvement are greater at certain achievement levels than others.

Another dimension for further study is the type of test. Perhaps more objective tests will show greater homogeneity than constructed-response tests such as MSPAP. Tests of non-achievement constructs may also show different results.

At the examiner level, this study may be extended by focusing on other characteristics than those studied here. Another approach could be to study mechanisms by which student scores may be affected (e.g., whether examiners look at student responses as they complete a test) and to manipulate these across examiners.

Several characteristics make this data set unique and important. The use of performance assessments with written, constructed student responses may allow a maximal opportunity for administrator effects to appear in what is nevertheless a large-scale, standardized testing context. The (quasi) random assignment of students, forms, and administrators to test groups administered in separate rooms guards against several internal validity threats that are normally present in field research settings. Random replications of within-school correlations should balance out confounding effects of test forms with administrator characteristics. The data are real, coming from an actual, high-stakes-for-schools statewide testing program. Finally, the samples are large enough for fairly precise estimation of effect sizes through the use of meta-analysis for the replicated field study design.

Use of meta-analysis allows the power of large sample sizes to be represented in the standard error of the average effect size and the narrowness of the confidence interval. Note that the correlation (effect size) in each data series has been estimated to be zero in the first decimal place at both ends of the 95% confidence interval. This result is useful since it implies that assessments may be given without fear that examiner characteristics will affect student scores appreciably. Given the nature of MSPAP, any effects that these examiner characteristics may have should have become apparent, and they did not.

The results presented here are useful findings for the assessment community since they suggest that the administrator characteristics studied have little effect on test scores.

Continuing the common practice of choosing examiners without regard to their gender, race, experience, or other, similar variables seems at least partially justified.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association:

Argeton, E., & Moran, M. (1995). Effects of race and dialect of examiner on language: Samples elicited from southern African-American preschoolers. *Journal of Childhood Communication Disorders,* 16 (2), 21-25.

Bishop, N., & Frisbie, D. (1999). *The effects of different test-taking conditions on reading comprehension test performance*. Iowa City: University of Iowa

Cronbach, L. (1970). *Essentials of Psychological Testing*. New York, Evanston, and London: Harper & Row, Publishers.

Franco, J., & LeVine, E. (1985). Effects of examiner variables on reported self-disclosure: implications for group personality testing. *Hispanic Journal of Behavioral Sciences, 7,* (2), 199-210.

Fuchs, D., & Fuchs, L. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *Review of Educational Research*, 56 (2), 243-262.

Hambleton, R. K., Impara, J., Mehrens, W., & Plake, B. S. (2000). *Psychometric review of theMaryland School Performance Assessment Program (MSPAP).* Baltimore, MD: Maryland State Department of Education.

Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Hedges, L. V. & Vevea, J. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*(4), 486-504.

Kiplinger, V., & Linn, R. (1993). *Raising the stakes of test administration: The impact of student performance on NAEP*. Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Schafer, W. D. (2001). Replication: A design principle for field research. *Practical Assessment*, *Research, & Evaluation*, *7*(15), available electronically at pareonline.net.

Figure 1. Third Grade Effect Sizes for Examiner Gender

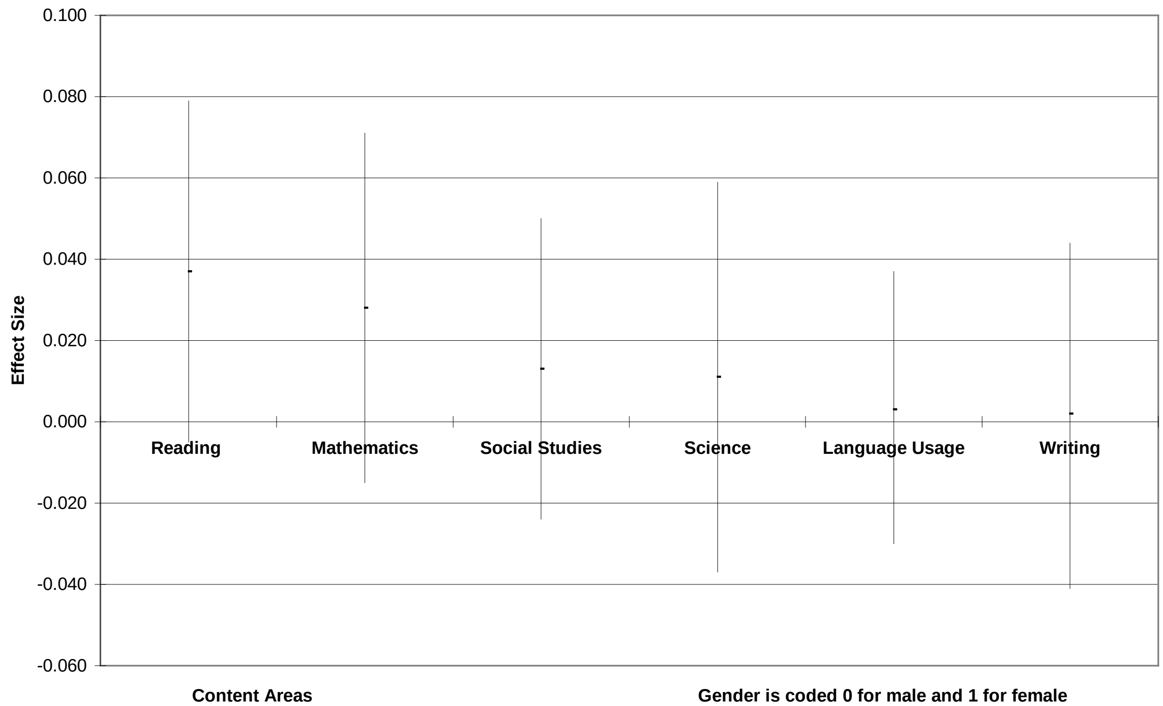


Figure 2. Fifth Grade Effect Sizes for Examiner Gender

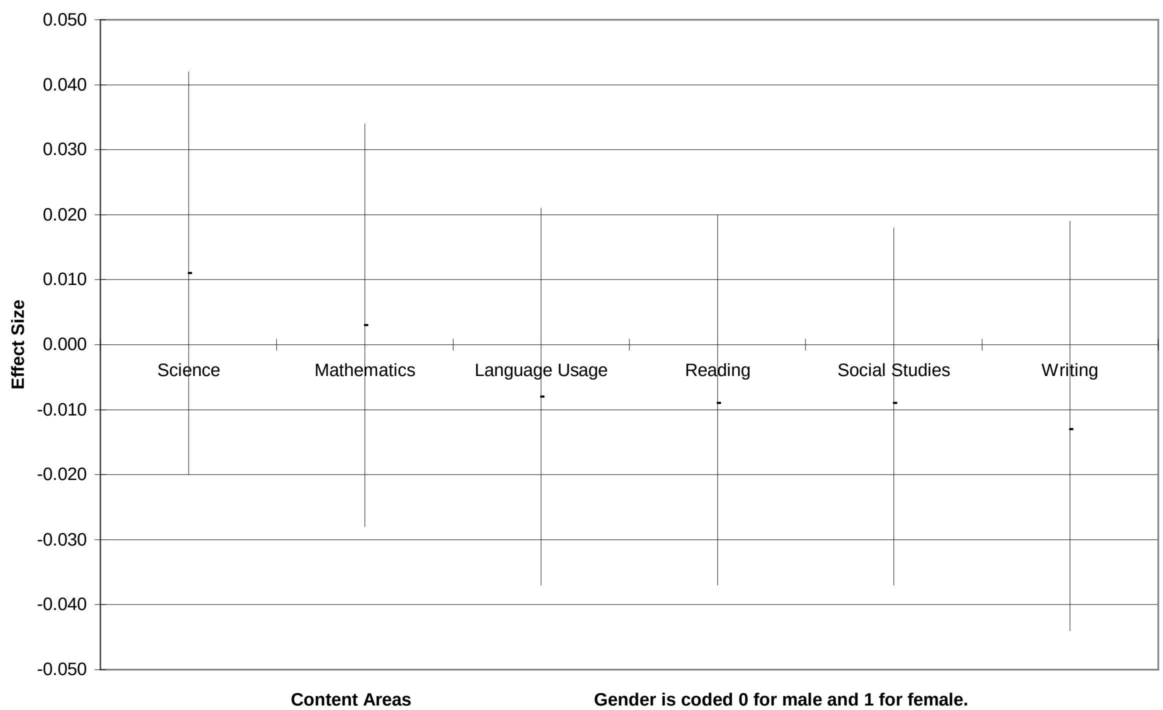


Figure 3. Third Grade Effect Sizes for Examiner Ethnicity

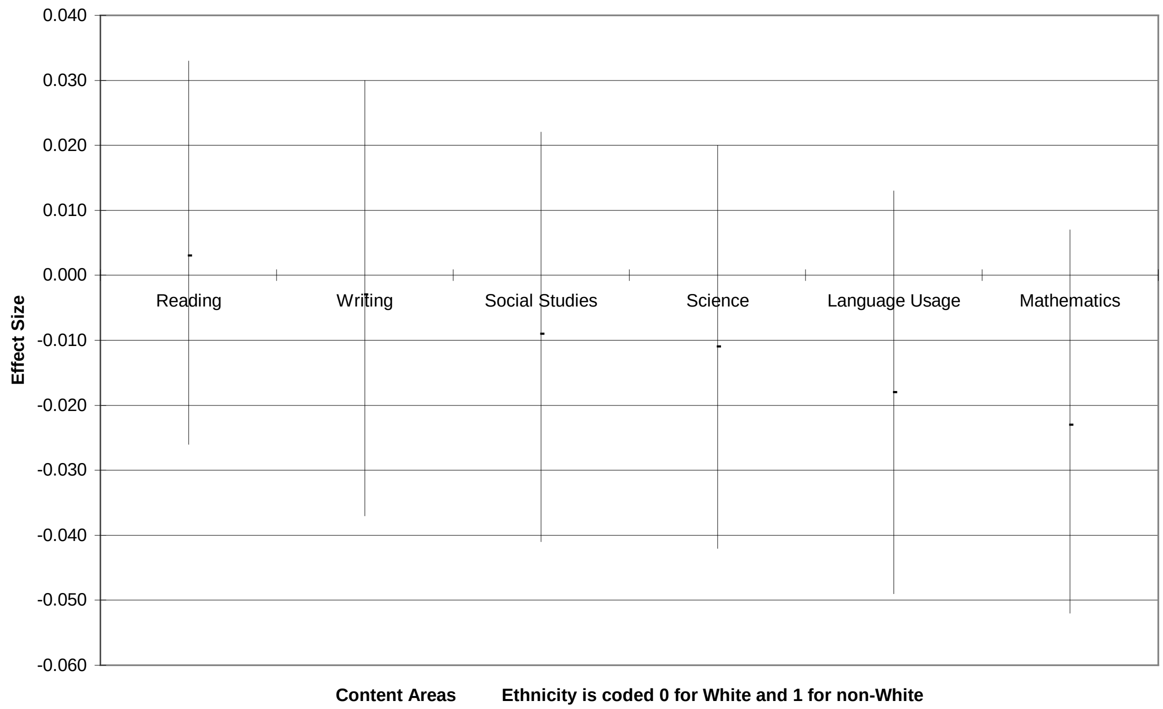


Figure 4. Fifth Grade Effect Sizes for Examiner Ethnicity

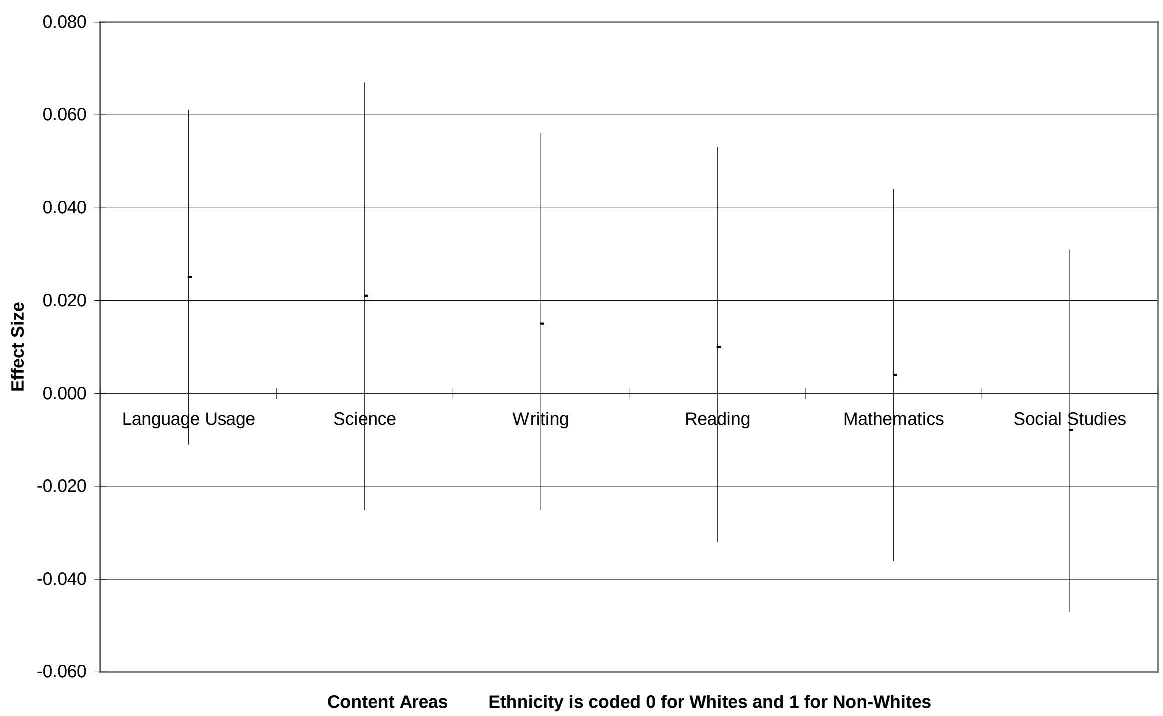
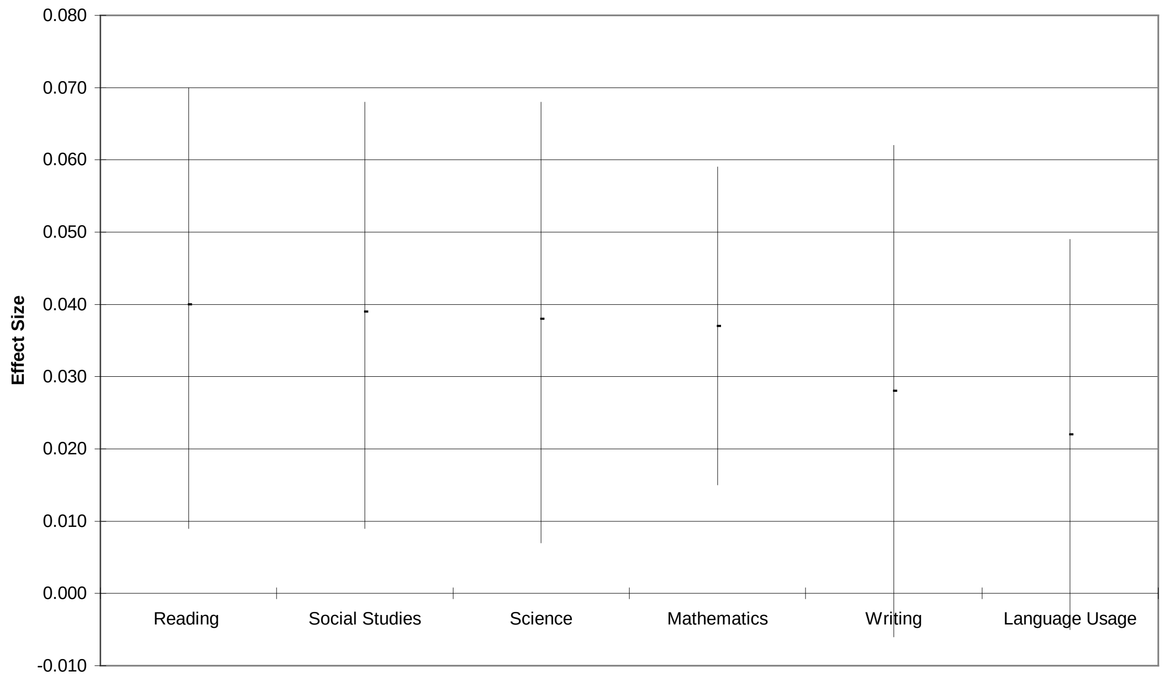
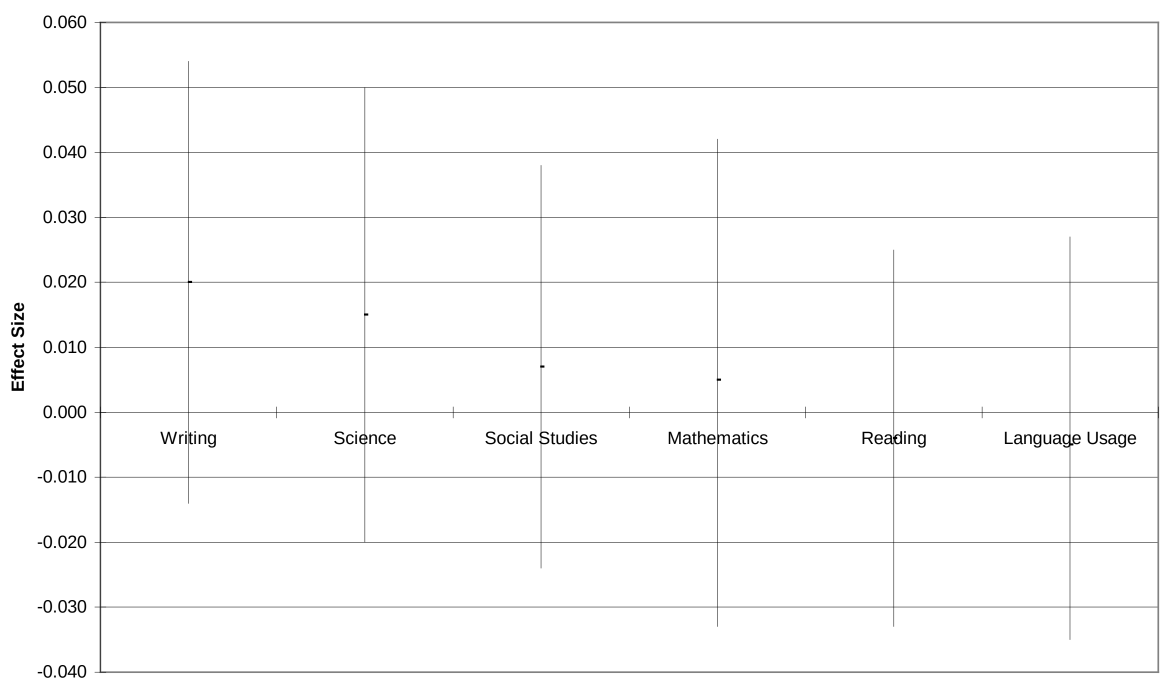


Figure 5. Third Grade Effect Sizes for Examiner Tenure



Content Areas Tenure is coded 0 for not tenured and 1 for tenured.

Figure 6. Fifth Grade Effect Sizes for Examiner Tenure



Content Areas Tenure is coded 0 for not tenured and 1 for tenured

Figure 7. Third Grade Effect Sizes for Prior Examiner Experience

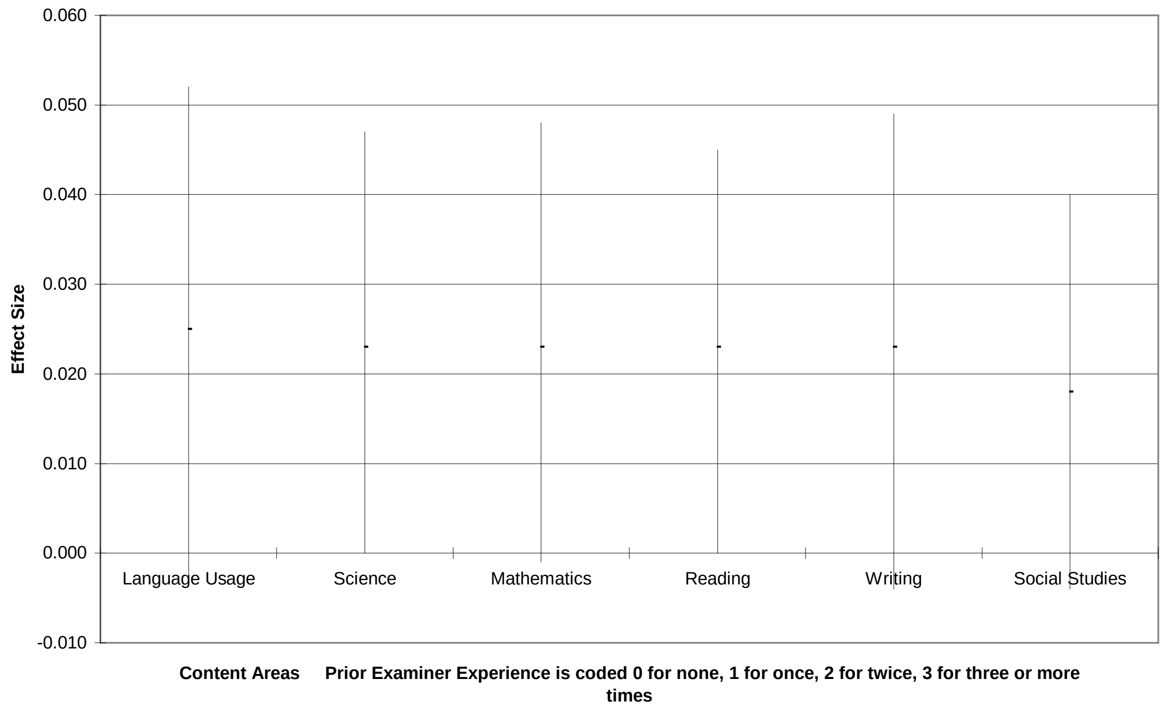


Figure 8. Fifth Grade Effect Sizes for Prior Examiner Experience

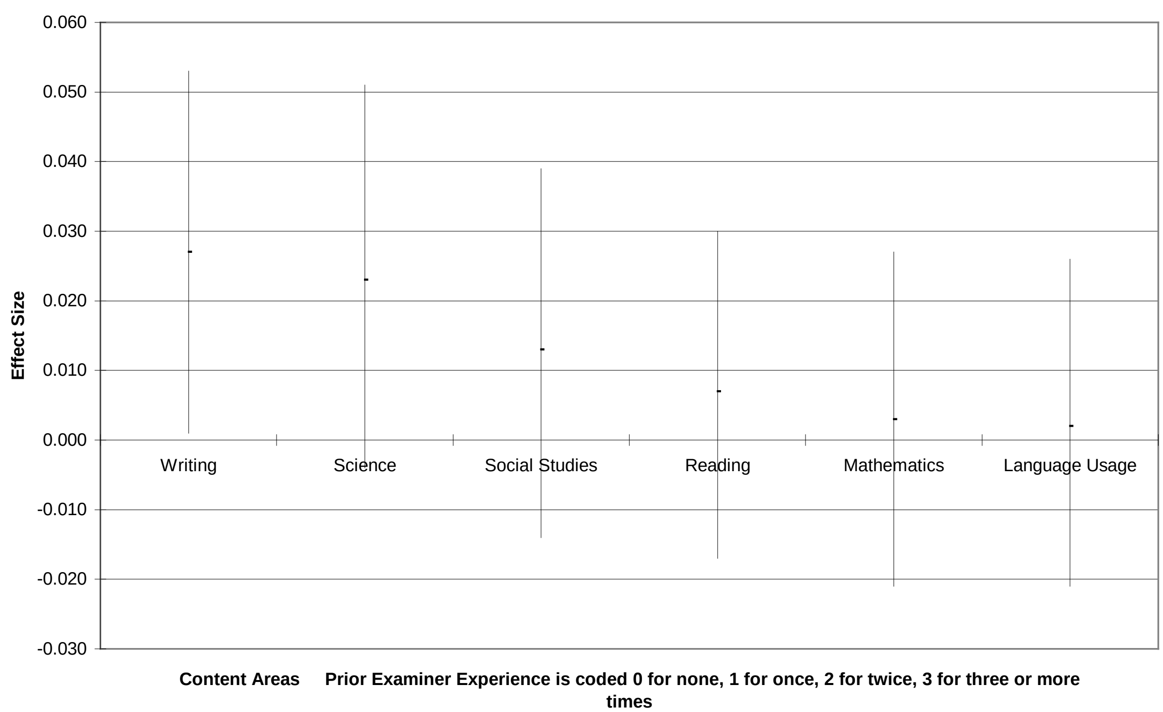


Figure 9. Third Grade Effect Sizes for Examiner Involvement

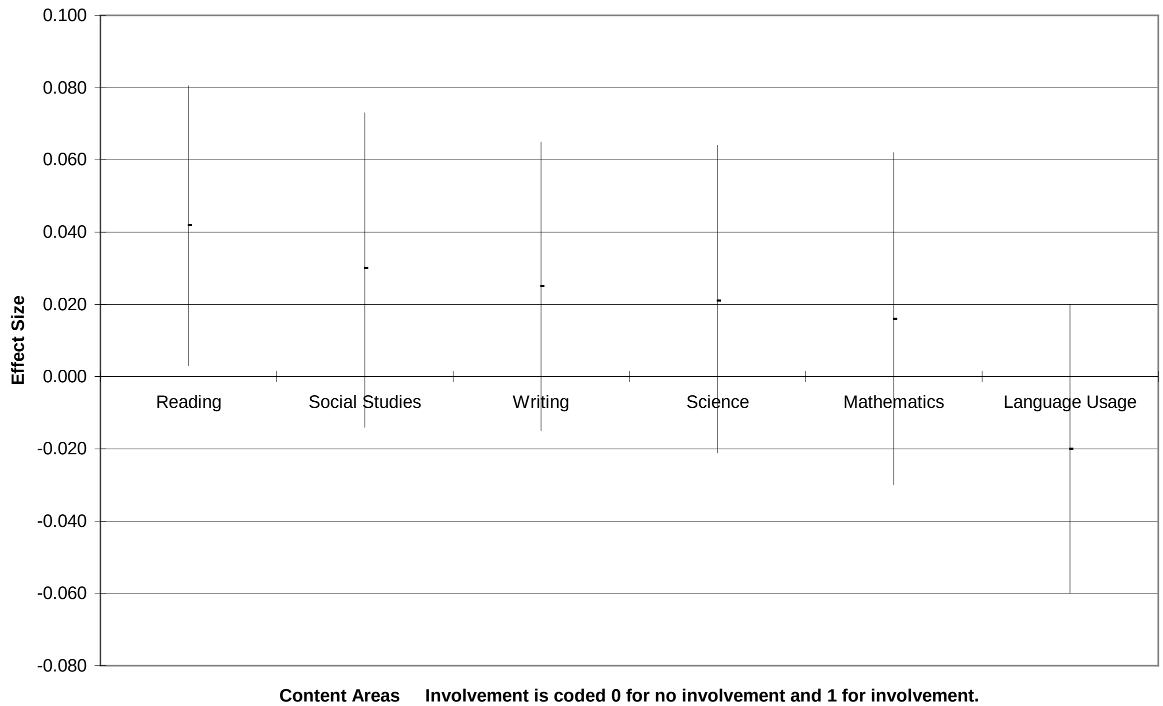


Figure 10. Fifth Grade Effect Sizes for Examiner Involvement