Generating Knowledge of Academic Language among Urban Middle School Students

Catherine E. Snow, Joshua F. Lawrence, Claire White

Background/context. The reading comprehension of post-primary grade students, in particular those attending urban schools, is a matter of recurrent concern. Performance of 8th graders on the NAEP reading assessment, for example, shows that 74% of all students perform at or above the basic (grade-appropriate) level, whereas only 60% of students in the large central urban districts perform at that level. Ethnic and language minority students in these districts perform even less well than Caucasian and native English-speaking students. While the specific reading challenges faced by students performing below basic level are no doubt heterogeneous, a ubiquitous issue mentioned by their teachers and confirmed by assessment is their limited vocabularies. It is not surprising that the many language minority students in urban districts show gaps in English vocabulary, but even native English speakers may fail to develop rich vocabularies if they have a history of low reading ability, limited comprehension, and low investment of time in reading, because much sophisticated vocabulary is acquired through reading (Anderson, Wilson, & Fielding, 1988; Stanovich, 1986).

Purpose / objective / research question / focus of study. Our purpose was to understand: 1) how well students participating in a vocabulary program learned target words relative to other students; 2) if treatment effects were better for language minority (LM) or English only (EO) students, and; 3) if improved vocabulary predicted improved scores on a state-mandated standardized assessment.

Setting. Five treatment and three comparison middle schools in Boston Public Schools (BPS) in Massachusetts. Roughly half the students in the schools were LM students.

Population / Participants / Subjects. This is a quasi-experimental study in which academic word-learning by students in five schools implementing the Word Generation program was compared to academic word-learning by students in three schools within the same system that did not choose to implement the program. Because the implementing schools were those that volunteered for the program, selection effects must be taken into account in interpreting the findings.

Participants and Setting

Schools. Word Generation was implemented in five schools during the 2007-2008 academic year, three middle schools and two K-8 schools in which only the 6th-8th grades used the program. Two schools, the Reilley and the Westfield, were completing their second year of implementation in 2007-2008, while the Mystic, Occidental, and Mercer Schools launched Word Generation in fall 2007 (pseudonyms are used for all schools). Demographics of the Word Generation and comparison schools reflect BPS more broadly, with a high incidence of poverty (ranging from a low of 79% to a high of 91% students receiving free or reduced-price lunch). BPS is characterized in general by rather high levels of special education designation, and all the schools shared this feature (between 16% and 33% of students with IEPs). A very high proportion of students at these schools come from second language homes, with percentages ranging from 32% to 70% across the schools. Four of the treatment schools offered Sheltered English Immersion (SEI) services to their limited English proficient (LEP) students; all students

enrolled in these sheltered classrooms (who represented between 6% and 26% of their school populations) received the Word Generation curriculum, albeit with modifications such as extended time and translation of key concepts.

The comparison schools looked somewhat less disadvantaged as a group than the intervention schools, and their average scores on the state accountability assessment at the start of this study were higher (mean of 45% failing in the comparison schools, compared to 56% in the treatment schools). This is not surprising; the schools volunteered to participate in the intervention; and those with lower scores were more likely to show an interest.

The five implementing schools participated in professional development activities to varying degrees, because of difficulties scheduling and organizing the required meetings. For example, the Mercer received only one brief PD session, whereas the Occidental participated in a four-day summer institute, received eight hours of PD prior to launch, and engaged in biweekly cross-grade school-site sessions devoted to feedback on and previewing of the materials and activities, with support from the Word Generation team at several of those sessions.

Students. Both pre- and post-test data were available on 697 6th, 7th, and 8th grade students in five treatment schools and 319 in three comparison schools. All students in the treatment schools received the intervention; those represented in this data set had completed usable test forms at both pre- and post-test. There were 349 girls and 348 boys in the treatment schools, and 162 girls and 157 boys in comparison schools. Of these, 438 were classified as LM (parents reported preferring to receive materials in a language other than English), 287 in treatment schools and 151 in comparison schools. The vast majority of students in both treatment and comparison schools were low-income.

Intervention / Program / Practice. Word Generation is a 24-week-long sequence of topics of current interest, each associated with five all-purpose academic words, and prescribed activities related to math, science, and social studies. The basic sequence of Word Generation activities was the following: On Monday a brief text in which the five target words were embedded was read by the students and teacher together, then discussed using guiding comprehension questions; this text presented arguments on both sides of some difficult controversy or dilemma. Then the five target words were highlighted and provided with student-friendly context-related definitions. This activity typically occurred in the English Language Arts classroom. On Tuesday, Wednesday, and Thursday, in an order determined by each school, the math, social studies, and science teachers respectively implemented activities provided for them, each of which embedded the same five target words. The math teacher assigned one or two problems related in content to the dilemma of the week; the format of these problems was modelled on the state math assessment. Math teachers then discussed the content as well as the math procedures. The science teacher presented a new text that focused on science content related to the dilemma of the week; students filled in target words left blank in the text, before the class discussed the text. The social studies teacher organized a debate about the dilemma of the week in one of several possible formats (fishbowl, pairs, whole class, four corners, etc.). On Friday, the students were asked to write a 'taking a stand' essay about the dilemma.

Various aspects of the Word Generation design respond to the local conditions in the district for which it was originally developed. Most 6th-8th graders in BPS attend separate middle schools where content area instruction is departmentalized, and teacher planning time built into the school schedule typically occurs within departments, limiting the opportunities for teachers to share information about student progress or curricular emphases across those departmental

boundaries. Transcending the boundaries to recruit participation by all the teachers in vocabulary teaching was one goal of the Word Generation design.

In addition, the extensively articulated state and district curriculum standards, as well as district pacing guides for math, science, and social studies, limited the classroom time available for focusing on vocabulary or on topics not explicitly included in the standards. Thus, to secure collaboration from the District leadership and the teachers, we agreed to design activities that could be completed in 15 min per day (thus taking only 15 min **per week** from math, science, or social studies). Furthermore, each school implementing the program had considerable leeway to decide on scheduling (which group of teachers was responsible for which day(s) of the week) and on extent of use. For example, one school excluded Structured English Immersion students during the first year of implmentation, but included them subsequently. The five schools reported on here used the program with all students in grades 6-8.

Research Design: This was a quasi-experimental evaluation of an intervention implemented at the school level. Five schools participated in the intervention and there were three comparison schools. Effects sizes were calculated by comparing the improvement students made from pre to post on a vocabulary assessment in treatment schools with improvement made in comparison schools. Secondary analyses examined language status as a moderator of word learning and word learning as a mediator of improved reading comprehension on a state mandated English language arts assessment.

Data Collection and Analysis: The efficacy of the intervention was assessed using a 48-item multiple choice test that randomly sampled two of the five words taught each week. The vocabulary assessment was not completed by all students in the time available. Because items at the end of the assessment had particularly low rates of completion, we dropped the last four items from our analysis of both pre- and post-test. The reliability of the test with the 40 items that remained was acceptable (Cronbach's alpha = .876).

This instrument was administered to students in all the treatment schools in October 2007, before the introduction of Word Generation materials. Because of difficulty recruiting the comparison schools, the pretest was not administered there until January. The post-test (identical to the pre-test except for the order of items) was administered in all the schools in late May. Because of the unfortunate disparity in interval between pre- and post-testing in the two groups of schools, we present analyses in terms of words learned per month as well as total words learned.

In addition to this curriculum-based assessment, we had access for most of the students to scores on the Massachusetts Comprehensive Assessment System (MCAS) ELA scores for spring 2008. Additionally, we had Group Reading and Diagnostic Evaluation (GRADE; Williams, 2000) for both spring and fall for a selection of students in all comparison (n = 133) and treatment (n = 256) schools. These scores were provided by the district for all the students for whom data were available. The decision to adminster the assessment was made at the school and classroom level. Thus, while these data are far from complete, we have no reason to think that there was a particular sampling bias across the schools.

Findings / Results: Descriptive statistics show that students in the Word Generation program learned approximately the number of words that differentiated 8th from 6th graders on the pretest—in other words, participation in 20-22 weeks of the curriculum was equivalent to two

years of incidental learning. Unfortunately, the relative improvements in the Word Generation schools will be exaggerated by the differences in timing of the pretest. In order to account for the differences in test administration times, the pre to post improvement in all schools was divided by the number of months between the pre and post test administration: the average improvement per months in the treatment schools was greater than that in the comparison schools. The average effect size on the researcher developed vocabulary assessment in the treatments schools was 0.49 (controlling for the improvement ascertained in the comparison schools).

Regression analysis was used to determine if participation in Word Generation predicted improved vocabulary outcomes, controlling for the pretest. Gender was a significant predictor of word learning (β = -0.052, $p < 0.007$), as was treatment ($\beta = 0.166$, $p < 0.001$). Language status (LM versus EO) was not a significant predictor, but the interaction of treatment and language status was at the margin of significance ($p = 0.055$), and including the interaction improved the overall model. Interestingly, student pretest vocabulary did not interact with treatment in predicting posttest scores. We split the data set to investigate the home language variable more closely. The first set of regressions used pretests and gender to predict post-test scores in the comparison schools (r squared = 0.62) and Word Generation schools (r squared = 0.64). In Word Generation schools LM status predicted improved vocabulary (β = -0.053, p = .022), but it was not a significant predictor in comparison schools.

In order to determine whether participation in Word Generation had any relationship to performance on the MCAS, a regression model was fit with MCAS scores in April, 2008 as the outcome, using gender, treatment status, pre-test and post-test scores as predictors. We added an interaction term to see if post-test scores interacted with treatment in predicting MCAS scores (controlling for pretest scores). The interaction term was significant (β = .21, p = 0.01) and its inclusion improved the model.

We further explored the interaction between treatment and vocabulary improvement by splitting the data and refitting the models to data from the treatment and comparison school separately. The fitted model for comparison school data did not predict MCAS achievement (R Square = 0.41) as well as the fitted model for the treatment school data (R Square = 0.49). In the Word Generation schools student post-test scores ($\beta = 0.527$, $p < .001$) were much stronger predictors of MCAS achievement than pre-test scores were ($\beta = 0.201$, $p < .001$), perhaps because the post-test scores captured not only target vocabulary knowledge at the end of the year, but also level of student participation in the Word Generation program.

Unfortunately, these analyses do not control for baseline reading achievement data, which were available only for a subset of students in our sample (n = 389). For that subgroup, we used fall standardized reading comprehension scores (on the GRADE) as a covariate to determine if the relation between improved vocabulary and MCAS persisted even when controlling for overall reading levels. Results demonstrate both that the GRADE is a strong predictor of spring MCAS scores ($\beta = 0.750$, $p < .001$) and that the interaction between treatment and improvement persists in the model controlling for GRADE. Split file analysis demonstrated the familiar pattern, with vocabulary improvement predicting MCAS scores for student in the treatment schools ($\beta = 0.151$, $p < .001$) but not for students in the comparison schools. GRADE scores were also used to determine if better readers learned words more efficiently than less able readers. Results demonstrate that GRADE baseline scores did not predict word learning and that there was no significant interaction between treatment and baseline reading achievement as measured on the GRADE.

Conclusions: The results of this initial trial of a novel approach to teaching academic language and vocabulary are promising. Students in schools implementing the program learned more of the targeted words than students in comparison schools, even though the latter group performed at a higher level at the start. Language minority students benefited more strongly than EO students, and improvement on the curriculum-specific assessment predicted performance on the state ELA assessment. Although the design of this study precludes making strong causal inferences, these preliminary results are encouraging. In particular, though the significant differences in the language demographics of different Word Generation schools makes it difficult to disentangle effect of student language status and school treatment effects, the LM-EO differences in word learning were replicated within one school. This analysis suggests that confounding effects of school-level effectiveness do not explain the faster word learning of LM students. Instead, we may need to contemplate the possibility that these students were benefiting from effective, engaging, vocabulary-focused pedagogy.

It is of interest to compare the effect size obtained with the Word Generation curriculum to that obtained in other vocabulary interventions. A similarly structured intervention, the Vocabulary Improvement Program (Carlo et al., 2004), obtained an effect size of .50. The Stahl and Fairbanks (1986) meta-analysis of vocabulary curricula reviewed studies with effect sizes ranging as high as 2 under short-term laboratory-teaching conditions, and as low as 0 under more authentic educational conditions. Thus, while Word Generation is not just a vocabulary intervention, and by design did not try to teach large numbers of words, its impact on students compares well with that of other successful programs.

It is particularly encouraging that post test scores on the Word Generation assessments strongly related to performance on the state accountability assessment. One might assume this reflects the coincidence that the words taught also occurred on the state test. However, this simple explanation is undermined by the absence of a similarly strong relationship in the treatment schools. Furthermore, while improvement in the Word Generation schools was significant, it was still modest – about four words out of forty tested. That translates into only about 12 words out of the 120 taught, which can hardly by itself explain a lot of variance on a long and challenging ELA assessment. Rather, we think it likely that improvement on our curriculum-based assessment represents an index of exposure to the Word Generation curriculum – a curriculum that taught new content, deep reading and comprehension skills, discussion, argumentation, and writing. Since the Massachusetts test is a relatively challenging one (arguably the best aligned with the NAEP of all the state assessments – McBeath, Reyes, & Ehrlander, 2007), performance on the MCAS is more likely to be related to those complex skills than to specific word knowledge.

Appendices

Appendix A References

Carlo, M., August, D., McLaughlin, B., Snow, C., Dressler, C., Lippman, D., et al. (2004). Closing the gap: Addressing the vocabulary needs of English-language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.

McBeath, J., Reyes, M.E., Ehrlander, M.F. (2007). *Education reform in the American states.* Charlotte, NC: IAP.

Stahl, S., & Fairbanks, M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research, 56*(1), 72-110.

Enhancing social studies vocabulary and comprehension for 7th grade English language learners: Findings from two Experimental studies

Sharon Vaughn, Leticia R. Martinez, Colleen K. Reutebuch, Coleen D. Carlson,
Sylvia L. Thompson and David J. Franci

Background/context. Adolescent ELLs who lack academic English language knowledge and who demonstrate low literacy levels are at risk for academic failure in content area classes (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006). Research regarding evidenced-based instruction for ELL adolescents in content learning is lacking. Our goal was to address the need for research in this area by examining the effects of an enhanced social studies instruction designed specifically for students who are ELLs that would benefit all students.

Purpose / objective / research question / focus of study. We identified instructional practices associated with improved outcomes for ELLs: (1) research-based vocabulary and concept instruction, (2) the use of media to build comprehension and concept knowledge, (3) the use of graphic organizers, and (4) structured peer-pairings. The purpose of our two studies was to examine the efficacy of incorporating instructional practices associated with improved outcomes into middle-school social studies instruction as a means of enhancing vocabulary knowledge and comprehension for English language learners. Our primary research question was: How does a multi-component instructional routine developed to enhance effective outcomes for English language learners and provided by classroom social studies teachers influence students' outcomes in vocabulary and comprehension?

Setting. Two middle schools from two districts in central Texas with large numbers of ELs.

Population / Participants / Subjects. Two different non-overlapping samples of classes of 7th grade students (N=381 and N=507) were randomly assigned at the classroom (i.e., section) level to a social studies intervention or to business as usual comparison groups. The number of sections assigned to treatment was seven and nine in experiments one and two, respectively.

Participants and Setting: Experiment 1

Schools. Participants were drawn from two middle schools in the same central Texas school district. Both schools were considered to have a substantial number of English language learners who were designated by the school as "Limited English Proficient (LEP)". At least 65% of the population at both schools was Latino with 11.45% of the students at one school designated as LEP and 13.80% of students at the other school as LEP. The proportion of students who qualified for the free or reduced-price lunch program ranged from 70 to 82%.

Teachers. The four participant teachers provided seventh-grade Texas History instruction to all the students in this study. Of the four teachers (two female, two male), two were first-year teachers who were certified to teach Social Studies (4-8) in the state of Texas. One of the male teachers had eight years of experience teaching social studies in secondary school settings. Additionally, one of the female teachers was Texas certified as a Generalist (4-8) and self-contained Bilingual/ESL teacher (1-8) and had six years of teaching experience. Classes were randomized to treatment and control conditions. Participating teachers, with support from research staff, implemented treatment conditions in intervention classes and continued with their typical instruction in comparison classes.

Participants and Setting: Experiment 2

Students. In the year after the completion of experiment 1, two middle schools from two districts in central Texas with large numbers of ELLs participated in experiment 2. Only one of

the schools had also participated in experiment 1. That school's LEP student population grew from 14% in year one to 20% in year two. The second school was new to the study and had 51% Latino students and 14% with a LEP status. The percentage of students who were eligible for free or reduced-price lunch was 68% at one school and 85% at the other school.

Teachers. During experiment 2, four teachers were identified by the principals as teaching social studies and participated in the study. All four teachers were males and certified to teach Social Studies (4-8) in the state of Texas. Two of the teachers were second-year teachers, one was a first-year teacher, and one had three years of teaching experience.

Intervention / Program / Practice. The treatment intervention was comprised of: a) overview and vocabulary instruction, b) the use of brief videos and purposeful discussion to build concepts, c) the use of graphic organizers and other writing activities to build comprehension and vocabulary through writing, and d) structure paired grouping. Students in the treatment classes received the intervention during their regularly scheduled seventh-grade social studies class. The intervention was implemented for 50 minutes a day, five days a week for approximately nine to twelve weeks. The number of lessons was the same across teachers and studies but interruptions in school schedule extended the number of weeks it took to complete the intervention. The researcher-designed lessons were used by teachers and included all the aforementioned intervention components Lesson plans identified the core subject matter and the "big ideas" that the students needed to learn in their social studies course as well as guided the teachers on the use of specific instructional practices to convey the subject matter. These practices were designed to enhance students' understanding of social studies content and of expository text by giving all their students opportunities to learn and use the vocabulary, concepts, big ideas, and issues associated with social studies. The lesson plans were not meant to be a script for teachers, but rather a guide for how to build in the study's strategies and materials. The unit lessons were designed around one or two central ideas that served as organizing concepts to help the teacher focus the events and ideas in each unit. Every lesson was organized similarly to encourage the teacher to develop a routine for the intervention.

Research Design. Two experimental studies in two successive school years with non-overlapping samples were conducted. Classes were randomly assigned to treatment and control conditions. The advantage to this design is that teachers were the same for both of our conditions and students in both the treatment and business as usual conditions covered the same material over the same period of time using the same textbook providing students in each condition equal access to learning content and key vocabulary.

Data Collection and Analysis. Prior to the intervention and after its completion all students were assessed with a researcher-developed Content-based measure (CBM). The measure was designed to cover students' understanding of the content taught during a nine to twelve week period and was meant to serve as an indicator of growth in social studies learning. It resembled traditional assessments of content area classes in that it consisted of vocabulary matching items and comprehension questions. The items were developed based on content in textbook and weekly quizzes. Students in both treatment and controls covered this same content. The vocabulary section had 20 items that included definitions that had to be matched with vocabulary terms used within the context of a sentence that contained social studies information. For example, the definition, to officially give up power or territory, had to be matched to the target

word *cede* in the sentence Mexico agreed to *cede* much of its northern territory to the United States. The second part of the assessment included 10 questions asking students to identify and explain the big ideas of the social studies units taught during the instruction. For example, one comprehension item required students to explain two ways in which slaves' human rights were violated. The content represented in these big ideas was part of the instructional materials and state standards and thus the content was part of the instructional materials for both treatment and control students.

Analyses of pre-and post-vocabulary and comprehension performance were examined separately for each study. The first step of the analyses examined differences in pre-test scores as a function of group (treatment or control). The second step examined group differences in post-intervention performance as a function of treatment group controlling for pre-test measures of the outcome variable. All analyses were conducted using three-level, hierarchical linear models in HLM 6.06 (Raudenbush, Bryk, & Congdon, 2008). For all HLM analyses, we report results for fixed effects of treatment based on robust standard errors. The three-level models included variability due to students within section, sections within teachers, and teachers. In all models, treatment was entered at level-2 (i.e., at the section level). We also tested for heterogeneity of regression between Treatment and Control sections, and examined models that allowed for pre-test regressions that varied by teacher, as well as the possibility that treatment effects differed across teachers. However, due to the small number of teachers in the study, we focus here on results from models where pre-test regressions and treatment effects were constrained to be fixed across teachers. Thus, in all reported models, random effects due to teachers were limited to effects on the intercept, i.e., the average value across all sections, both treatment and control, for that teacher.

Findings / Results.

Results for Study 1

ELL and non-ELL students differed at the pre-test, although these differences were comparable for Treatment and Control sections. More importantly, pre-test scores were not different between Treatment and Control sections. A three-level HLM analysis of pre-test scores showed no differences between Treatment and Control sections for either Comprehension ($t_{(13)}$ = -0.970, p = .350) or Vocabulary ($t_{(13)}$ = 0.552, p = .590). Analysis of post-test scores using a three-level analysis of covariance revealed statistically significant differences between students in Treatment and Control sections for both Comprehension and Vocabulary. For comprehension there was a significant effect for treatment condition indicating that students in treatment sections were performing at significantly higher levels than students in control sections post-intervention ($t_{(13)}$ = 14.31, $p \leq .001$). The estimated difference in comprehension scores between treatment and control sections was 1.57 with a standard error of 0.109. Similar results were found for vocabulary, although the absolute difference between treatment and control groups was found to be somewhat larger (Treatment Effect Estimate = 2.53, s.e. = 0.629, $t_{(13)}$ = 4.026, p = .002). For both vocabulary and comprehension, treatment effects did not interact with student status as ELL or non-ELL indicating that ELL and non-ELL students benefitted equally from participation in treatment sections. Translating the above differences into effect sizes shows that the effects of the intervention were large for both vocabulary and comprehension. For comprehension, we estimated the effect size to be g = 1.12. For vocabulary, we estimated g = 0.53.

Results for Study 2

Analysis of pre-test performance scores indicated no significant, differences between treatment and control sections for either vocabulary ($t_{(15)} = 1.450$, $p = .168$) or comprehension ($t_{(15)} = 0.934$, $p = .366$. Similar to study 1, students performed more poorly on the comprehension measure at both time points than on the vocabulary assessment. However, one must be cautious in interpreting these differences between domains as the assessments were not developed in a manner that would guarantee that they are equivalent in difficulty. Results from the three-level HLM analysis with student pre-test scores as a covariate yielded results highly similar to those from study 1. As in study 1, performance at the pre-test significantly predicted student performance at the post-test for both vocabulary ($\beta = 0.515$, s.e. = 0.107, $t_{(363)} = 4.82$, $p < .001$) and comprehension ($\beta = 0.596$, s.e. = 0.030, $t_{(361)} = 19.46$, $p < .001$). More importantly, students in treatment sections again outperformed those in control sections on both the comprehension and vocabulary measures (Comprehension: $\beta = 1.09$, s.e. = 0.403, $t_{(15)} = 2.71$, $p = .016$; Vocabulary: $\beta = 1.94$, s.e. = 0.550, $t_{(15)} = 3.53$, $p = .003$). As in study 1, treatment effects did not interact with students' ELL status for either outcome, indicating that both ELL and non-ELL students benefited equally from being assigned to sections that were randomly assigned to the treatment condition. Expressing the treatment – control differences as effect sizes using *g* as in study 1, we find that effect sizes are overall somewhat smaller, but again are classified as large or moderate, and are somewhat larger for comprehension than for vocabulary. In the case of vocabulary, there was some suggestion that pre-test regressions might differ across teachers and that treatment effects might differ across teachers. However, because significance tests for random effects may be misleading when the number of sampling units is small, in this case $n_{teachers} = 4$, we have focused above on the average treatment difference between treatment and control sections, averaged across teachers from the three-level HLM model that constrained the regression effect for the pre-test to be the same for all teachers.

Conclusions. Although this intervention was developed to address the instructional and language needs of ELLs, the students who were not limited English proficient in the intervention classes also benefited. Students who were limited English proficient outperformed their counterparts in the comparison condition on both the vocabulary and comprehension measures. When both the target group (English language learners) and their classmates benefit from an intervention or practice, it meets the criteria for universal design. This finding is particularly relevant for teachers who have both ELLs and non-LEP students in their classrooms and who may be concerned about the possible detrimental effect for other students of instruction that targets ELLs. If effective instructional practices for ELLs also benefit non-ELLs, teachers have a strong rationale for implementing the instructional practice. Furthermore, ELLs in the comparison condition made the least gains and lagged behind all other groups on both the vocabulary and content comprehension measures, providing further support for interventions such as the one in this study to alter the course for ELLs.

Appendices

Appendix A. References.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & Rivera, H. (2006). *Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions (Under cooperative agreement grant S283B050034 for U.S. Department of Education).* Portsmouth, NH: RMC Corporation, Center on Instruction, Center on Instruction. Available online at http://www.centeroninstruction.org/files/ELL1-interventions.pdf.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2008). HLM 6 (Version 6.06). Lincolnwood, IL: Scientific Software International.

The Impact of an Instructional Intervention on the Science and Language Learning of Middle Grade English Language Learners

Diane August, Lee Branum-Martin, Elsa Cardenas-Hagan, and David Francis

Background/context: Current national educational policy embodied by the No Child Left Behind Act requires that all students, including English language learners, meet high standards in language arts, as well as science and math. While expectations for content area achievement are high, findings from recent international and national studies underscore the importance of improving the literacy and science achievement of middle grades students, including English-language learners (ELLs) (Rampey, Dion, & Donahue, 2009; National Center for Education Statistics, n.d). While the need is great to improve instruction for these students, the research base is very thin compared with that for English proficient students (August & Shanahan, 2008). Our goal in this paper was to address the need for research in this area by examining the effects of an enhanced science curriculum designed specifically for students who were English-language learners (ELLs)

Purpose / objective / research question / focus of study: The goal of this study was to assess the effectiveness of an intervention--Quality English and Science Teaching (QuEST)--designed to develop the science knowledge and academic language of middle grades English language learners studying science in their second language and their English-proficient classmates. Both English language learners and English-proficient students are incorporated in our study design, and we have tested explicitly for interactions of language status with experimental and traditional forms of instruction. An overriding principle in our research was to make science instruction effective for both English language learners and English-proficient students because these two groups of students are often together in the same classrooms in the middle grades. Thus, the interventions we developed used what we know about high-quality science instruction for students in the middle grades as a starting point but made modifications to ensure ELLs understood the content of the lessons delivered in English and concurrently developed their English proficiency. For this reason, we also drew on research about the role of English language proficiency, learning in a second language, and knowledge acquired in the first language, in this case Spanish, to tailor the intervention to meet the language and literacy needs of English language learners.

Setting: Five middle schools located in a large, high poverty district in the Rio Grande Valley with a high percentage of Latino English-language learners.

Population / Participants / Subjects: Participants in this study included 890 students; 562 were English-language-learners and 328 were English proficient. The research involved forty sections[1] of students receiving their science instruction from one of ten teachers at the five middle schools. Each teacher contributed four sections, two of which were randomly assigned to treatment and two of which were assigned to typical instruction. Class sizes of the four sections for each teacher were comparable in size within and between teachers, ranging from a minimum of 11 to 13 to a maximum of 29 to 32 for treatment and control, respectively

Intervention / Program / Practice: In the district, the science curriculum at the sixth grade level consisted of the Prentice Hall textbook and workbook, and district developed labs aligned with text book content and state and district standards. The intervention, called Project QuEST, was comprised of two components that were not present in the district at the time of the intervention:

[1] At the middle grades level, a section is a classroom of students. In this district, middle grades teachers were generally assigned six sections of students a day.

instructional materials and professional development. Both components sought to ensure that both English-language learners and English proficient students acquired high levels of science knowledge and concurrently developed language and literacy skills. District and state standards were used to set science, language, and literacy goals. The intervention materials included a teacher guide and instructional charts, a student guide and instructional charts, and supplies for hands-on science activities. The intervention materials and instructional practices built on a highly rated inquiry approach to teaching science to monolingual English speakers developed by the Biological Science Curriculum Study (BSCS)[2] The approach puts a premium on hands-on experimentation that aids students in building their own understanding of new concepts.

Additionally, the intervention required that teachers use scaffolding techniques that previous research has shown to foster English language learners' understanding of academic content and has enhanced their language and literacy development (August & Shanahan, 2008). First, visuals were consistently used in science lessons, including illustrations of vocabulary concepts and graphic organizers. Second, students were given a preview of the activities they would be conduct to ensure that they understood the goals and procedures. Third, explicit vocabulary instruction of both general and discipline-specific vocabulary was emphasized. Fourth, ELLs were paired with English proficient students who served as language models. Professional development included three training sessions and ongoing weekly mentoring. Teachers learned how encourage expression of students' own ideas, build upon information students provided and experiences they have had, and guide students to increasingly sophisticated levels of understanding and language.

Research Design: The intervention was implemented in 20 sixth grade science classrooms taught by 10 teachers in five middle schools, while the districts standard curriculum was implemented in 20 sixth grade science classrooms taught by these same 10 teachers. Teachers taught two of their science sections using the QUEST materials and strategies for a period of 9 weeks, and taught two other sections of science using the district standard curriculum over the same time period. For each teacher, the four sections were randomly assigned to Quest and to the district curriculum so that each teacher taught two sections under each condition. The sample of students included English-language learners, former English-language learners, and fluent English-speaking students.

Treatment effects were tested separately for science knowledge and vocabulary using analysis of covariance, with the analogous pre-test serving as the covariate. Analyses included fixed effects of treatment assignment and the covariate, and random effects for section and teacher. Treatment effects were tested at the level of the section, which was the unit of assignment. All models were fit in HLM 6.06 (Raudenbush, Bryk, & Congdon, 2008).

Data Collection and Analysis: All students in the 40 science sections were assessed using the vocabulary and passage comprehension subtests of the GRADE assessment, form A, level M prior to the onset of treatment as a safeguard against possible unhappy randomization. In addition, students were assessed in English for science and vocabulary knowledge using researcher-developed assessments aligned with the district's sixth grade science curriculum covered during the nine weeks the intervention was in place, namely Living Systems and the

[2] BSCS is a nonprofit corporation that endeavors to improve all students' understanding of science and technology by developing exemplary curricular materials, supporting their widespread and effective use, providing professional development, and conducting research and evaluation studies.

Environment. The items for the science tests were selected from items released from past state science tests as well as from an item bank available from the publisher of the text book both treatment and control classrooms were using. The researcher developed vocabulary measure consisted of both the general academic and discipline specific words that appeared in the textbook. Reliability estimates based on coefficient alpha indicate that the assessments functioned as desired. A fidelity/quality of instruction observation instrument was developed by the research team and used to document whether and how well the specific components of the intervention were being implemented in the treatment classrooms, the activities that were taking place in the control classrooms, and the general quality of classroom instruction. The same instrument was used in both treatment and control classrooms. All student assessments were group administered by trained research assistants prior to or following each curricular unit. Each teacher was observed teaching treatment and control sections by trained observers for the sake of rating fidelity to treatment and quality of instruction.

To examine the effects of treatment in this cluster-randomized trial, we fit separate three-level, multi-level analysis of covariance models for the vocabulary and science outcomes, using the corresponding pre-test as the covariate. We also examined evidence for treatment differences on the pre-tests using a three-level analysis of variance model. In both the analysis of pre-tests and post-tests, the levels of the model corresponded to students, sections, and teachers, with students nested within sections, and sections nested within teacher, and treatment assigned at the section level. Thus, the effects of treatment are tested within teacher and averaged across teachers. In this sense, treatment is a repeated, or "within subjects," factor when viewed from the teacher level, and a "between subjects" factor when viewed from the section level. The intercept was allowed to vary randomly at each level.

Findings / Results: On average, the QuEST intervention produced positive gains in performance for students, regardless of their status as English Language Learners or native speakers of English. In addition, the QuEST intervention produced positive effects for both Science and Vocabulary outcomes for students as reflected in the curriculum based measures, which reflected the material being taught in both the treatment and control sections. Standardized effect sizes for the covariate adjusted means were in the small to moderate range for Science (.15 to .24) for the entire sample, as well as for the ELL students when examined separately (g = .16 to .25), depending on whether pre-test or post-test standard deviation is used as the baseline for the effect size computation. Results for vocabulary were somewhat more favorable, but still in this same general range, although more toward the moderate end (g = .28 to .37) for the entire sample, and for the ELL students alone (g = .26 to .37). To put these effect sizes in perspective, at g = .25, if schools could achieve this return for each of the three years of middle school, the net effect over three years would be a gain of .95 standard deviation units (1.25^3), or almost a full standard deviation of improvement over treatment as usual. These gains would be added to the annual gains due to instruction, which in the present study produced gains of only.27 standard deviations for Vocabulary for native speakers of English, and gains of .75 standard deviation units for Vocabulary for ELL students. Annual Science gains were somewhat larger, namely .88 standard deviation units for native speakers of English, and 1.01 standard deviation units for ELL students. These gains are estimated from the change observed in students in the control sections relative to the pre-test standard deviation. Thus, treatment gains in Vocabulary from three years of QuEST would be roughly double what native speakers of English would be expected to gain over that same time frame in traditional instruction, while ELL students would gain better than

four years of vocabulary growth in three years. Both native speakers of English and ELL students would gain in three years what might have been expected over four years of traditional instruction. These estimates of gain assume only that the effect size of .25 could be maintained in each of the three years of Middle School instruction. That is, they assume no compounding of content are or vocabulary knowledge as students acquire more knowledge. In so far as effect sizes for QuEST were comparable for ELL and native English speakers, this assumption appears reasonable for the time being, although one could easily argue that gains in language should lead to a compounding of effects on the acquisition of other language skills. If the QuEST strategies could be adapted to other content areas with similar efficacy, one could obviously expect a compounding effect with respect to content area knowledge and vocabulary gains in so far as effects would be observed across multiple domains of knowledge acquisition, and words learned in one content area that have general purpose utility, would obviously transfer to other content areas and not need to be relearned in that other context. Whether such gains are possible and could be obtained across multiple content areas is a matter of speculation at this point. However, it seems clear, that gains of .25 standard deviations, if sustainable year over year can result in a very significant shift in the achievement and language distributions for both ELL students and students who are native speakers of English.

Conclusions: Overall, it appears that the implementation of QuEST improves the quality of teachers' science instruction and raises student performance on curriculum based measures of Vocabulary and Science. Project QuEST differed from the practices in the control classrooms by making alterations to accommodate the needs of ELLs and build on their strengths. Consistent with the literature on effective L2 instruction, the content was made clear to students through the use of visuals, modeling, and ongoing discussion. Additionally, students' English oral proficiency was developed in the context of science instruction through explicit vocabulary instruction, guided reading, and partnering with classmates who were more English proficient.

Project QuEST makes an important contribution to the field in that there is very little research that explores whether enhancements to traditional practices are necessary or improve the traditional versions, and importantly, whether modifications to traditional practices to make them more effective with English-language learners also make them more effective with monolingual English students. To be optimal, ESL-enhanced instructional practices must enhance the learning of English-language learners in the classroom AND must be no less effective than traditional methods of instruction for monolingual English students. Because ELL students are often placed in classrooms with native speakers of English it is critical that the development of instructional methods to specifically benefit the EL students in mixed classrooms, cannot disadvantage those students in the same classrooms who are not identified as EL students. The present study of QuEST shows that such instructional improvements are not merely theoretically possible, they can be achieved in real school settings with actual Middle School science teachers. While much work remains to be done, to our knowledge this study represents the first such demonstration in a randomized controlled experiment that gains in content area knowledge in Science and gains in Vocabulary are possible for both ELL and English proficient students using a common approach to instruction that is designed to be optimal for the EL students. Of all research published between 1982 and 2009 in English in settings where English is the main medium of science instruction in elementary and secondary schools this is only published experimental study we know of that has found significant intervention effects for both ELLs and English proficient students in science knowledge and science vocabulary.

Appendices

Appendix A. References

August, D & Shanahan, T. (Eds.) (2008). *Developing Reading and Writing in Second-Language Learners.* NY: Routledge.

Rampey, B. D., Dione, G.S., & Donahue, P.L. (2009). *NAEP 2008 Trends in Academic Progress* (NCES 2009-479). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Retrieved May 28, 2009 from http/nces.ed.gov/pubsearch/pubsinfo.asp?pubid+2009479.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2008). HLM 6 (Version 6.06). Lincolnwood, IL: Scientific Software International.

National Center for Education Statistics. (n.d.). *National Assessment of Educational Progress*. Retrieved December 15, 2008 from *http://nationsreportcard.gov/science_2005/s0115.asp?tab_id=tab2&subtab_id=Tab_1&printver=#chart*