Running head: The poor get richer

The goal of this study was to test the impacts of a brief discussion-based, vocabulary-focused intervention on students' knowledge of taught vocabulary, general vocabulary, and reading comprehension. The program being evaluated, Word Generation, involves students in a variety of reading, writing, and discussion activities, all designed to offer contexts for using target vocabulary in an academic register (Schleppegrell, 2001; Author, 2009b). We know that vocabulary is strongly related to reading comprehension (Freebody & Anderson, 1983a, 1983b; Gough & Tunmer, 1986; Paris, 2005; RAND Reading Study Group, 2002) and that vocabulary interventions find effects on comprehension in texts that include the taught words (e.g., McKeown, Beck, Omanson, & Perfetti, 1983; McKeown, Beck, Omanson & Pople, 1985). Thus, the Word Generation program focuses on high frequency academic words that are required when reading secondary school texts across disciplines (Townsend, Filippini, Collins, Biancarosa, 2012). We also consider the possibility that the contexts that support students' exposure to and use of academic words in discussion and writing are likely to support general vocabulary and reading development.

Word Generation was originally developed as a product of a collaborative partnership between the Strategic Education Research Partnership (SERP) and the Boston Public Schools (see www.serpinstitute.org). The goal of the partnership was established by the district, whose leadership team had identified middle school literacy outcomes as a persistent problem of practice. Intensive reading-focused instruction was determined to be too resource-intensive, so the research-practice partnership developed a program focused on teaching all-purpose academic vocabulary – a domain that the teachers identified as problematic for students and as an impediment to reading success for many of them. Teacher input from the inception of the program insured that Word Generation fit with the on-the-ground needs of teachers and administrators, a feature that has led to its adoption and use in fifty states and a dozen

countries. Findings from an initial quasi-experimental study of the program (Author, 2009) were encouraging, as was the response of practitioners, who reported high levels of student interest and engagement. In follow-up studies we found that gains for participating students were still evident even a year after the end of instruction (Author, 2012; Author, 2014). In the first year of an IES-funded randomized trial the program had significant impact on taught academic vocabulary, but not on standardized measures (Author, 2015). These studies examined treatment-by-student profile interactions, but were not powered to explore interactions at the school level. The current study analyses data from more than 8000 students attending 44 schools in three districts. This is the largest study of the program to date, and allows us to examine student and school-level interactions.

Vocabulary Instruction

Vocabulary has long been recognized as an important skill for developing readers (Freebody & Anderson, 1983b) and as an outcome that can be improved through targeted intervention. The National Reading Panel (National Institute of Child Health and Human Development [NICHD], 2000) identified 47 vocabulary studies with reliable experimental or quasi-experimental results. Even though vocabulary is increasingly predictive of reading comprehension outcomes as children age (Author, 2007), and students with strong word identification and fluency skills can struggle if they don't have comparable vocabulary skills (Buly & Valencia, 2002), only a few studies have evaluated vocabulary interventions designed for secondary students in urban schools. Successful vocabulary programs share some key design features. They target high-leverage academic vocabulary items, such as those found in the *academic word list* (Coxhead, 2000). The target words are presented in text so that students have the opportunity to infer something about their meanings (Author, 2004; Fukkink, Blok, & de Glopper, 2001). Exposures to target words across contexts are reinforced with learner-friendly definitions that help students establish the words' abstract

meanings (Bolger, Balass, Landen, & Perfetti, 2008). The words are presented multiple times in varying local semantic contexts, and students are given multiple opportunities to learn them (e.g., McKeown, Beck, Omanson, & Perfetti, 1983; Frishkoff, Perfetti, Collins-Thompson, 2011). Students are explicitly taught word-analysis strategies, such as how to leverage knowledge of etymology, morphology, or cognates to determine meanings of newly encountered words (Lesaux, Kieffer, Faller, & Kelley, 2010; see Ford-Connors and Paratore, 2014 for a systematic review).

<<< **Table 1** – vocabulary programs>>>

The Word Generation Program

Despite sharing core principles with other adolescent vocabulary interventions, Word Generation has some unique features (Table 1). It takes relatively little time to implement the entire program (roughly 30 hours a year). It is taught by teams of teachers, so most teachers implementing the program only allocate 7 ½ hours to the program across the entire school year (more program details are provided below in the Methods section). A related feature is that each daily lesson only takes about 15 minutes. This has important consequences for teachers, since they can implement Word Generation lesson and still teach curricular material in the same period even if they are not teaching in a school with extended scheduling. Implementation of Word Generation does not require intensive training. Each summer we host several summer institutes that are usually attended by a cohort of 3 – 5 teachers from each implementing school, and there may be some limited follow up at the school site. Some schools send no teachers to the summer training. The level of professional development that we report in this study and elsewhere is roughly consistent with what districts implementing the program on their own can provide with the resources available online. Since the materials are available for downloading at no cost and since only limited professional development is required, , we believe that it is a relatively cost effective program (Levin, 1988).

Our goal is to understand the impact of the program in the low-cost version that is being adopted in schools nationally, rather than in ideal conditions. To understand how the program was actually implemented during the intervention year we conducted a total of 482 classroom observations in all Word Generation (n = 271) and comparison (n = 211) schools. We found that discussion was richer in ELA and social studies than math or science classes, and found that one of our districts (district 3) had stronger average classroom discussion than the other two. During the same observations we looked for evidence of Word Generation program implementation in randomly sampled treatment classrooms. Results of these observations suggested widespread program implementation. However, we also asked schools to provide Word Generation workbooks for a randomly drawn list of 25% of students from each grade. We coded the student notebooks for evidence of work being done in them. Virtually all notebooks evidenced some level of student program participation; participation rates were strong at the beginning of the year but dropped precipitously towards the end of the year (as we have seen in previous studies, Authors, 2011). Overall, there was only evidence of program participation on about 40% of the workbook pages coded.

We anticipated being able to help students improve their knowledge of targeted vocabulary, but we also wanted to test for gains in general vocabulary and reading comprehension. We had only cautious expectations for impacts on these distal measures: even relatively intensive literacy-focused curricular enhancements often generate only small effects after the primary grades (Wanzek, Vaughn, Scammacca, Metz, Murray, Roberts, & Danielson, 2013). For example, neither the Enhanced Reading Opportunities Study (Somers, Corrin, Stepanik, Salinger, Levin & Zmach, 2010) nor any of the eight evaluations funded under the Striving Readers initiative (Abt Associates, 2010) showed robust or educationally significant impacts. Impacts are limited in particular when instructional interventions are

evaluated at large scale – across multiple districts and schools – as was the case in the Enhanced Reading Opportunities study, and in the work reported here.

On the other hand, given the well understood relationship between vocabulary and reading comprehension, there is reason to think improved knowledge of target words might result in improved reading comprehension for some students. Specifically, we know that children who know more vocabulary learn newly encountered words better (Fukkink, Blok, & de Glopper, 2001). We also know that knowledge of academic words correlates with core academic language skills, reading comprehension, and word reading fluency (Uccelli, Galloway, Barr, Meneses and Dobbs, 2015). These studies suggest students who learn target words in the Word Generation program may be able to leverage this knowledge in the learning of other general words not directly taught.

Word Generation might also provide an alternate pathway to improved comprehension by introducing richer discussion into some classrooms. An analysis of data from the first year of the randomized trial demonstrated that classes implementing the program has higher quality classroom discussion, and that program treatment effects on word learning were mediated by higher quality classroom talk (Authors, 2015). We also know that discussion is associated with improved reading comprehension (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009), so it is possible that program participation might result in improve reading outcomes at the school level, especially for schools where academically productive talk is uncommon.

Thus, our specific research questions were:

1) What is the effect of participation in WG on three outcomes (a curriculum-based test of word knowledge, a standardized test of word knowledge, a standardized test of reading comprehension)?

2) Do all students benefit equally from participation in the program? Are there differences according to baseline knowledge of target words?
3) Do all schools benefit equally from program participation? Do students benefit differentially according to the baseline score of the school they attend?

Methods

School District Settings

Three districts participated in this evaluation study. Two large urban districts in the northeast US had joined the evaluation study the year before that reported on here. One large urban district in the west was implementing the program for the first time during the year we report on. We include second-year and first-year implementing schools in a single analysis because they were all using the same version of the curricular materials during the year of the study. The curricular content for the previous year's implementers had been different, though the teaching practices emphasized were the same.

Recruitment started with district leaders, who then invited their school-level leadership teams to participate in the study. To be considered, teams had to accept the prospect of being randomly assigned either to implement the program the following fall ("phase 1 schools") or only after two years ("phase 2 schools"). Before conducting the randomization, we created composite scores for each school by taking a linear combination of the following covariates: percent minority, percent free and reduced lunch, percent English language learners, and prior mean achievement using the state accountability data. We ranked the schools on the composite within district. Each sequential pair of schools formed a dyad within which randomization occurred, in order to maximize comparability of treatment and control schools. This strategy minimizes group differences and reduces the potential for unhappy randomization (Raudenbush, Martinez, & Spybrook, 2007). The school-level scores were also

used as covariates in the analysis, which served to reduce the intra-class correlation (ICC) and increase power for detecting treatment effects (Bloom, Richburg-Hayes & Black, 2007).

School Settings

<<< **Table 2** - school descriptives>>>

A total of 44 schools participated in this study.[i] The first and second data column in Table 1 describes which district each school is in and the school code used to identify it in this study. There were no differences between schools in the treatment and control condition in number of students or the percentage of students receiving free or reduced lunch (data columns 3 and 4, Table 1). Because the districts were in different states, we could not directly compare schools across districts on state assessments of reading or language arts proficiency except by considering percentage of students who met state-established proficiency benchmarks (data columns 6 - 8). The differences in each district between proficiency levels in treatment and control schools were low (a 9%, -5% and -3% difference in districts 1, 2 and 3 respectively). The right hand side of Table 1 shows the number of valid contributions (i.e. students who contributed both pre and posttest scores) by grade level in each school, the total contributions from each school, and the total number of within school grade level clusters that contributed data from more than 10 students to the study. For instance, school 1 had 2 grade level teams (in grades 6 and 8) that had more than 10 students complete the curricular (WG) vocabulary pre- and posttests. There were also students who completed only the first wave of the WG vocabulary (26.2 %), general vocabulary (16.7%), or reading comprehension assessments (19.0%). The students who completed both waves of data did better at each wave than students who only completed one wave.[ii]

The Word Generation program is implemented by cross-content teams of teachers. For the most part the teaching teams were organized at the grade level within schools and in

all schools except a couple of the largest ones a single team served all the students within a grade. We therefore treat grade-level teams in each school as teaching teams. Our analysis assumes that students are nested in grade-level teaching teams, which are nested in schools. Since teaching teams with small numbers of students cannot implement the discussion and debate components of the program as designed, we only included teaching teams in the analysis if they had more than ten students who contributed data at the pretest. This resulted in 84 students (less than 0.001% of the sample) being excluded from the analysis.[iii]

The Intervention

Each Word Generation weekly curricular unit is organized around an engaging civic, moral, or philosophical dilemma, e.g., *What is the function of school? Should students be required to wear uniforms? Should undocumented immigrants be granted amnesty?* A brief introductory reading passage explains the importance of the issue and provides a few arguments in support of different positions on it. Five all-purpose academic words (such as *confer, implement,* or *priority)* embedded in the passage are called out for special instructional attention. The text and target words are introduced on Mondays in a shared-read-aloud and discussion context, usually by the ELA teacher. On subsequent days the math and science teachers lead activities around authentic math and science problems that are related to the content of that week's dilemma and incorporate the week's target words. On Thursdays the social studies teacher leads a classroom discussion or debate on the dilemma. On Fridays students write a 'taking a stand' paragraph, in which they argue their own point of view on the dilemma, incorporating the information accumulated across the week to defend their claim. Each of the activities is designed to support small-group, or whole-class discussion, or both, providing opportunities for the students to produce the newly taught words and to formulate and defend arguments. More information about the Word Generation approach and freely

downloadable copies of the curricular materials can be found at http://wordgen.serpmedia.org/.

Optimal implementation of Word Generation, with its relatively novel focus on classroom discussion and on teaching academic language in science, math, and social studies, requires teachers to implement some new practices. Schools that had been randomly assigned to implement Word Generation in each of three participating districts were invited to send teams of teachers to a 3-day Summer Institute prior to first implementation. For fiscal and practical reasons, few schools sent full teaching teams; instead, one or two "Word Generation leads" from each school participated. Leads were study-recruited school staff who agreed to be the primary study liaisons for a modest stipend. These were usually individuals working at their schools as literacy coaches, assistant principals, or in other instructional leadership roles. Follow-up distance-learning and on-site professional development sessions were offered and provided when requested by schools or groups of schools. In some cases leads organized and held their own school-site sessions. At a minimum, teachers implementing the program received an introduction that lasted a few hours. At a maximum, they had a total of several days' preparation to use the program as well as support sessions throughout the year.

Measures

We describe each of the vocabulary and reading measures below:

WG vocabulary knowledge. The research team developed a multiple-choice vocabulary synonym task to assess students' knowledge of the taught academic words at pretest and posttest. Academic words (such as *relevant, obtain,* and *invoked*) are rarely encountered in everyday speech but are frequently used across academic genres. The Educator's Word Frequency Guide (Zeno, Ivens, Millard, & Duvvuri, 1995) provides standardized measures of word frequency in a corpus of reading materials (with over 17 millions words) that a typical student could encounter by their first year in college. We found

that tested Word Generation words in the curriculum used this year occurred less frequently (*M* = 137 occurrences per million words, *SD* = 183) than those on the Gates-MacGinitie test (*M* = 179 occurrences per million words, *SD* = 1048), and yet were much more widely dispersed across academic genres (*M* WG dispersion = .67, *M* Gates dispersion = .49).

Underlined target words were used in simple sentences, and students had to choose the synonym for the target word from four options. One or two Word Generation target words were selected at random from each week of the program to ensure the assessment did not only assess recently taught words. The pretest scale reliability was acceptably high for these 40 items (0.88); pretest raw scores ranged from 0 to 40 (*M* = 21.01, *SD* = 8.33).

General vocabulary knowledge. Participants completed either level 6 or level 7/9 of the Gates-MacGinitie vocabulary assessment (depending on their grade level). Assessment items presented students with a sentence or clause, which included an underlined target word. Students were required to select a synonym for the underlined word from five options. The words assessed in this test include frequently used vocabulary words, high leverage academic words, and also very rarely used words. Kuder-Richardson Formula 20 reliability coefficients were high (0.91 and 0.90 for level 6 and level 7/9 respectively). All analysis was completed with the extended scaled scores, which were developed according to Item Response Theory using the Rasch model (MacGinitie, MacGinitie, Maria, & Dreyer, 2000). The assessment uses 45 items and had high reliability in our sample (0.90). Pretest raw scores ranged from 367 to 661 (*M* = 519.33, *SD* = 36.72).

Reading Comprehension. We administered level 6 or level 7/9 of the Gates-MacGinitie reading comprehension assessment, depending on student grade level. A total of 48 multiple-choice questions were used to assess student comprehension of short reading passages. Kuder-Richardson Formula 20 reliability coefficients were high (0.92 and 0.91 for level 6 and level 7/9 respectively; Maria, Hughes, MacGinitie, MacGinitie, & Dreyer, 2007).

The extended scale scores were used in this analysis, because this score allows progress in reading to be tracked over time and across grades on a single, continuous scale. The internal reliability of the test in our sample was high (alpha = 0.91). Raw scores ranged from 361 to 643 (*M* = 522.36, *SD* = 38.39).

Analytic Data

We prepared the assessment data for analysis by creating teaching-team-centered individual scores, school-centered teaching team scores, and school-mean scores for each measure.

School-mean scores. We calculated the school-mean pretest scores in WG (academic) vocabulary (ACA_VOC _SM_W1), general vocabulary (GEN_VOC _SM_W1), and reading comprehension (READ _SM_W1) at each school using pretest scores from all students who contributed data at that wave. Across the all schools at the pretest, school mean WG vocabulary scores ranged from 14.1 to 25.7 (*M* = 18.37, *SD* = 3.05), school mean general vocabulary scores ranged from 473.41 to 537.76 (*M* = 508.90, *SD* = 13.15), and school mean reading comprehension scores ranged from 488.06 to 545.82, (*M* = 510.82, *SD* = 13.82).

School-mean-centered teaching team scores. We calculated the mean score of each grade-level team in each school at pretest and posttest. We calculated the school-mean-centered teaching team scores by finding the difference between the mean scores in each teaching team and the mean scores at each team's school at pretest. WG vocabulary pretest scores ranged from -4.64 to 4.42 (*M* = 0, *SD* = 2.24), general vocabulary pretest scores ranged from -38.81 to18.46 (*M* = 0, *SD* = 9.76), and reading comprehension scores pretest scores ranged from -4.64 to 5.12 (*M* = 0, *SD* = 2.21).

Teaching-team-centered individual scores. We calculated the teaching-team-centered score of each student on WG vocabulary (ACA_VOC_TTC_W1), general vocabulary (GEN_VOC_TTC_W1), and reading comprehension (READ_TTC_W1) by

finding the difference between each student's score and the mean non-centered score of each student's teaching team. Teaching-team-centered individual scores range from -25.79 to 21.54 (M = 0, SD = 7.55) in WG vocabulary, -150.97 to 141.95 (M = 0, SD = 33.37) in general vocabulary, and -25.79 to 21.54 (M = 0, SD = 13.73) in reading comprehension.

Treatment. TREAT is a categorical variable indicating if a school was participating in the Word Generation program (TREAT = 1) or not (TREAT = 0).

Grade level. Each district provided us with information about each student's grade level. We used these data to create two variables to allow a non-linear parameterization of differences across grade levels.

School percent free and reduced lunch scores. We established the percentage of students eligible for free and reduced lunch from publicly available sources and used it to create the school-level covariate PERCENT_FARM. PERCENT_FARM values ranged from moderate (PERCENT_FARM = 49) to quite high (PERCENT_FARM = 96) in our sample of urban schools.

School-level proficiency scores. The districts that participated in this study were in different states, and each state used its own assessment to determine student reading proficiency. It was not possible for us to scale across the state achievement measures. Instead, we used data about the percentage of students who reached proficiency as defined by local state standards at each grade level in each school to create a school-level covariate. School level percentage of students who scored proficient (PERCENT_PPROF) ranged from 24% to 100%.

School district. We used the district codes to create dummy variables used to specify which district a student was in.

Analysis

We used two methods to determine the effect of participation in the Word Generation program on student knowledge of taught academic WG words, general vocabulary, and reading comprehension. First, we examined pretest-to-posttest differences in treatment and control schools at the school level and calculated treatment effect sizes. Secondly, we fit a series of hierarchical linear models (HLM), which accounted for how individual students are nested in grade levels within schools. Student achievement data were collected at the student level, but we were primarily interested in understanding the treatment effect at the school level. We used multilevel modeling techniques to estimate the effect size of participation in the Word Generation program. These techniques allow us to appropriately account for the nested structure of the data in both our primary analyses and our secondary analyses, which explore heterogeneous treatment effects across and within schools (Raudenbush & Bryk, 2002). All mixed models were fit with the *xtmixed* command in STATA version 12 using full-information maximum likelihood estimation.

We explored Word Generation treatment effects with models based on the following:

$$ACA\_VOC\_W2_{ijk} = \beta_{0jk} + \beta_{1ijk}\text{ACA\_VOC\_TTC\_W1}_{ijk} + \varepsilon_{ijk} \quad (1)$$

where $ACA\_VOC\_W2_{ijk}$ is the posttest score of child *i* in teaching team *j* at school *k*; β_{1ijk} is the difference in the outcome associated with a one point difference in the child's teaching-team-centered score on the same measure at the beginning of the year ($\text{ACA\_VOC\_TTC\_W1}_{ijk}$); and ε_{ijk} is the residual error term for child *i* in teaching team *j* at school *k*. The intercept β_{0jk} is modeled using the level-2 model:

$$\beta_{0jk} = \gamma_{11k} + \gamma_{12}\text{ACA\_VOC\_TM\_W1}_{jk} + \gamma_{13}GRADE7_{jk} + \gamma_{14}GRADE8_{jk} + \xi_{jk} \quad (2)$$

where γ_{12} is the difference in the outcome associated with a one-point difference in school-mean-centered grade-level scores at pretest ($\text{ACA\_VOC\_TM\_W1}_{jk}$);

$\gamma_{12}GRADE7_{jk}$ and $\gamma_{13}GRADE8_{jk}$ are the differences in achievement between students in grades 6, 7 and 8 respectively, controlling for all other achievement variables; and ξ_{jk} is the variance component associated with teaching teams. The intercept γ_{11k} is modeled at the school level by:

$$\gamma_{11k} = \lambda_{11}ACA\_VOC\_SM\_W1_k + \lambda_{12}\text{DISTRICT}\_2_k + \lambda_{13}\text{DISTRICT}\_3_k + \lambda_{14}\text{TREAT}_k + \xi_k \quad (3)$$

.

where λ_{11} parameterizes the predicted posttest differences associated with a one-point difference in school mean pretest scores at school k ($ACA\_VOC\_SM\_W1_k$); $\lambda_{12}DISTRICT\_2_k$ represents differences in the outcome associated with being in district 2 over district 1; $\lambda_{13}DISTRICT\_3_k$ represents differences in the outcome associated with being in district 3 over district 1 controlling for all pretest variables; $\lambda_{14}TREAT_k$ represents the effect of school participation in the Word Generation program; and ξ_k represents the unexplained variance at the school level. All of our hierarchical linear models use a three-level nested structure to predict student-level outcomes from grade-level-centered individual scores, school-centered teaching team scores, and school-mean scores, and specify a nested variance structure.

In addition to determining the treatment effect of the program on key student outcomes, to answer our second research question we explore which students, teaching teams, and schools benefited more from program participation using interaction terms and visual displays of the data.

Results

<< **Table 3** – Effect Sizes>>>

We calculated a preliminary estimate of the Word Generation program effect by comparing improvement from pre- to posttest on our three outcomes in both the treatment and control schools[iv] (see Table 3). Treatment effects estimated at the school level should be used

in the numerator of the effect size equation if that is the level of randomization. However, this difference should always be expressed relative to a measure of the student-level, within-group pooled standard deviation[v] in the denominator (What Works Clearinghouse, 2008). These preliminary estimates do not account for nesting of student data correctly or use covariate data. They suggest a small treatment effect on taught WG vocabulary (*Hedge's g* = 0.130), a negligible negative effect on general vocabulary (*Hedge's g* = -0.015), and a small treatment effect on reading comprehension (*Hedge's g* = 0.061). Estimates from fitting HLM models allow us to account for nesting of students and teaching teams, and control for a host of pretest covariates.

<<<**Table 4** – HLM >>>

Table 4 presents nine HLM models. The first three (Models 1A, 1B and 1C) predict student WG vocabulary scores. Models 2A, 2B and 2C predict student general vocabulary scores. The last three models (3A, 3B and 3C) predict student reading comprehension. The first model in each series is the most basic: it predicts student posttest scores from achievement at pretest on the predicted measures and treatment status. Thus, Model 1A predicts student posttest WG vocabulary from student, teaching-team and school-mean academic vocabulary and treatment but does not control for other achievement measures or explore interactions. The second model in each series (1B, 2B and 3B) is similar to the first in each series, but also includes controls for each of the other achievement measures (at the individual, teaching-team and school levels), grade level, and other covariates. We examined the coefficient associated with treatment in these models to answer research question 1. The third model in each series (1C, 2C and 3C) explores interactions; we will examine these models when we turn to research question 2.

RQ1. What is the effect of participation in WG on three outcomes (a curriculum-based test of word knowledge, a standardized test of word knowledge, a standardized test of reading comprehension)?

Model 1A predicts students' WG vocabulary at posttest from school mean scores (β = 1.095, $p < 0.001$), teaching-team centered scores ($\beta = 0.614$, $p < 0.001$), teaching-team centered individual pretest scores ($\beta = 0.779$, $p < 0.001$), and treatment ($\beta = 0.931$, $p < 0.05$). The treatment effect calculated from this estimate (0.931 / 8.33 = 0.11) is slightly higher than that at which we arrived arithmetically in Table 3. Model 1B is similar, but controls for general vocabulary, reading comprehension, grade level, district, school-level percent free and reduced lunch, and school-level percent of students who score proficient on the state-mandated test. Unsurprisingly, this model is a better fit (deviance = 49680.26) than Model 1A (deviance = 53259.18). With controls for WG vocabulary, general vocabulary and reading comprehension (at the school and teaching-team level), it is also not surprising that other school-level covariates were not significant in this model nor interactions between grade level and treatment in any of our models. The estimate of the effects of treatment on WG vocabulary ($\beta = .780$, $p < 0.05$) is smaller when including all covariates. This estimate divided by the within-group pooled standard deviation[vi] provides our best estimate of the effect of the Word Generation program on targeted WG vocabulary (*Hedge's g* = 0.094; see far right column on Table 4).

Model 2A presents the reduced model predicting general vocabulary ability from general vocabulary pretest ability at the school, teaching team, and individual level. The estimate of treatment impact for general vocabulary is small and not statistically significant (β = -0.016, p = *n.s.*). Model 2B includes school-, teaching-team-, and individual-level covariates. Parameter estimates from Model 2B show that schools with higher percentages of free and reduced lunch had lower test scores of general vocabulary than would have been

predicted from pretest scores (we discuss interactions between percent free and reduced lunch and treatment below). There was no main effect of treatment (β = -0.090, p = *n.s.*) on general vocabulary outcomes.

Model 3A predicts student reading comprehension outcomes from pretest ability at the school, teaching-team and individual level and suggests no main effect of treatment on reading comprehension (treatment β = 1.74, p = *n.s.*). Model 3B controls for a host of covariates. We find that seventh-grade students improved more in reading comprehension than would have been predicted from pretest covariates (β = 7.01, $p < 0.01$) and that schools with higher proficiency levels on state mandated tests had better reading outcomes than would have been predicted from other achievement measures alone (β = 18.40, $p < 0.01$). Treatment did not predict improved reading comprehension in this model (treatment β = 2.67, p = *n.s.*).

2) Do all students benefit equally from participation in the program? Are there differences according to baseline knowledge of target words?

In order to understand how participating in the Word Generation program might have supported students (RQ2) and schools (RQ3) with different baseline profiles, we conducted secondary analyses including interaction terms (Models 1C, 2C and 3C in Table 4). Although we were primarily interested in interactions with treatment, we first explored interactions between baseline student achievement and school achievement (coefficients are reported under the heading *School Mean by Student Interactions*). We found that the relationship between students' pretest and posttest scores in general vocabulary is stronger in schools with higher mean general vocabulary scores. (β = 0.003, $p < 0.001$; Model 2C). Similarly, the relationship between students' pretest and posttest scores in reading comprehension is stronger in schools with higher mean reading scores (β = 0.003, $p < 0.001$ Model 3C). These findings replicate the general pattern of increasing gaps in achievement between high- and low-performing students across the school years (Reardon, Valentine & Shores, 2012); this

fanning of student scores is unexpected, since regression to the mean, or convergence toward the mean, is the normal outcome.

We explored interactions with treatment at the teaching-team level and the individual level. There were no student-level pretest by treatment interactions in models predicting WG vocabulary, but interactions were significant in our models of general vocabulary (Model 2C) and reading comprehension (Model 3C). In both cases the reported interaction term is the product of the mean WG pretest vocabulary score of each student (ACA_VOC _W1) and treatment status (TREAT). We did not find any interactions between treatment and teaching-team-level baseline scores or grade level. We did find a relationship between uncentered individual baseline WG vocabulary scores and treatment in predicting general vocabulary (β = -0.175, $p < 0.01$). We found similar though less pronounced trends predicting individual standardized reading comprehension scores (β = -0.129, $p < 0.5$). In order to interpret these findings we fit linear models of general vocabulary and reading comprehension improvement indexed by baseline academic vocabulary scores in each school. The estimated impact of academic word knowledge on general vocabulary was lower in the comparison schools (WG vocabulary coefficient estimate *M* = -.186) than in the treatment schools (WG vocabulary coefficient estimate *M* = .130). The estimated impact of academic word knowledge on reading comprehension was similar in each group of schools (*M comp* = .073; M *treat* = .098).

<<<**Figure 1** >>>

<<<**Figure 2** >>>

We plotted the results of these models (Figure 1 & 2). Figure 1 plots the predicted general vocabulary improvement of students in each school by their baseline individual WG vocabulary scores. Visual inspection of the plot in control schools (left hand side of Figure 1) suggest that in most comparison schools, student with higher baseline academic vocabulary scores tended to improve on the general vocabulary measure more than those with weaker

baseline scores. This *Matthew effect* is blocked in the treatment schools, where lower performing baseline students tend to improve more than their high baseline peers (Figure 1, right hand side). Figure 2 plots the predicted reading comprehension improvement for students in each school indexed by WG vocabulary pretests. Although the HLM model suggests an interaction, and there is a modest difference between average within-school estimates of the impact of baseline WG vocabulary on reading comprehension, these plots confirm the interpretation that the individual-level interaction between baseline WG vocabulary and reading comprehension is not strong. This contrasts strongly with the school-level interactions we turn to now.

3) Do all schools benefit equally from program participation? Do students benefit differentially according to the baseline score of the school they attend?

<<< **Figure 3** – Scatter plot Vocab pre and post GATES Vocab>>>

There was no pretest by treatment interaction in predicting academic vocabulary; all schools benefited roughly equally in their average WG vocabulary regardless of where they started (Model 1C)[vii]. There were similar school-level pretest by treatment interactions in models predicting our distal outcomes, general vocabulary (Model 2C) and reading comprehension (Model 3C). In both cases, the reported interaction term is the product of the mean WG pretest vocabulary score at each school (ACA_VOC _SM_W1) and treatment status (TREAT). The estimate of this coefficient was significant in predicting student posttest general vocabulary scores (Model 1C; β = -0.899, $p < 0.05$). In order to interpret this interaction we plotted raw school-level general vocabulary improvement scores by school-level baseline WG vocabulary scores for treatment and control schools at pretest (Figure 3). This figure demonstrates that across control schools (left hand side) there is a strong positive relationship between baseline school-mean scores in WG vocabulary and pre- to posttest improvement in general vocabulary. In the Word Generation schools (plotted on the right

hand side), the relationship between initial vocabulary levels and improvement demonstrated in the control schools is disrupted.

<<<**Figure 4** – Scatter plot Comprehension pre and post>>>

The school level WG vocabulary by treatment interaction variable is also significant in Model 3C predicting reading comprehension (β = -1.58, $p < 0.001$). Again we plotted the improvement in scores (this time in reading achievement) by school-level WG vocabulary pretest (Figure 4) and found a strong relationship between school-level academic achievement and pre- to posttest reading comprehension gains in the control schools (left hand side). Although Word Generation schools did not have higher mean gains across the sample of schools, the predictive effect of low baseline WG vocabulary scores is eliminated in treatment schools. Evidently a school-level version of the Matthew effect was playing out for the control schools, i.e., poorer starting performance was associated with slower growth at the school level. In contrast, in the treatment schools with lower baseline scores, students had more opportunity to benefit from instruction: the Matthew effect was blocked. These differential results across treatment and control schools are also not consistent with regression to the mean, because regression effects would be expected to affect treatment and control schools equally due to randomization.

Discussion

The analyses reported here generated one unsurprising and two much more remarkable findings. Unsurprisingly, effects of a brief and only modestly supported vocabulary intervention program on student learning of words taught are significant but small (effect size of about 0.1). On average, students in control schools improved 1.46 points on the test of WG vocabulary, while students in treatment schools improved roughly 2.37 points. The small main treatment effect on taught vocabulary confirms the difficulty of finding big effects of programs implemented across schools and districts with varying levels of commitment to the

program and with varying quality and intensity of implementation. Much more interestingly, in control schools there is a relationship both within schools and between schools of pretest WG vocabulary knowledge to improvements in general vocabulary and reading comprehension, which is blocked in schools participating in the Word Generation program.

<< Insert Table 5 >>

<< Insert Figure 5 >>

In order to understand how we could have a main effect on academic taught words but a moderated impact on general academic vocabulary and reading comprehension, we looked at the relationship between these measures using quantile regression (analyses reported here are with the control school data only, but results are similar for the whole sample). First we regressed students' general vocabulary and reading comprehension scores on their WG vocabulary scores (Table 5; Models 1 and 2 respectively). The models are very similar; one point higher WG Vocabulary score predicted higher scores in general vocabulary ($\beta = 2.77$, $p < .001$) and reading comprehension ($\beta = 2.91$, $p < 0.001$). In model three we use quantile regression to understand this relationship for students at the 25^{th}, 50^{th}, and 75^{th} percentile in general vocabulary ability (Gould, 1993; Hao & Naiman, 2007). The relationship between WG vocabulary knowledge and general vocabulary knowledge is stronger for student with lower general vocabulary scores than it is for students with higher general knowledge. This trend is demonstrated in Figure 5 which plots the intercept, WG Vocab regression coefficients and confidence interval for students in each of nine general vocabulary quantiles. This figure makes clear both how uneven the relationship between the measures across performance bands is, and how much information is lost looking only at the OLS model. These models suggest that although all students and schools participating in the Word Generation program improve on WG vocabulary (on average), the impact is differential across students: the same

improvement in WG Vocabulary will have stronger cascading effects on low-baseline students' general vocabulary skills.

<< Insert Figure 6 >>

Model 4 in Table 5 replicates this analysis, this time regressing reading comprehension on WG Vocabulary. In this case, although we see the same general trend, it is not as pronounced. Differences across performance bands are not as stark, and confidence intervals are generally close to, or overlap with, those arrived at by OLS regression (see Figure 6). These models suggest that while WG vocab may be somewhat more predictive of reading ability in the lower quantiles, this trend is not as strong for reading comprehension as it is for general vocabulary. These findings align with our reported estimates of the individual-level academic vocabulary-by-treatment interaction in Table 4; the interaction term in the model of general vocabulary (Model 2B) was much stronger than in the model of reading comprehension (Model 3B β = -.175, p < .01).

On the other hand, we noted that the school level WG vocabulary-by-treatment interactions were stronger in the models for reading comprehension (β = -1.580, p < .001) than in models for general vocabulary (β = -0.899, p < .05). While we have not been able to interpret this finding to our full satisfaction, we believe that it is related to the emphasis in the WG program on classroom discussion and debate. There is considerable evidence that discussion, though rare in U.S. classrooms (Applebee, Langer, Nystrand, & Gamoran, 2003; Gamoran & Nystrand, 1991), is related to reading comprehension outcomes (Murphy, Wilkinson, Soter, Hennessey, & Alexander, 2009). In an analysis of data from the first year of the Word Generation randomized trial, we found that participating classrooms had higher quality discussions than control classrooms, and improved classroom discussion mediated 14% of the program impact on academic vocabulary scores (Author, *2015*). In the current study we also found strong program impact on discussion quality (Hedge's g = .377).

Unfortunately, we do not have baseline observation data of classroom discussion. There is, however, positive correlation between WG vocabulary and observed discussion quality in the 20 control schools (r = .35); not surprisingly, given the small sample size, it is not significant. This suggests that school-level measures of WG vocabulary at pretest index, to some extent, school debate and discussion. If so, WG would constitute a desirable instructional change in low baseline schools, but a distraction in high WG vocab schools where rich discussion already occurs. Both kinds of schools would achieve the narrow goal of improved WG vocabulary, but only those schools with low baseline WG vocab (indexing low discussion) *also* benefit from a programmatic school-wide emphasis on classroom discussion and debate.

Limitations

There are several important limitations of this study. First, we know that quality and intensity of implementation of any program relates to its impact on student learners; while we did not incorporate direct measures of implementation in this analysis, the inclusion of the teacher-team variable did address this issue to some extent. Second, the teachers implementing the program would have benefited from more and more intensive professional development than we were able to provide. We were thus not testing a 'gold standard' version of the treatment, but rather a real world version with highly variable levels of quality and intensity across different settings. We know that many of the schools implemented fewer than the full 24 weeks of the program, with interruptions for test preparation, state accountability testing, outings, snow days, and other circumstances.

Nonetheless, these results reflect the stark reality of U.S. urban schools: in the absence of programs designed to focus teacher attention on vocabulary, academic language, and opportunities for engaging discussion, schools characterized by poor academic achievement fail to improve student scores and so students in those schools fall farther and

farther behind. By studying the characteristics of programs that are successful in narrowing the disparities in rate of progress between higher and lower achieving schools, we may gain insights into the features of curriculum and pedagogy necessary to shrink the achievement gap.

References

Abt Associates Inc. (2010). *Summary of 2009 Striving Readers Projects*. Washington, DC: Department of Education, Office of Elementary and Secondary Education.

Applebee, A., Langer, J., Nystrand, M., & Gamoran, A. (2003). Discussion-based approaches to developing understanding: Classroom instruction and student performance in middle and high school English. *American Educational Research Journal*, *40*, 685-730.

Author. (2004).

Author. (2007).

Author. (2009a).

Author. (2009b).

Author. (2009c).

Author. (2009d).

Author. (2011).

Author. (2012).

Author. (2013).

Author. (2014).

Author. (2015).

Biemiller, A. (2007). Vocabulary development and instruction: A prerequisite for school learning. In D. Dickinson & S. Neuman (eds.), *Handbook of Early Literacy Research*, Volume 2 (41-51). New York: Guilford Press.

Bloom, H., Richburg-Hayes, L., & Black, A. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30.

Bolger, D., Balass, M., Landen, E., & Perfetti, C. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes, 45*(2), 122.

Buly, M., & Valencia, S. W. (2002). Below the bar: Profiles of students who fail state reading assessments. *Educational Evaluation and Policy Analysis, 24*(3), 219-239.

Author. (2004).

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

De La Paz, S., Ferretti, R., Wissinger, D., Yee. L., & MacArthur (2012). Adolescents' disciplinary use of evidence, argumentative strategies, and organizational structure in writing about historical controversies. *Written Communication,* 29(4), 412-454.

Ford-Connors, E., & Paratore, J. R. (2015). Vocabulary Instruction in Fifth Grade and Beyond Sources of Word Learning and Productive Contexts for Development. *Review of educational research, 85*(1), 50-91. doi: 10.3102/0034654314540943

Freebody, P., & Anderson, R. C. (1983a). Effects on text comprehension of differing proportions and locations of difficult vocabulary. *Journal of Literacy Research,* 15(3), 19-39. doi:10.1080/10862968309547487

Freebody, P. & Anderson, R.C. (1983b). Effects of vocabulary difficulty, text cohesion, and schema availability on reading comprehension. *Reading Research Quarterly, 18,* 277-294.

Frishkoff, G. A., Perfetti, C. A., & Collins-Thompson, K. (2011). Predicting robust vocabulary growth from measures of incremental learning. *Scientific Studies of Reading, 15*(1), 71-91.

Fukkink, R., Blok, H., & de Glopper, K. (2001). Deriving word meaning from written context: A multicomponential skill. *Language Learning, 51*(3), 477-496.

Gamoran, A., & Nystrand, M. (1991). Background and instructional effects on achievement in eighth-grade English and social studies. *Journal of Research on Adolescence, 1*, 277-300.

Gough, P. & Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education, 7*, 6–10.

Gould, W. (1993). Quantile regression with bootstrapped standard errors. *Stata Technical Bulletin*, *2*(9). Retrieved from http://ideas.repec.org/a/tsj/stbull/y1993v2i9sg11.1.html

Graves, M. F. (2000). A vocabulary program to complement and bolster a middle-grade comprehension program. In B. M. Taylor, M. F. Graves & P. Van den Broek (Eds.), *Reading for meaning: Fostering comprehension in the middle grades*. Newark, DE: International Reading Association.

Hao, L., & Naiman, D. Q. (2007). *Quantile regression*. Thousand Oaks, Calif: Sage Publications.

Kelley, J. G., Lesaux, N. K., Kieffer, M. J., & Faller, S. E. (2010). Effective academic vocabulary instruction in the urban middle school. *The Reading Teacher, 64*(1), 5 - 14.

Lesaux, N. K., Kieffer, M. J., Faller, S. E., & Kelley, J. G. (2010). The effectiveness and ease of implementation of an academic vocabulary intervention for linguistically diverse students in urban middle schools. *Reading Research Quarterly, 45*(2), 196-228.

Lesaux, N. K., Kieffer, M. J., Kelley, J. G., & Harris, J. R. (2014). Effects of Academic Vocabulary Instruction for Linguistically Diverse Adolescents Evidence From a Randomized Field Trial. *American Educational Research Journal, 51*(1), 1159-1194. doi: 10.3102/0002831214532165

Levin, H. M. (1988). Cost-effectiveness and educational policy. *Educational Evaluation and Policy Analysis, 10*(1), 51-69.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, N.J: Wiley.

Lively, T., August, D., & Carlo, M. (2003). *Vocabulary Improvement Program for English Language Learners and Their Classmates*. Baltimore, MD: Paul Brookes Publishing Company.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading test technical report: Forms S and T*. Chicago, IL: The Riverside Publishing Company.

Maria, K., Hughes, K. E., MacGinitie, W. H., MacGinitie, R. K., & Dreyer, L. G. (2007). *Lexile conversions for the Gates-MacGinitie reading tests, 4th Edition*. Rolling Meadows, IL: Riverside Publishing.

McKeown, M., Beck, I., Omanson, R., & Perfetti, C. (1983). The effects of long-term vocabulary instruction on reading comprehension: A replication. *Journal of Reading Behavior, 15*(1), 3-18.

McKeown, M., Beck., I., Omanson, R. & Pople, M. (1985). Some effects of the nature and frequency of vocabulary instruction on the knowledge and use of words. *Reading Research Quarterly, 20*, 522-535. doi: 10.2307/747940

Murphy, P., Wilkinson, I., Soter, A., Hennessey, M., & Alexander, J. (2009). Examining the effects of classroom discussion on students' comprehension of text: A meta-analysis. *Journal of Educational Psychology, 101*(3), 740.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly, 47*(1), 91-108.

National Institute of Child Health and Human Development [NICHD]. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: U.S.: U.S. Government Printing Office.

Paris, S. G. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, 40(2), 184-202.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R&D program in reading comprehension.* Santa Monica, CA: RAND.

Raudenbush, S. & Bryk, A. (2002). *Hierarchical linear models: Applications and data analysis methods*. New York: Sage.

Raudenbush, S., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*(1), 5.

Reardon, S., Valentine, R., & Shores, K. (2012). Patterns of literacy among U.S. students. *The Future of Children, 22*, 17-38.

Schleppegrell, M. (2001). Linguistic features of the language of schooling. *Linguistics and Education, 12*(4), 431-459.

Somers, M. A., Corrin, W., Sepanik, S., Salinger, T., Levin, J., & Zmach, C. (2010). The enhanced reading opportunities study final report: The impact of supplemental literacy courses for struggling ninth-grade readers. NCEE 2010-4021. *National Center for Education Evaluation and Regional Assistance*.

Townsend, D. (2009). Building academic vocabulary in after-school settings: Games for growth with middle school English-language learners. *Journal of Adolescent & Adult Literacy, 53*(3), 242-251.

Townsend, D., & Collins, P. (2009). Academic vocabulary and middle school English learners: An intervention study. *Reading and Writing, 22*(9), 993-1019.

Townsend, D., Filippini, A., Collins, P., & Biancarosa, G. (2012). Evidence for the importance of academic word knowledge for the academic achievement of diverse middle school students. *The elementary school journal, 112*(3), 497-518.

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., & Dobbs, C. L. (2015). Beyond Vocabulary: Exploring Cross-Disciplinary Academic-Language Proficiency and Its Association With Reading Comprehension. *Reading Research Quarterly*.

Uccelli, P., Dobbs, C. L., & Scott, J. (2013). Mastering academic language organization and stance in the persuasive writing of high school students. *Written Communication*, *30*(1), 36-62.

Wanzek, J., Vaughn, S., Scammacca, N., Metz, K., Murray, C., Roberts, G., & Danielson, L. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*, 83, 163-195.

What Works Clearinghouse. (2008). *Evidence standards for reviewing studies Technical Report*. Washington, DC: Institute for Educational Sciences. Available at http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_version1_standards.pdf.

Zeno, S., Ivens, S., Millard, R., & Duvvuri, R. (1995). *The educator's word frequency guide*. Brewster, N.Y.: Touchstone Applied Science Associates.

[i] Not all schools participated in ways that had been agreed upon with the district leaders. School 32 participated as both a control and Word Generation school, where eighth graders were in the control condition and sixth graders in the treatment condition. School 37 was assigned to TX but did not implement at all. Two other schools assigned to control (24 and 36) and one other school assigned to treatment (33) dropped out of the study and did not provide data.

[ii] In order to be sure that missing data did not unduly influence our results we replicated descriptive tables and basic models with multiply imputed (MI) datasets created using the multivariate normal model (Little and Rubin, 2002). We imputed pre and post test scores for each student with missing data on any of the three achievement measures using all of the achievement data from that student, plus information about district and grade level, as well as demographic information about the student's school (PERCENT_FARM PERCENT_ELL). Despite important developments in the field in respect to conducting MI in multilevel contexts, we were not confident in our models that imputed school-level or teaching-team-level data, so we fit basic models. Each model used the mi estimate command with xtmix in Stata on MI data sets predicting each outcome measures from each pretest, controlling for grade level (school was the grouping variable). The fully imputed data set included 11015 students in each model. We found that the models looked very similar to those that we have presented here. The coefficient for the variable associated with Word Generation participation were TREAT (WG vocabulary) = 0.972, $p = 0.15$, TREAT (general vocabulary) = 1.04, $p = n.s.$, and TREAT (reading comprehension) = 0.147, $p = n.s.$ These estimates do not properly account for student nesting, so we have more confidence in the full HLM models we present in table 3; however, these MI models align with and complement our other models and we take them to suggest that our models are not being unduly influenced by nonrandom missingness.

[iii] This resulted in 11 very small teams being eliminated from the analysis (including teams with only two or three students who contributed pre and post test data in the grade level). We fit models with different exclusion criteria at the team level and found consistent results in our estimates of treatment effects on taught words. With no limit on the number of students per teaching team group TREAT = 0.872 ($N_{students}$ = 8465, N_{teams} = 118), when we limit the sample to only teams with more than four students, TREAT = .883 ($N_{students}$ = 8459, N_{teams} = 116), when we limit the sample to only teams with more than eight students, TREAT = .954 ($N_{students}$ = 8421, N_{teams} = 110), when we limit the sample to only team that contributed more than 12 students, TREAT = 1.03 ($N_{students}$ = 8338, N_{teams} = 102).

[iv] We calculated effect sizes with the following equation:

$$\Delta = \delta_T - \delta_C = \frac{\left(\mu_{T,\,post} - \mu_{T,\,pre}\right) - \left(\mu_{C,\,post} - \mu_{C,\,pre}\right)}{\sigma\ pooled}$$

[v] We calculated pooled standard deviation with the following equation:

$$\sigma\ pooled = \sqrt{\frac{\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_1^2}{\left(n_1 + n_2 - 2\right)}}$$

[vii] In our exploratory models we also found school-mean general vocabulary by treatment interactions in predicting general vocabulary posttests (β = -.156, $p < 0.01$, model deviance = 66830.3) and reading comprehension (β = -.352, $p < 0.01$, model deviance = 71415.1). We also found an interaction between treatment and school-level percent of free and reduced lunch in predicting posttest reading comprehension (β = 0.400, $p <$ 0.01, model deviance = 71416.8). However, none of these interactions were significant when we also included interactions with baseline WG vocabulary. In each case, the interactions with baseline WG vocabulary resulted in better model fit (based on fit statistics on Table 3), as we anticipated. We present the best fitting models.