Practical Assessment, Research, and Evaluation

Volume 12 *Volume 12, 2007* Article 13

2007

The Performance Effects of Word Locator Cues on the NAEP Reading Assessment

Howard T. Everson

Steven J. Osterlind

Enis Dogan

William Tirre

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 12, Number 13, October 2007 ISSN 1531-7714

The Performance Effects of Word Locator Cues on the NAEP Reading Assessment

Howard T. Everson, Fordham University
Steven J. Osterlind, University of Missouri-Columbia
Enis Dogan, American Institutes for Research
William Tirre, National Center for Education Statistics

Beginning with the 2009 National Assessment of Educational Progress (NAEP) reading assessment, a new subset of items will be introduced with the intent of measuring vocabulary in context. The assessment's item format requires an examinee to locate a targeted word in the reading passage. It was reasoned that presenting these items along with 'word locator cues' might help reduce construct irrelevant variance due to students' differential ability in searching the targeted word. Using a sample of 1323 fourth and eight grade students, this study investigated the effects of two such 'word locator cues' on student performance: numbering the lines of the passage, and printing targeted words in boldface type. The results indicated that various format conditions (with and without cues) do not influence student performance on the vocabulary items after controlling for reading comprehension. On the other hand, at both fourth and eighth grade, we detected interactions between format conditions and race/ethnicity, which suggested that word locator cues appear to hurt the performance of certain subgroups. Implications of these findings for NAEP's future reading assessments are discussed.

Beginning with the 2009 NAEP Reading Assessment, a new subset of test items will be introduced with the intent of measuring vocabulary in context. These new test items are designed to gauge students' ability to infer the correct meaning or sense of a particular word in context. The context for the targeted words is a reading passage, usually about one paragraph long. The 2009 NAEP Reading Framework (National Assessment Governing Board, 2005) defines the construct as follows:

> *Vocabulary assessment will occur in the context of a passage, that is, vocabulary items will function both as a measure of passage comprehension and as a test of readers' specific knowledge of the word's meaning as intended by the passage author. A sufficient number of vocabulary items at each grade will provide reliable and valid information about students' vocabulary knowledge.* (p. iv)

In 2009 NAEP assessment, reading passages will be presented first, followed by a four-response, multiple-choice item for each meaning vocabulary word. For each item examinees selected the meaning that most closely matches the author's intention. This format requires an examinee to locate a targeted word in the reading passage, and then select from among a number of alternatives the definition that most closely matches the word. The search process itself can be a source of score variance because students may differ in their ability to search for and find the targeted word. Some students, for example, may rely on their short-term memory to identify the targeted word, while others may use visual cues and clues. Still others may engage another visual or semantic search strategy. Thus, the search process alone, though unintended, could be a source of construct-irrelevant variation in the reading assessment scores. Messick (1989) identifies this and other variables as sources for error that can erode valid appraisal.

To mitigate this source of potential measurement error, an advisory panel suggested using word locator cues—such as line numbers or boldface print for targeted words—to help

students more easily find the targeted words in a reading passage.

This study seeks to address these concerns by presenting meaning vocabulary words to examinees in varying formats (i.e., with and without line numbers or boldface location cues) and systematically investigating their effects on examinees' scores.

More specifically, the study was designed to investigate the effects of target word location cues on students' ability to respond correctly to meaning vocabulary test items in context. In addition to examining the overall effect of word location cues on students' meaning vocabulary scores, we were interested in discovering if word location cues produce differential effects on these scores depending on students' reading comprehension ability, or if the effects differ for students from different gender or racial/ethnic groups.

Related Literature

While there is not extensive literature on studying the contextual clues replied upon by examinees, there are some significant studies and a body of related work. For example, Glynn and DiVesta (1979) identified two main forms of cues used by examinees: 1) instructional, and 2) typographical. According to these researchers, instructional cues include advance organizers and adjunct questions, while typographical cues are such stylistic devices as underlining, highlighting, italics and indentation. Our work is most closely aligned with the typographical form of cue. Working with undergraduates in educational psychology courses, the researchers presented students with one of four underlining formats in a systematic manner. They concluded that cues were both preferred by the students and that the presence of cues enhanced performance, an important finding relevant to our investigation.

Meeks (1977) conducted a study examining the effects of imbedded aids—clusters of attention-heightening cuing methods—and noted their effects on seventh graders' reading scores. Meeks concluded that imbedded reading aids did little to improve their reading comprehension.

In a study by Lorth and Chen (1986) the students were asked to identify pieces of supporting evidence for an argument, where reading materials were numbered for reference. They theorized that, "If readers use number signals to organize their representations of a text, they should be able to use the numbers as cues to guide the retrieval process" (p. 264). And, their conclusions were useful in guiding our work in that these researchers concluded that overall, students recall more information with numerical cues than without cues.

METHODOLOGY

Sample

The study included samples of students from the fourth and eighth grades who were recruited from volunteer schools at five states across the country. The NAEP state coordinators were asked to recruit volunteer schools to participate in the study. Each school was given a laptop computer as an incentive. The number of schools that participated was seventeen and seven at fourth and eighth grade respectively. Students were told the test was part of a regular NAEP assessment. Of the students recruited, 86 % participated in the fourth and 91% participated in the eighth grade. The participation rates by demographic groups were not available in this study.

As a result, the fourth grade sample was composed of 730 students from all five states . More than half (53%) were girls. The sample was racially and ethnically diverse and included White (61%), African-American (18%), Hispanic/Latino (11%), Asian-American (4%) and Native American (3%) students. Moreover, nearly four of ten students (37%) were eligible for either free or reduced price lunch.

The eighth grade sample was composed of 593 students from four of the five states. 51% of the students were girls. Again, the sample was racially and ethnically diverse: White (61%), African-American (26%), Hispanic/Latino (8%), Asian-American (4%), and Native American (.3%). Roughly 37% of the students in the eighth grade sample were eligible for free or reduced price lunch.

Assessment Instruments

Our assessment instruments included three standardized, 10-item multiple-choice tests: a meaning vocabulary test that was identical for students in both the fourth and eighth grade samples, and two different multiple-choice reading comprehension tests, one each for grades four and eight. The assessment instruments were designed to be identical to operational NAEP assessments in format, and were administered in the same manner as operational NAEP assessments in order to maximize generalizability of the results. The meaning vocabulary test items were designed as cross-grade items, appropriate for both fourth and eighth graders. On the meaning vocabulary test, all of the items were in the four-response, multiple-choice format and each item was scored dichotomously. The format of the reading passages presented in the meaning vocabulary test, of course, varied and included three different conditions: (1) standard format, (2) line numbering, and (3) bold-faced targeted words.

The reading comprehension passages, in contrast, had no format modifications. This measure was included to provide an indicator of the students' reading comprehension abilities and used exclusively as a covariate in the statistical analyses. For the eighth grade test, four of these items were in the four-response, multiple-choice format, and each was scored dichotomously. The remaining six test items were constructed-response type items. One of these was dichotomously scored, while the remaining five items were scored using a partial credit rubric. Similarly, the 4th grade

reading comprehension test was composed of five multiple-choice and five constructed-response type items. The multiple-choice items were in the standard four-response format and scored dichotomously. For the remaining five constructed-response items, students were assigned partial credit depending on the completeness of their responses.

Reliability for the measures was investigated by computing the coefficient alpha statistic (Cronbach, 1951). The reading comprehension and vocabulary measures used with the fourth graders had reliability coefficients of .68 and .72, respectively. The eighth grade reading comprehension and vocabulary measures had reliability estimates of .66 and .74, respectively. Given that these tests had few items (only 10 on each instrument), the reliability estimates were regarded reasonably good.

Also, background questionnaires were administered to all students in our study. The fourth graders answered 11 questions asking about their race/ethnicity, home environment (e.g. number of books at home, language spoken at home), school attendance and homework habits. The eighth graders were asked the same set of questions, plus two additional ones asking about their parents' education levels.

Procedures

As noted earlier, students at both grade levels were assigned randomly to either one of three format conditions when taking the meaning vocabulary assessment: (1) the standard condition, (2) line-numbered passages, or (3) targeted words in the passages were highlighted using a bold-face format. Random assignment to test condition was assured by spiraling the test booklets within each classroom. Students were allowed 25 minutes to complete each of the two tests, and five minutes to complete the background questionnaire. The experimental test administrations occurred between May 2006 and June 2006.

The reading comprehension test was administered first and the meaning vocabulary test followed immediately. The background questionnaire was given to the students at the end.

Units of Analyses

In our study students' scores were the unit of analysis. We recognize that this is unlike the national administrations of NAEP where individual student scores are not possible because of the matrix sampling design. In our research design all students took both tests in their entirety. The metric we used throughout our analysis was total raw score for both the meaning vocabulary and the reading comprehension measures. Thus, our scores ranged from 0 to 10 for each test.

Analytic Approach

For the most part we relied on general linear modeling methods to analyze the data in this study. More specifically, the raw scores on the meaning vocabulary tests served as the dependent variables in our analyses. The reading comprehension scores served as covariates in the ANCOVA s and multiple regression analyses. In addition, the main independent variable was passage format, which had three levels: (1) standard, (2) line numbering, and (3) bold faced targeted words.

All main effects and interactions were tested. Analyses were conducted separately for each grade level. Where appropriate, we also conducted follow-up contrast tests in which each focal group was contrasted with a reference group. We used the standard format condition as the reference group in some analyses, and for racial investigations we used White students as the reference group.

In all analyses we adopted the standard criterion of .05 for judging statistical significance, though we adjusted the stringency of this criterion in the simple contrast tests to guard against Type I errors. Following the strong recommendation of the American Psychological Association's *Task Force on Statistical Inference* (Wilkinson, 1999) we paid special attention to effect size. The educational measurement literature includes several measures for effect size. In this study we used partial-η^2 (partial eta-squared) as our measure. The partial-η^2 is useful here because it captures the proportion of the variance in the dependent measure accounted for by the independent variable, and it is computationally simple. Values for partial-η^2 approximately correspond to the following effect size conventions: small (.01), medium (.06), and large (.14) (Cohen, 1988).

Sample distributions

Prior to our analyses the distribution of the dependent variable (i.e., total vocabulary score) was explored using both traditional and IRT distributional statistics, including standard checks on linearity and homoscedasticity. The independence of observations was assumed. The eighth grade data showed a moderate negative skew with kurtosis higher than normal. We explored transformation but the distributions were not substantially improved. However, linearity and homoscedasticity (for homogeneity of variance) of eighth grade scores were acceptable. The distribution of scores at the fourth graders, on the other hand, had virtually no skew and only minimal kurtosis. The linearity and homoscedasticity of scores at the fourth grade were also acceptable.

Power analysis

Overall, the power for the ANCOVAs was .50, a value that is considered moderately acceptable for experimental research (Keppel & Wickels, 2004). The values for the separate ANCOVAs were also very close to this, and varied by only a decimal place or two.

We used one-tailed analyses to determine power. As mentioned above, our alpha level was .05 for all models

discussed. As is routinely done in variance analyses, we examined power *post hoc*, given our alpha level, sample size and the resultant effect size. For calculation of power, the *G*Power* software was used (Faul & Erdfelder (1992).

RESULTS

To reiterate, we computed total vocabulary and reading comprehension scores for each student by simply summing their correct responses to the ten items presented in each test. These raw scores, which ranged from 0 to 10, served as the dependent measures and covariates in all subsequent analyses.

Tables 1 and 2 present the means and standard deviations on both the meaning vocabulary and reading comprehension measures for the fourth and the eighth grade student samples under each of the three format conditions.

Table 1: Fourth Grade Means, Standard Deviations and Sample Sizes for the Vocabulary and Reading Comprehension Scores, by Format Condition

	Standard			Line Number			Bold-Face		
Variable	M	SD	n	M	SD	n	M	SD	n
Reading Comprehension Score	5.46	2.14	246	5.68	2.14	240	5.48	2.22	244
Vocabulary Score	4.91	2.63	246	5.13	2.53	240	4.98	2.60	244

Table 2: Eighth Grade Means, Standard Deviations and Sample for the Vocabulary and Reading Comprehension Scores, by Format Condition

	Standard			Line Number			Bold-Face		
Variable	M	SD	n	M	SD	n	M	SD	n
Reading Comprehension Score	5.68	2.09	197	5.68	2.18	195	5.72	1.91	201
Vocabulary Score	8.55	1.87	197	8.46	1.99	195	8.60	1.82	201

For both 4th and 8th graders the reading comprehension scores were quite similar, hovering around 5.50, with a standard deviation of about 2.10. It is also reassuring to note, given our overt attempt at assigning students randomly to format condition, that the reading comprehension scores did not vary significantly across format conditions in either grade level.

As expected the correlations between the various reading comprehension and vocabulary measures were moderately high and consistent across both grade levels. The reading comprehension scores correlated with the prototype meaning vocabulary scores for both the fourth and eighth graders, r = .60 and .53, respectively.

The central research question of this study was to examine whether the test item format—use of line numbered reading passages or bold-faced targeted vocabulary words—affected performance on meaning vocabulary measure. We explored this question for the entire group and disaggregated results by gender and race/ethnicity. The reading comprehension measure was used as a covariate to adjust for initial differences.

Analysis of Covariance

Table 3 displays the adjusted means (Pedhazur & Schmelkin, p. 572) for the meaning vocabulary scores by format condition, controlling for reading comprehension, for the fourth and eighth grade samples. The interaction between the reading comprehension score and format condition was not significant at either fourth ($F_{(2, 724)} = 0.31$,

p=.73) or at eighth grade ($F_{(2, 587)} = 0.70$, p=.50). Moreover, the results of the ANCOVAs suggested that once the influence of reading comprehension had been controlled for, various format conditions did not influence the meaning vocabulary scores at either grade four ($F_{(2, 726)} = 0.06$, p=.96) or grade eight ($F_{(2, 589)} = 0.30$, p=.74).

Table 3: Mean Vocabulary Scores by Format Condition, adjusted for Reading Comprehension

	Standard	Line number	Bold
Grade 4	4.97	5.02	5.02
Grade 8	8.56	8.47	8.59

Also the effect sizes associated with the format conditions were negligible (partial-η^2 <.01) at both grade levels. Thus, at both grade levels, the effect size metrics and the results of the ANCOVAs suggest no difference in the meaning vocabulary scores across all three format conditions. To ensure thorough analyses of all available data, we took additional steps to investigate whether there were interactions between gender and format conditions and between race/ethnicity and format conditions. These analyses are reported next.

Interactions Among Examinee Race/Ethnicity, and Gender, and Format

Gender and race are both important considerations for the fairness and validity of standardized tests. To explore the effects of gender and race and their possible interactions with test formats, we repeated the ANCOVA approach used earlier, which specifically looked at the whether there are differences between the gender and race/ethnicity groups on the vocabulary test under varying format conditions. After controlling for differences in overall reading comprehension, we found no significant main effects for gender ($F_{(1, 723)} = .26$, p=.61) or gender by format condition interactions ($F_{(2, 723)} = .06$, p=.95) at the fourth grade, or at the eighth grade ($F_{(1, 586)} = 2.23$, p=.14, and $F_{(2, 586)} = .77$, p=.46) respectively.

When we examined the effects for race/ethnicity, however, the results were more complex. Our analyses focused on understanding the effects of variations in vocabulary test format on the performance of three racial/ethnic groups: White, African American and Hispanic. Note that there were too few examinees from the remaining two racial/ethnic groups to consider in these analyses.

Looking only at the White, African American and Hispanic students in the fourth grade sample, the ANCOVA (controlling for reading comprehension) revealed a significant race/ethnicity by format condition interaction ($F_{(4, 647)} = 2.41$, p=.05). At this point, we conducted four separate ANCOVAs contrasting combinations of the format conditions (Standard versus Line number and Standard versus Bold) and the racial/ethnic groups (White versus African American and White versus Hispanic) to get a better understanding of the race by format condition interaction(s). These analyses indicated a significant format by race/ethnicity interaction when Standard and Line number conditions were contrasted along with a White versus African-American racial contrast ($F_{(1, 377)} = 6.48$, p=.01).

Figure 1 displays the nature of this interaction. As shown in this figure, the difference between the format conditions depends on racial group and vice versa.

Next, we conducted two additional analyses comparing the relative performance of White and African American students under the standard and the line number conditions separately. The purpose was to see if significant differences existed between the average meaning vocabulary scores of these two racial groups under any one of these two format conditions, after controlling for their reading comprehension scores. Results of these additional analyses indicated that although there was no difference in the meaning vocabulary performance of the White and African American fourth graders under the standard format condition, the performance of the African American students was significantly lower than that of the White students under the line number format condition ($F_{(1, 377)} = 14.07$, p=.00). Table 4 below summarizes the results of this analysis.

Figure 1: Plot of fourth Grade White and African-American Students' Adjusted Meaning Vocabulary Scores under Standard and Line Number Conditions, Controlling for Reading Comprehension

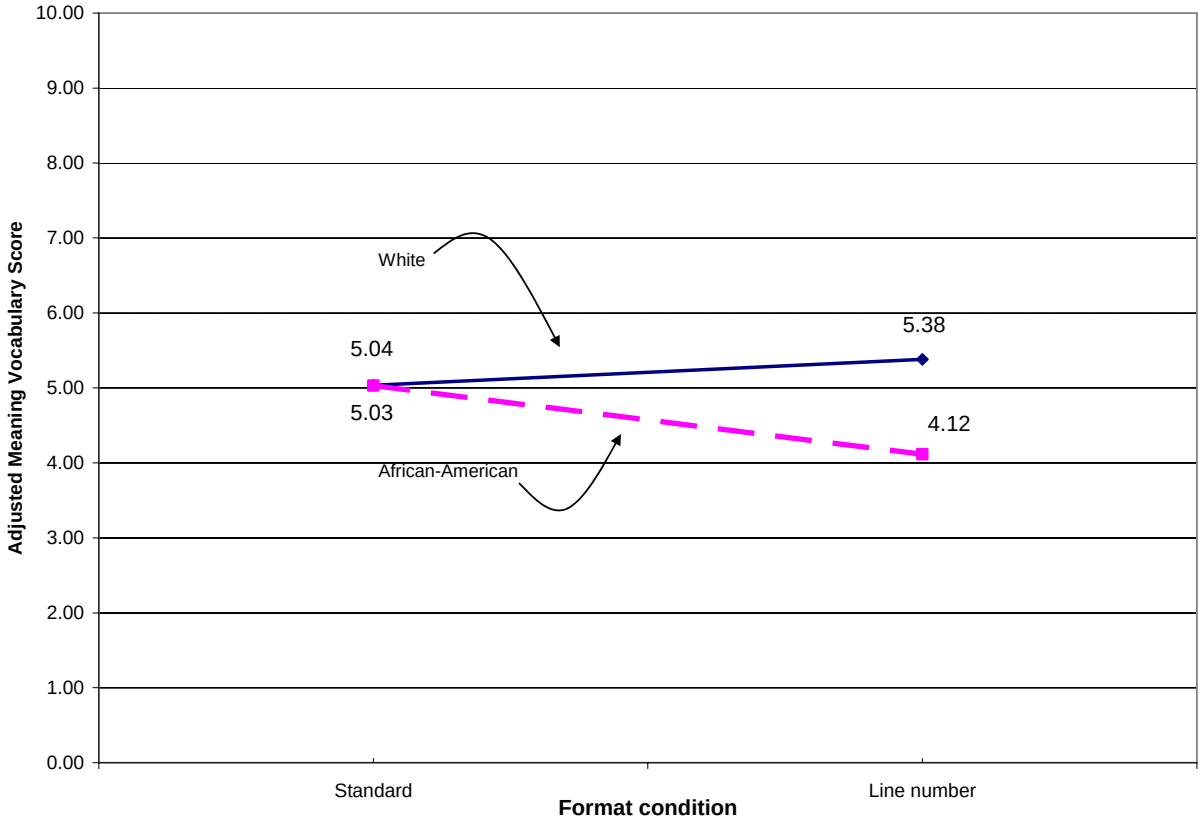


Table 4: Summary of Regression Analysis Comparing Meaning Vocabulary Performance of White and African-American students under Line Number Condition at fourth Grade, controlling for Reading Comprehension

Variable	B	*SE* B	*β*
Reading Comprehension score	0.73	0.07	0.58*
Contrast code (White v. African-American)	1.28	0.34	0.21*

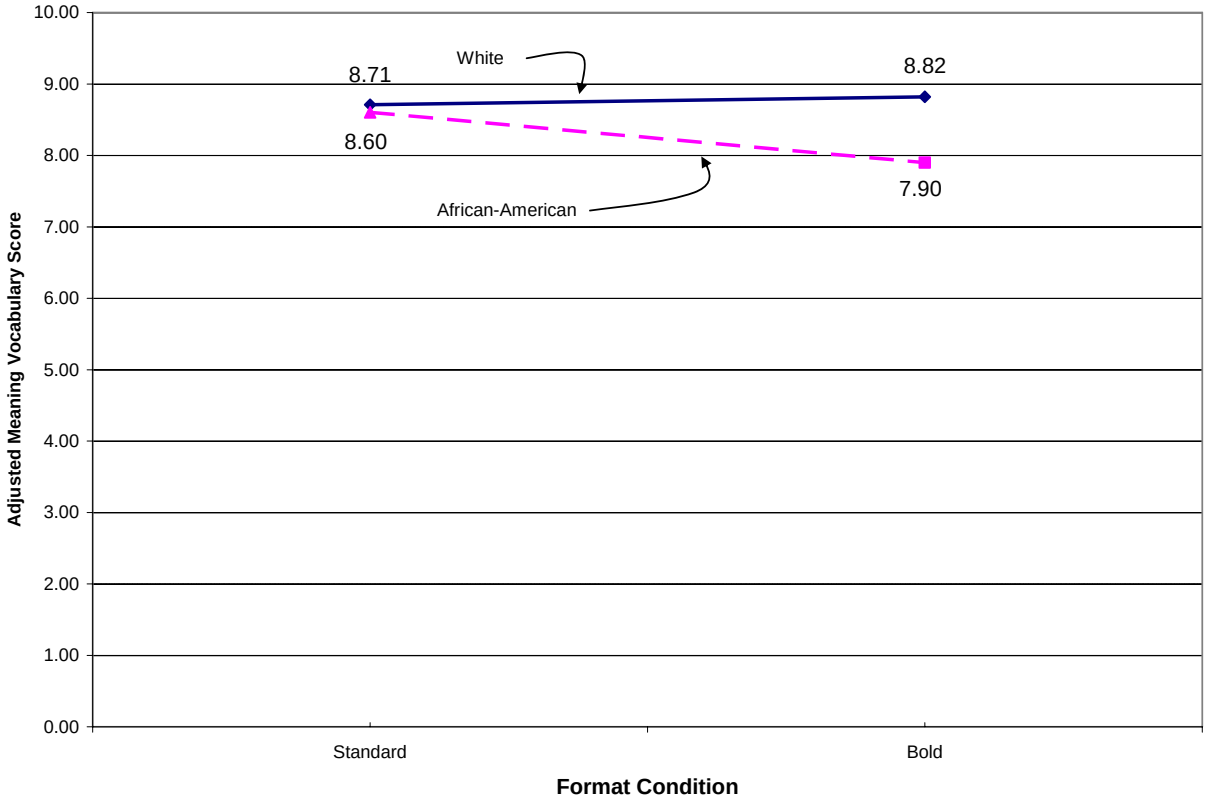
Note. $R^2 = .46$, $F_{(2, 335)} = 76.81$; * $p < .01$

The standardized *Beta* coefficient for the White versus African-American comparison under the Line number condition was 0.21. The magnitude of this coefficient suggests an increase (above and beyond initial differences on reading comprehension) in the White-African American performance gap of roughly one-fifth of a standard deviation under the line number format with an estimated effect size of 0.07 (a medium effect).

Regarding the eighth grade sample, we found a similar, though not identical, pattern of format by race interactions. Overall, our analyses of eighth graders' performance on the meaning vocabulary test revealed significant format by race/ethnicity interactions ($F_{(8,576)} = 2.74$; p=.01). We again conducted four separate ANCOVAs contrasting combinations of the format conditions (standard versus line number and standard versus bold) and the racial/ethnic groups (White versus African American and White versus Hispanic) to better understand the race by format condition interaction(s). These analyses indicated a significant format by race/ethnicity interaction when Standard and Bold conditions were contrasted along with a White versus African-American racial contrast ($F_{(1, 335)} = 5.53$, p=.02). Figure 2 displays the nature of this interaction.

Practical Assessment, Research & Evaluation, Vol 12, No 13
Everson, Osterlind, Dogan & Tirre, Word Locator Cues

Figure 2: Plot of fourth Grade White and African-American Students' Adjusted Meaning Vocabulary Scores under Standard and Bold Conditions, Controlling for Reading Comprehension Score



Following the analyses that indicated significant race/ethnicity by format interactions, we conducted two additional analyses comparing the relative performance of White and African-American students under the standard and the bold conditions separately. The purpose was to see if significant differences existed between the average meaning vocabulary scores of these two racial groups after controlling for their reading comprehension scores under any one of these two format conditions.

The results of these additional analyses indicated that although there were no differences in the vocabulary performance of the White and African American eighth graders under the Standard format condition, the performance of the African American students was significantly lower than that of the White students under the Bold format condition ($F_{(1, 166)} = 10.61$, p=.00). Table 5 summarizes the results of this analysis.

Table 5: Summary of Regression Analysis Comparing Meaning Vocabulary Performance of White and African-American students under Bold condition at Eighth Grade, controlling for Reading Comprehension

Variable	B	*SE* B	β
Reading Comprehension score	0.52	0.06	0.53*
Contrast code (White v. African-American)	0.89	0.27	0.21*

Note. $R^2 = .38$, $F_{(2, 166)} = 49.74$; * $p < .01$

The standardized Beta coefficient for the White versus African-American comparison under the Bold condition was 0.21, suggesting an increase (above and beyond initial differences on reading comprehension) in White-African American performance gap of roughly one fifth of a standard deviation under the bold format condition with an estimated effect size of .06 (a medium effect).

The Effect of Format Condition across States

Another important question we explored was the following: Do students from different states perform differently on the meaning vocabulary items under various format conditions? The information we received from the five participating states did not suggest that they differed substantially in their use of 'word locator cues' in their state assessments. Nevertheless, we explored whether the relative performance of the students from these states differed across format conditions after controlling for reading comprehension. In order to find an answer to this question we conducted an ANCOVA at each grade, where we treated both format condition and state as fixed effects. Tables 6 and 7 display the adjusted means (Pedhazur & Schmelkin, p. 572) of the meaning vocabulary scores for each state across format conditions for the fourth and the eighth grade samples, respectively.

Table 6: Fourth Grade Meaning Vocabulary Means by Format Condition and State, adjusted for Reading Comprehension

	Standard	Line number	Bold
State 1	4.57	4.94	5.15
State 2	4.94	5.36	5.15
State 3	5.36	4.98	4.95
State 4	4.82	4.64	4.92
State 5	5.25	5.06	4.88

Table 7: Eighth Grade Meaning Vocabulary Means by Format Condition and State, adjusted for Reading Comprehension

	Standard	Line number	Bold
State 1	8.54	8.34	8.50
State 3	8.93	9.09	9.36
State 4	8.51	8.37	8.38
State 5	8.56	8.75	8.91

Results of the ANCOVAs indicated no state by format condition interaction at either fourth ($F_{(8,714)} = 0.65, p = .73$) or eighth grade ($F_{(6,580)} = 0.31, p = .93$). Results also indicated no format condition effect at either grade, ($F_{(2,714)} = 0.01, p = .99$ and $F_{(2,580)} = 0.38, p = .68$, respectively). The effect sizes for these analyses were all of negligible size (partial-η^2<.01).

DISCUSSION

The results of our analyses suggest that variations in test format for the proposed meaning vocabulary measure do not have a main effect on test scores for either fourth- or eighth-grade students. Nonetheless, we hasten to point out consistent evidence suggesting race/ethnicity by format condition interactions. For example, at the fourth grade we observed that African-American students performed less well on the meaning vocabulary test under the line number format condition when contrasted with their White counterparts. Similarly, eighth grade African-American students performed less well when contrasted with White students under the boldface condition. The magnitude and direction of the statistical effects of these analyses indicated a moderate level of association. On the other hand, our analyses produced no evidence suggesting that males versus females or students from separate states perform differently on the meaning vocabulary items under various test format conditions.

Like all research concerning student performance on standardized tests, there are a number of limitations to be mindful of when interpreting this study's results. For instance, although we detected a number of significant interactions, it is difficult to draw clear conclusions from the apparent non-significant effects and interactions because of the small sample sizes in certain comparisons and the low statistical power associated with them.

The operational differences between the design and implementation of this study and a typical NAEP assessment also impose certain limitations to the generalizability of our results. The effects of word locator cues we observed in this study might have been due to some characteristics of the passages and items that would not be present in other passages. With a larger sample of possible passages and items, we might have found that the word locator cues would have produced different effects. Our evidence suggests, however, that at least for some passage/item combinations word locator cues introduce unwanted error variance.

Moreover, the samples used in this study were not drawn randomly from the population of fourth and eighth grade students in the participating states and that the ethnic and socio-economic diversity of our particular samples do not necessarily mirror the typical national samples taking part in NAEP.

The nature of the format by race/ethnicity interactions found in this study requires further investigation. Specifically, the reasons behind these interactions can be explored. For instance, future research may explore whether word location cues increase the "cognitive load" of an item. It is also possible to explore if students use these cues as shortcuts by moving directly to the question before reading the entire passage. Obviously, these and other possible hypotheses could be tested empirically. Approaches incorporating the use of cognitive interviews in which students are queried about their test-taking strategies and behaviors may be a productive route to explore for answers, too.

In sum, the results of this study suggested that placing word locator cues into regular NAEP assessments might create adverse effects for some students.

REFERENCES

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crouse, J. H. & Idstein. P.(1972). Effects of Encoding Cues on Prose Learning. *Journal of Educational Psychology,* 68(4), 309-313.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28, 1-11.

Faul, F., & Erdfelder, E. (1992). GPOWER: A priori, post-hoc, and compromise power analysis for MS-DOS [Computer program]. Bonn, FRG: Bonn University, Dept of Psychology. Retrieved July 6, 2006; Access: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/ .

Glynn, S. M. & Di Vesta, F.J. (1979). Control of Prose Processing via Instructional and Typographical Cues. *Journal of Educational Psychology,* 71(5), 595-603.

Keppel, G. & Wickens, T. D. (2004). *Design and Analysis: A Researcher's Handbook, 4th ed.* Upper Saddle River, NJ: Pearson/Prentice Hall.

Lorch, R. F. & Chen, A.H. (1986). Effects of Number Signals on Reading and Recall. *Journal of Education Psychology,* 78 (4), 263-270.

Lorch, R. F. Jr., & Lorch, E. P. (1996). Effects of Organizational Signals on Free Recall of Expository Text. *Journal of Educational Psychology,* 88(1), 38-48.

Manzo, A.V., Manzo, U. C., & Thomas, M. M. (2005). *Content Area Literacy: Strategic Teaching for Strategic Learning*. Hoboken, NJ: John Wiley & Sons, Inc.

Meeks, J. W. (1977). *An Investigation into the Effects of Imbedded Aids*. Paper presented at the Annual Meeting of the National Reading Conference, New Orleans, Louisiana, December 1-3, 1977.

Messick, S. (1989). Validity. In R. Linn (Ed.). *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Pedhazur, E. J., & Schmelkin, L. P. (1991). *Measurement, Design and Analysis: An Integrated Approach.* Hillsdale, NJ: Lawrence Erlbaum Associates.

National Assessment Governing Board (2005). *Reading framework for the 2009 National Assessment of Educational Progress*. Pre-publication edition. Washington, DC: American Institutes for Research.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54, 594–604.*

Citation

Everson, Howard T., Steven J. Osterlind, Enis Dogan and William Tirre (2007). The Performance Effects of Word Locator Cues on the NAEP Reading Assessment. *Practical Assessment Research & Evaluation*, *12*(13). Available online: http://pareonline.net/getvn.asp?v=12&n=13