Preliminary Report on the 2004–05 Evaluation Study of the ASCD Program for Building Academic Vocabulary

by
Robert J. Marzano
September 2005

This report summarizes the initial findings from the 2004–05 evaluation study of the program titled Building Academic Vocabulary (BAV). The theory and research supporting the development of this program are articulated in *Building Background Knowledge for Academic Achievement* (Marzano, 2005). Briefly, the basic assumption underlying the BAV program is that teaching standards-based academic terminology using a specific six-step process can enhance students' abilities to read and understand subject-area content and ultimately help students build a store of academic background knowledge that enhances academic achievement.

Design of the Evaluation Study

At the basic level, this evaluation study asked two questions:

1. What is the effect of the BAV program on students' ability to read and comprehend subject-area content?

2. Does the effect of BAV differ from grade level to grade level?

In more technical terms, the study employed two primary independent variables: (1) whether teachers used the BAV program (referred to as the *experimental* teachers) or some other set of vocabulary instruction techniques (referred to as the *control* teachers), and (2) the grade level of the students (grades 1–9). In research and evaluation terminology, the independent variables are those factors assumed or hypothesized to have an effect on some outcome, which is usually referred to as the dependent variable. (See Technical Note 1 for further consideration of independent and dependent variables.) In this case, the dependent variables were students' abilities to read and understand information regarding mathematics, science, and general literacy.

The Sample

The sample was drawn from volunteer schools and teachers across the United States. Specifically, in the summer of 2004 the Association for Supervision and Curriculum Development (ASCD) issued an invitation for schools to participate in an evaluation of the BAV program. Participating schools were required to furnish both experimental and control teachers and students. In all, five districts—involving 11 schools, 118 teachers, and 2,683 students—accepted the invitation. These districts, schools, teachers, and students represented a broad range of demographic and socioeconomic factors.

The numbers of students by experimental/control conditions by grade levels are reported in Figure 1.

Figure 1: Subjects in Experimental/Control Conditions by Grade Level

Grade Level	Control	Experimental (BAV)
Kindergarten	48*	0
1	118	52
2	296	135
3	180	103
4	100	171
5	130	183
6	145	341
7	143	131
8	132	0
9	158	117
Total	**1450**	**1233**

*__Note:__ Sixteen pre-kindergarten students involved in the study were included in the kindergarten count.

The Intervention

In September of 2004, teachers participating in the experimental classes met in Denver, Colorado, for a two-day training on the BAV protocols. Each member received a draft copy of the BAV teacher's manual (Marzano & Pickering, 2005). Each district brought a contact person whose task it was to coordinate data collection for the project and communicate with ASCD. The two-day training was the only formal training provided.

The Dependent Measures

Four dependent variables were addressed:

- Ability to read and understand information about mathematics.
- Ability to read and understand information about science.
- Ability to read and understand information about general literacy.
- Aggregated ability to read and understand information across the three subject areas.

Each of the four dependent *variables* was assessed in two formats: multiple-choice and constructed response. In effect, then, eight dependent *measures* were employed in this study:

- Mathematics assessed using multiple-choice items.
- Mathematics assessed using a constructed-response item.
- Science assessed using multiple-choice items.
- Science assessed using a constructed-response item.
- General literacy assessed using multiple-choice items.
- General literacy assessed using a constructed-response item.

- The combined scores for multiple-choice items regarding mathematics, science, and general literacy.
- The combined scores for the constructed-response items regarding mathematics, science, and general literacy.

For each dependent measure, four levels of assessments were constructed. Level 1 assessments were designed for students in grades K–2; level 2 assessments were designed for students in grades 3–5; level 3 assessments were designed for students in grades 6–8; and level 4 assessments were designed for students in grades 9–12. Two versions of each assessment were developed: one to be used as a pre-test and one to be used as a post-test for both experimental and control subjects.

ASCD staff members constructed all assessments. They began by identifying appropriate reading material in mathematics, science, and general literacy for each of the four levels. They then constructed first-draft versions of multiple-choice and constructed-response items for each subject-matter passage at each level. Marzano & Associates (M&A) reviewed all items for face validity. On the basis of the feedback from M&A, final versions of the items were constructed.

Pre-tests were administered to experimental and control students in October of 2004, and post-tests were given in April of 2005. Upon completion, pre-tests and post-tests were sent to ASCD, where all multiple-choice items were scored by ASCD staff members. To a great extent, the ASCD scorers did not know whether a given student was in the experimental or control condition.

To determine the reliability of the pre- and post-tests, the responses for 100 students in each level were used. Cronbach's coefficient alpha was computed for each level and considered to be the reliability estimate for each assessment. (For a discussion of Cronbach's alpha, see Technical Note 2.) Alpha coefficients are reported in Figure 2 for the pre-test and post-test at each level.

Figure 2: Reliability Estimates Using Cronbach's Coefficient Alpha

Dependent Measure	Pre-test	Post-test
Total for Multiple-Choice Response	L4: .79 L3: .77 L2: .71 L1: .68	L4: .88 L3: .82 L2: .81 L1: .83
General Literacy Multiple-Choice	L4: .69 L3: .68 L2: .73 L1: .68	L4: .79 L3: .77 L2: .79 L1: .74
Math Multiple-Choice	L1: .70 L3: .68 L2: .69 L1: .64	L4: .82 L3: .75 L2: .74 L1: .71
Science Multiple-Choice	L4: .73 L3: .71 L2: .72 L1: .68	L4: .83 L3: .79 L2: .78 L1: .77

Note: L1= level 1; L2=level 2; L3=level 3; L4= level 4.

The constructed-response items for the pre-test and the post-test were scored by a consultant hired by M&A. Because the consultant was proficient in both English and Spanish, scoring responses in both languages was possible. The consultant employed a four-point rubric designed by M&A for each of the constructed-response items. To estimate the interrater reliability of judgments regarding the constructed-response items, 30 student assessments were randomly selected from the pre-tests, and two raters independently scored the constructed-response items for mathematics, science, and general literacy. The correlations between those ratings were considered to be estimates of the interrater reliability of judgments for the constructed-response items. The estimates were .77, .87, and .86 for mathematics, science, and general literacy, respectively.

These estimated reliabilities for the multiple-choice and constructed-response dependent measures are within an acceptable range for the social sciences. For example, Osborne (2003) found that the average reliability reported in psychology journals is .83. Lou and colleagues (1996) reported a typical reliability on standardized achievement tests of .85 and a reliability of .75 for nonstandardized tests of academic achievement.

Data Analysis and Findings

For each of the eight dependent measures, the data were analyzed using the general linear model as employed by the SPSS, version 12.0. The two independent variables (experimental/control condition and grade level) were analyzed as fixed effects. (See Technical Note 3 for a discussion of fixed effects.) In each case, the pre-test was used as the covariate. In effect, a fixed-effects analysis of covariance (ANCOVA) was executed for each of the eight dependent measures.

The advantage of using an analysis of covariance with the pre-test as the covariate is that it statistically adjusts for students' initial status on the measure in question. (For a discussion of the use of ANCOVA, see Technical Note 4.) Metaphorically, one might say that ANCOVA starts all students at the same point relative to the dependent measure. In the absence of random assignment to groups or classes (which this study did not employ), use of the ANCOVA design with the pre-test acting as the covariate is a commonly used technique to address the issue of students' prior achievement relative to the dependent variable.

General Effect of BAV

The study's first evaluation question (What is the effect of the BAV program on students' ability to read and comprehend subject-area content?) addresses the general effect of BAV aggregated across grade levels. Figure 3 provides a brief summary of the findings for each of the dependent measures.

Figure 3: Summary of Fixed-Effects ANCOVAs for Dependent Measures

Dependent Measure	Experimental Group Mean	Control Group Mean	Significance	Differences in Percentage Passing a 50/50 Test
Total for Written Responses	5.87	4.62	.001	14.8%
Total for Multiple-Choice Responses	11.84	10.73	.001	11.0%
General Literacy Written	11.85	11.44	.001	12.3%
General Literacy Multiple-Choice	3.93	3.48	.001	9.5%
Math Written	2.03	1.70	.001	8.4%
Math Multiple-Choice	4.31	3.91	.001	10.0%
Science Written	2.00	1.46	.001	13.0%
Science Multiple-Choice	3.67	3.19	.001	8.4%

As depicted in Figure 3, the mean score for the experimental group (i.e., teachers who used the BAV program) was greater than the mean score for the control group (i.e., teachers who did not use the BAV program) for all eight dependent measures. The "Significance" column in Figure 3 is of particular importance. It indicates that each of the differences between experimental and control means was significant at the .001 level. The standards typically employed in educational

evaluation studies for statistical significance are the .05, .01, and .001 levels. The .05 level of significance is generally interpreted as an indication that the observed difference in the means could have occurred fewer than five times in 100 studies if there is no "true difference" between experimental and control group means. The .001 level of significance is generally interpreted as an indication that the observed differences could have occurred fewer than one time in 1,000 if there is no true difference between experimental and control means. (For a detailed discussion of the meaning of statistical significance, see Harlow, Mulaik, & Steiger, 1997). Taking these conventions at face value, Figure 3 indicates that the differences between experimental and control means typically would be considered "highly" significant.

Statistical significance indicates that observed differences between the experimental group means and the control group means most probably aren't simply a function of chance. However, it does not address how strong the relationship is between the use of the BAV program and student scores on the eight dependent measures. To address this issue, consider the last column of Figure 3, "Differences in Percentage Passing a 50/50 Test." Technical Note 5 explains how the values in this column were computed. Briefly, though, to interpret this column, consider the dependent measure of the multiple-choice items for mathematics (row six of Figure 3). The value for the last column in that row is 10 percent. To interpret this value, assume that students in both the experimental and control classes in this study took the same multiple-choice test after reading the same mathematics content. Also assume that when the responses of all students were examined, the overall passing rate was 50 percent; that is, half of the students in the combined sample passed the assessment and half failed it. However, if one were to separate out the experimental group (the BAV group) from the control group, important differences would be observed, as depicted in Figure 4.

Figure 4: Expected Passing Rate for Experimental and Control Groups on Multiple-Choice Mathematics Test

	Expected to Pass	Expected to Fail
BAV	55%	45%
Control	45%	55%

As illustrated in Figure 4, the BAV group would have an expected passing rate of 55 percent and the control group would have an expected passing rate of only 45 percent—a 10 percent difference, as depicted in the last column of Figure 3. For comparative purposes, Figure 5 reports the expected passing rates between the BAV and control groups as well as the net difference in passing rates for all eight dependent measures.

Figure 5: Expected Passing Rates for Eight Dependent Measures

Dependent Measure	Group	Expected to Pass	Net Difference in Passing Rate
Total for Written Responses	BAV	57.4%	14.8%
	Control	42.6%	
Total for Multiple-Choice Responses	BAV	55.5%	11.0%
	Control	44.5%	
General Literacy Written	BAV	56.15%	12.30%
	Control	43.85%	
General Literacy Multiple-Choice	BAV	54.75%	9.5%
	Control	45.25%	
Math Written	BAV	54.2%	8.4%
	Control	45.8%	
Math Multiple-Choice	BAV	55.0%	10.0%
	Control	45.0%	
Science Written	BAV	56.5%	13.0%
	Control	43.5%	
Science Multiple-Choice	BAV	54.2%	8.4%
	Control	45.8%	

As illustrated in Figure 5, the difference in expected passing rates between students in the BAV and control groups is substantial for all eight dependent measures.

To put these findings into perspective, it is useful to consider interventions that are considered part of the Comprehensive School Reform Program (CSRP), a federally funded initiative that provides grants to schools to adopt proven comprehensive reform models (see Borman, Hewes, Overman, & Brown, 2003). The U.S. Department of Education (2002) defines a comprehensive school reform (CSR) model in terms of a number of criteria, many of which center on the research supporting the program's effect on student achievement. According to the meta-analysis by Borman and his colleagues, the average effect of 29 CSRP models in terms of the metric presented in the last column of Figure 3 is 7.5 percent (see Technical Note 6 for a discussion). Comparing this general finding with the values in the last column of Figure 3 indicates that BAV had a greater effect than the average for all eight dependent measures. This said, it is important to realize that the studies reviewed by Borman and colleagues typically involved standardized achievement tests, whereas this BAV evaluation study used assessments specifically designed for the study. When "curriculum-specific" assessments (such as those employed in this study) are

used, the estimated effect sizes are typically much larger than those estimated using standardized achievement tests. It must also be noted that the meta-analysis by Borman and colleagues identified a number of studies within their set of 1,111 that exhibited much larger effect sizes than the average effect size of 7.5 percent.

Another interesting comparison between the finding in this evaluation study and those found in studies of CSR models involves the issue of cost. Because teachers in this study's experimental group were involved in a two-day training only, this might be considered a relatively short intervention that would not involve great cost to a school or district. In contrast, among the 29 CSR models reviewed by Borman and his colleagues, first-year (start-up) personnel costs (e.g., for training and new hires) **per school** ranged from a low of $0 to a high of $208,361, with a median cost of $13,023. First-year nonpersonnel costs (e.g., for materials and equipment) ranged from $14,585 to $780,000 per school, with a median cost of $72,926.

Effects at Different Grade Levels

This study's second question addressed the differential effect of BAV at various grade levels. Within the fixed-effects ANCOVA design used in the study, this issue is addressed in the interaction effect for the treatment condition and grade levels. (See Technical Note 7 for a discussion.) Briefly, within the context of this study, an interaction effect indicates whether the differences between BAV and control group means are different across the grade levels. For all eight dependent measures, the interaction effect was significant at the .05 level or greater. This implies that the pattern of differences between BAV and control means was not the same from grade level to grade level. Figure 6 depicts some of the findings regarding these interaction effects.

Figure 6: Group with Greater Mean at Various Grade Levels for Dependent Measures

	9	7	6	5	4	3	2	1
Total for Written Response	**BAV**	**BAV**	**BAV**	C	**BAV**	**BAV**	**BAV**	**BAV**
Total for Multiple-Choice Response	C	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**
Reading Written	**BAV**	**BAV**	**BAV**	C	**BAV**	**BAV**	**BAV**	**BAV**
Reading Multiple-Choice	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**
Math Written	**BAV**	**BAV**	**BAV**	**BAV**	C	**BAV**	**BAV**	**BAV**
Math Multiple-Choice	**BAV**	**BAV**	**BAV**	C	**BAV**	**BAV**	**BAV**	**BAV**
Science Written	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**
Science Multiple-Choice	C	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**	**BAV**

Note: For grades K and 8, data with which to construct a comparison were not available for the experimental or control group.

For each of the eight grade levels for which experimental and control means could be computed, Figure 6 identifies which group (BAV or control) had the greater mean. In the 64 cases where data were available to compare BAV with control means, BAV had a greater mean in 58 instances or about 91 percent of the cases. One might interpret these findings as an indication that, with a few exceptions, BAV students exhibited greater means across all grade levels for all dependent measures. However, given the significant interaction effects mentioned earlier, these differences were not uniform from grade level to grade level for a given dependent measure.

Conclusions

This evaluation study addressed two basic issues: (1) whether the BAV program exhibited a positive effect on students' abilities to read and comprehend subject-area content, and (2) whether such an effect was similar from grade level to grade level. Relative to the first issue, the BAV program exhibited a statistically significant positive effect for all eight dependent measures. Relative to the second issue, students in the BAV classes exhibited greater mean scores about 91 percent of the time for grades 1, 2, 3, 4, 5, 6, 7, and 9. However, the differences between BAV and control means were not uniform from grade level to grade level within a given dependent measure.

This report addresses the findings of the evaluation study in broad terms only. Subsequent reports will address more specific features of the study.

Technical Notes

Technical Note 1: Some independent variables can be manipulated. In this case, the experimental/control variable can be manipulated in that subjects might or might not be exposed to the BAV program. Other independent variables are not manipulated but represent factors that might have a relationship of interest with the dependent variable. The second independent variable in this study—grade level—is this type of variable. The dependent variable is hypothesized to be influenced by the independent variables. In this case, the hypothesis was that the experimental/control condition and the grade level of students might have an effect on the dependent variables. The dependent variables in this study were students' abilities to read and understand mathematics information, science information, and general literacy information. The aggregated scores for these three subject areas in two item formats (multiple-choice and constructed-response) were also considered to be dependent variables.

Technical Note 2: Cronbach's coefficient alpha is generally considered an estimate of the internal consistency of the items in a test. In general, alpha is a function of the interitem correlations and the number of items on a test. The post-test reliabilities reported in Figure 2 are higher than the pre-test reliabilities probably because the post-tests had more items than the pre-tests.

Technical Note 3: Independent variables can be analyzed as fixed effects or as random effects. When independent variables are analyzed as random effects, the intent is to generalize results beyond the boundaries of the independent variables employed in the study. When fixed effects are employed, one typically does not generalize beyond the boundaries of the independent variables in the study. In this case, the BAV program was contrasted with the approaches to vocabulary instruction employed by the control teachers. Since the BAV versus control condition was considered a fixed effect, generalizations should be made only to the instruction used by experimental and control teachers involved in this study.

Technical Note 4: With an ANCOVA design, the covariate is used to predict students' performance on the post-test. The residual scores for each student are then used as the dependent measure. To illustrate, consider the multiple-choice scores for mathematics. Using ANCOVA, each student's post-test score was predicted using the student's pre-test score. The difference between the predicted post-test score and the observed post-test score, which is referred to as the residual score, was then computed for each student. This score represents the part of each student's post-test score that can not be predicted from the pre-test score. Theoretically, use of residual scores based on pre-test predictions is an attempt to equate all students on the dependent measure prior to execution of the intervention—in this case the BAV program. Berk (2004), however, warns that in actual practice this interpretation is not always appropriate.

Technical Note 5: The Binomial Effect Size Display (BESD) was used to compute the values in the last column of Figure 3. According to Rosenthal and Rubin (1982), the BESD is a translation of the Pearson product moment correlation (*r*) into a situation in which the independent and dependent variables are considered dichotomous. In this study, the independent variable is thought of as two distinct groups—the experimental group (or BAV group) and the control group. Additionally, the dependent variable in this study (performance on the eight dependent measures, all of which are continuous variables) is thought of in terms of two distinct groups—those who

pass and those who *fail*. Ideally, the *passing* students and the *failing* students constitute normal distributions. With both independent and dependent variables dichotomized, the proportion or percentage of subjects from the two groups represented by the independent variable who would be expected to pass and fail the test represented by the dependent measure can be computed. Generally, a passing rate of .50 is assumed for the dependent variable. Given these assumptions, the BESD is easily computed from *r* by simply dividing *r* by 2 and then adding to and subtracting from .50. For example, if *r* = .50, then .50 divided by 2 is .25. The proportion of subjects in the experimental group who would be expected to pass the test represented by the dependent variable would be the expected passing rate (i.e., .50) plus one-half of *r* or .75 in this case. The proportion of subjects who would be expected to fail the test represented by the dependent variable would be the expected passing rate minus one-half of *r* or .25. In this study, partial *eta*, as opposed to *eta*, was used as the estimate of *r* for each dependent measure. *Eta* is appropriate for balanced designs in which there is no confounding of effects. In such cases, *eta* is defined as the square root of the sum of squares for the effect—experimental/control condition—divided by the total corrected sum of squares (i.e. [SS effect/SS corrected total]^.5). In contrast, the formula for partial *eta* is the square root of the sum of squares for the effect divided by the sum of squares for the effect and the sum of squares for error (i.e. [SS effect/ (SS effect + SS error)]^.5).

Technical Note 6: The Borman (2003) study examined the results of 1,111 experimental/control comparisons across 29 CSR models. The average effect size was found to be .15 (Cohen's *d*), which translates roughly into a correlation of .075. Using the BESD as described in Technical Note 5, one can estimate that the expected passing rate of students in schools who use the CSR models would be 53.75 percent, as compared to an expected passing rate of 46.25 percent in schools where CSR models were not used.

Technical Note 7: In this study, the interaction effect addressed the pattern of differences between BAV and control means at different grade levels. If an interaction is deemed not to be significant, a common interpretation is that the pattern of differences is similar from grade level to grade level. In this study the interaction effect was significant at the .05 level or higher for all dependent measures, indicating that the pattern of differences between BAV and control means was not the same from grade to grade. However, this does not mean that control means were greater than experimental means. As illustrated in Figure 6, the BAV mean was greater than the control mean in about 91 percent of the cases. However, the significant interaction effect does indicate that the difference between BAV and control means might have been quite substantial at one grade level but relatively small at another grade level.

References

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks, CA: Sage Publications.

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2003). Comprehensive school reform achievement: A meta-analysis. *Review of Educational Research, 73*(2), 125–230.

Harlow, L. L., Mulaik, A. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

Lou, Y., Abrami, P. C., Spence, J. C., Poulsen, C., Chambers, B., & d'Appollonia, S. (1996). Within-class grouping: A meta-analysis. *Review of Educational Research, 66*(4), 423–458.

Marzano, R. J. (2004). *Building background knowledge for academic achievement.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J., & Pickering, D. J. (2005). *Building academic vocabulary: Teacher's manual.* Alexandria, VA: Association for Supervision and Curriculum Development.

Osborne, J. W. (2003). Effect sizes and disattenuation of correlation and regression coefficients: Lessons from educational psychology. *Practical Assessment, Research and Evaluation, 8*(11). [Online]. Retrieved September 28, 2005, from http://PAREonline.net/getvn.asp?v=8&n=11.

Rosenthal, R., & Rubin, D. B. (1982). A simple general purpose display of magnitude and experimental effect. *Journal of Educational Psychology, 74*, 166-169.

U.S. Department of Education. (2002). *Guidance on the comprehensive school reform program.* [Online]. Retrieved September 28, 2005, from www.ed.gov/offices/chiefltr/html (search for comprehensive school reform).