Symposium Title: Conditions for Success: Variation in Treatment Effects for a School-Based Randomized Controlled Trial of Word Generation

Organizer: Catherine Snow

Symposium Justification

Randomized controlled trials are increasingly used to evaluate promising educational programs, but the results are often not as expected, with many RCTs producing largely inconsistent effects. In recent years, two reports were prepared by experts in education and policy research on the topic of variation in program effects, commissioned by the William T. Grant Foundation (Weiss, Bloom, & Brock, 2013) and by the Institute of Education Sciences (Schochet, Puma, & Deke, 2014). These reports offer recommendations responding to increasing concern about the focus on average program effects in the face of research indicating (1) programs vary in their effectiveness, (2) a program that works for one group may not work for another, and (3) a program that is effective under certain conditions may not be effective under other conditions (Weiss, et al. 2013). To explain variation in program outcomes, we must first understand what happens when well-designed and rigorous RCTs meet the real world of school settings. In three papers, this symposium contributes to this effort and focus on identifying the conditions and mechanisms of impact variations in a widely-implemented adolescent literacy intervention program, Word Generation.

*Word Generation (WG)* is a tier-one, discussion-based program for middle school students designed to build academic literacy and academic practices through language arts, math, science, and social studies classes. The program consists of weekly units that introduce five high-utility target words through brief passages designed to spark active examination and discussion of contemporary issues. WG was designed with the understanding that promoting classroom discussion can result in particular kinds of academic benefits, such as improved word knowledge, reasoning, and expression.

Capitalizing on the data collected as a part of the IES-funded project *Catalyzing Comprehension through Discussion and Debate* (CCDD), which included a school randomized trial of WG, the three papers examine various sources and mechanisms of WG impact variation. The first paper focuses on examining the correlates of the variation in program implementation at the classroom level and student engagement with the program, and tests whether these variations are predictive of variation in the student outcomes. The second paper tests whether the WG program has differential impacts on ELL students and non-ELL students, in an effort to identify effective academic interventions that can provide leverage to advance the academically disadvantaged ELL population. The third paper reports on variation in performance on an argumentative writing measure completed by WG students and controls, examining differences in lower-level skills such as writing mechanics as well as higher-level features such as perspective-taking.
Paper 1

**Title:** Engaging early adolescents in urban schools: Variation in implementation and outcomes in an experimental trial of Word Generation

**Authors and Affiliations:**

Maria LaRusso, University of Connecticut  
Ha Yeon Kim, New York University  
Stephanie M. Jones, Harvard University  
James Kim, Harvard University  
Suzanne Donovan, Strategic Education Research Partnership (SERP)  
Catherine Snow, Harvard University

**Background:**

Randomized controlled trials have become the gold standard for evaluating educational programs, but analyses rarely reveal consistently positive effects. Researchers are recognizing the need to shift focus from average program effects toward a deeper understanding of the variation in program impacts across individuals and settings. Key to this endeavor is the systematic study of program implementation, including students’ level of engagement with and exposure to program activities. However, few program evaluation studies examine implementation in predicting outcomes, despite compelling evidence of its importance (Durlak and DuPre, 2008; Swanson, et al., 2013). In particular, it is widely known that programs are rarely implemented fully. In a review of 500 evaluations of health and education programs; no studies reported 100% implementation and few reported levels greater than 80%. Implementation also varies across sites, typically from 20% to 40%, and higher levels of implementation are associated with stronger program outcomes (Durlak and DuPre, 2008). To identify what works for students in urban public schools, it is critical that experimental studies examine program implementation and engagement, including how it varies across students, teachers and classrooms and how it relates to program outcomes.

**Research Questions:**

In this paper we examine:

1. variation in student engagement and classroom level program implementation for the Word Generation (WG) curriculum in 4th through 7th grade urban classrooms,
2. student and teacher characteristics as predictors of student engagement with WG, and
3. classroom level program implementation and student level engagement as predictors of variation in program impacts on student outcomes.

**Setting:**

As part of the IES-funded project *Catalyzing Comprehension through Discussion and Debate* (CCDD), the data for this study were collected as part of the impact evaluation of Word Generation (WG) conducted with students in 4th through 7th grades in 25 schools. In this paper...
we focus on two grade groups: 4th and 5th graders (elementary cohort), and 6th and 7th graders (middle-grades).

Population / Participants / Subjects:

Participants included 5,285 fourth to seventh graders from 23 schools (275 classrooms; 44% in treatment schools). Across grade levels, student characteristics varied as follows: 4-12% English language learners (ELLs), 15-17% classified as special education, 79-86% eligible for free or reduced lunch, 36-47% Black, 18-31% Latino/a, 23-35% White, 2-3% Asian, and 1-3% Mixed/Other.

Intervention / Program / Practice:

Word Generation (WG) is a tier-one, discussion-based program designed to build academic literacy and academic practices with a focus on controversial issues selected to be compelling to early adolescents.

Research Design:

The IES funded evaluation of WG is a school-level experimental study that included 25 schools randomized to treatment and control conditions. Of the 25 schools, 23 were K-8 schools; one middle school with an elementary feeder school was treated as a single unit in the randomization. Data were collected on the experimental sample for two years (2012-2014) although one randomized pair (2 schools) dropped out after the first year. For this paper we employ data from Year 2 when workbooks were collected and coded for all students in the treatment condition.

Data Collection and Analysis:

Student Outcomes: All participating students completed group-administered assessments in the Fall and Spring of each year, including: academic vocabulary (WG Academic Vocabulary), academic language (Core Academic Language Skills-Instrument), deep reading comprehension skills (Global Integrated Scenario-based Assessments), and social perspective taking skills (perspective positioning and articulation: Social Perspective Taking Acts Measure-Revised; administered all waves except fall of Year 2).

Student Engagement with WG Workbook Activities: Student workbooks were collected at the end of the school year. Each activity was coded for completion and a workbook completion rate was calculated by averaging the % activity completion across all units.

Classroom Program Implementation: Workbook activities also reflect the amount of curriculum covered in each classroom. The mean level of workbook completion was calculated for each classroom.
Teacher Data: A teacher survey was administered at the end of each year and included teacher self-reports of years teaching experience and burnout as assessed with Maslach Burnout Inventory (Emotional Exhaustion, Depersonalization, and Personal Accomplishment.)

Analyses were conducted using multi-level regression models with classrooms nested in schools for models predicting student engagement and school-pair fixed effects, with classroom nesting, for models predicting student outcomes.

Findings / Results:

The mean percentage of workbook activity completion was 40.2% (SD=20.7) for elementary WG classrooms was 40.2% (SD=20.7) and 31.2% (SD=20.5) for middle-grade classrooms. In elementary classrooms, student engagement with workbook activities was significantly greater for students in classrooms with more experienced teachers (coefficient=.51, p<.05) and significantly lower for students eligible for free/reduced lunch (coefficient=-2.63, p<.05), receiving special education services (coefficient=-5.98, p<.001), classified as English Language Learners (coefficient=5.65, p<.001), and in classrooms with teachers reporting higher levels of emotional exhaustion (coefficient=-4.64, p<.05). In middle-grade, student engagement was significantly greater for students in classrooms with teachers reporting a greater sense of personal accomplishment as an educator (coefficient=6.90, p<.01) and significantly lower for students receiving special education services (coefficient=-4.25, p<.001), classified as ELLs (coefficient=-6.15, p<.001), and in classrooms with more experienced teachers (coefficient=-.76, p<.01).

Student engagement was a significant predictor of vocabulary knowledge (elementary: coefficient= 11.02, p<.001, middle-grade: coefficient= 9.36, p<.001) and perspective positioning (elementary: coefficient=.27, p<.001, middle-grade: coefficient=.33, p<.05) in all grades and deep reading comprehension in elementary only (coefficient=17.96, p<.05). Controlling for student engagement, classroom level implementation significantly predicts vocabulary knowledge (elementary: coefficient=11.67, p<.001, middle-grade: coefficient=9.09, coefficient<.001) and perspective positioning (elementary: coefficient=.29, p<.01, middle-grade: coefficient=.40, p<.001) in all grades and deep reading comprehension in the elementary cohort only (coefficient=18.30, p<.001). When classroom implementation is included in the models, variation in student engagement is no longer significant, except for deep comprehension in the middle-grade cohort, which is predicted only by student engagement (coefficient=66.06, p<.05).

Conclusions:

These results provide evidence that variation in program implementation and student engagement with the curriculum are important for understanding variation in program outcomes. The findings also suggest that efforts to improve student engagement in program activities should focus on English Language Learners and students receiving special education services, as well as teachers who are experiencing higher levels of burnout.

References

Paper 2

Title: Reducing the Academic Inequalities for English Language Learners: Variation in Experimental Effects of Word Generation in High Poverty Middle Schools

Authors and Affiliations:

Ha Yeon Kim, New York University
Maria LaRusso, University of Connecticut
Stephanie M. Jones, Harvard University
Suzanne Donovan, Strategic Education Research Partnership (SERP)
Catherine Snow, Harvard University

Context:

English language learners (ELLs) lag behind their non-ELL peers in course grades and standardized test scores (Han, 2008; 2012). For this reason, many academic intervention programs have focused on improving ELLs’ academic skills, and have found encouraging evidence that such programs can provide the leverage ELLs need. However, most of these studies have been done with language-minority (LM) students, including both students who currently have less-than-proficient English skills (current-ELLs) and students from non-English speaking homes who are proficient in English (e.g. Lawrence et al., 2012; Lesaux et al., 2014). Little research focuses on current-ELLs, who may need many more years to acquire academic English skills comparable to those of their native English-speaking peers (Cummins, 1991). As this transitional period presents greater vulnerability to their learning (Suárez-Orozco et al., 2010), it is important to identify effective programs that can support current-ELLs’ learning.

Word Generation (WG) is a research-based vocabulary program for middle school students intended to teach words through brief passages designed to spark active examination and discussion of contemporary issues. Previous evaluations of the WG showed positive impacts on vocabulary for all students, including those with special education classifications (Lawrence et al., 2014) and for LM students (Lawrence et al., 2012). This study is the first to test the impacts of WG on the students who are currently designated as ELLs.

Objective:

The analysis builds on to the WG intent-to-treat impact study (authors, under review), and examines the variation in impacts by ELL status. Specifically, we test whether the ELL and non-ELL students in WG classrooms experience differential gains in performance and knowledge of vocabulary, social perspective taking skills, academic language, and reading comprehension.

Setting:

This study employs a subsample of data from the IES-funded two-year (2012-2014) randomized controlled trial (RCT) evaluating the effectiveness of the WG program. The sample of the
current study included 11 schools from three school districts located in Massachusetts. The districts include one major city serving ethnically diverse, low income students; one small city serving ethnically diverse, primarily low income students; and one suburban district serving a primarily white, low to middle income population.

Participants:

Participants included the 4,167 fourth to seventh grade students: 3,082 students from 11 schools (141 classrooms; 47% in treatment schools) in Year 1 of the study and 2,840 students from 9 schools (136 classrooms; 48% in treatment schools) in Year 2. Reflecting the characteristics of the school community, a large proportion of the participants were racial/ethnic minorities and from low-income households. Students classified as ELLs accounted for 13% of the sample in Year 1 and 12% in Year 2.

Intervention:

See symposium justification for general description of WG program. The original WG program consisted of week-long units, each of which introduced five high-utility target words drawn from the Academic Word List (Coxhead, 2000) in the context of a brief reading passage that defined a controversial issue selected to be compelling to adolescents, and that provided some evidence on each side of the issue. In the CCDD study, an enhanced version was evaluated, which differs from the original version in two major ways. First, 10-day long units incorporating the same principles were developed for use in fourth and fifth grades (WordGen Elementary). In sixth and seventh grades, the program includes 12 of the original one-week units described above, as well as 12 weeks devoted to the enhanced, more intensive units tied to curricular content standards in science and social studies (WordGen Weekly, SciGen, and SocGen).

Research Design:

Eleven schools (five school pairs) were randomized within pairs that had been matched on district, size, socio-demographic variables including percent of language minority students. One randomized pair (2 schools) of the current sample dropped out of the study in the Year 2 (2013-2014). Students were assessed in fall and spring of each year.

Data Collection and Analysis:

All participants completed the group-administered assessments of targeted outcomes, including: academic vocabulary (WG Academic Vocabulary: Lawrence et al., 2015), social perspective taking (positioning and articulation: Social Perspective Taking Acts Measure-Revised, authors, under review), academic language (Core Academic Language Skills-Instrument, Uccelli et al., 2015), and reading comprehension skills (Global Integrated Scenario-based Assessments, Sabatini et al., 2014).

In this study, the analyses are conducted with all fourth to seventh grade students. We report exploratory analyses for fourth and fifth graders and sixth and seventh graders separately. All analyses were conducted using multi-level regression models with school-pair fixed effects,
with consideration of the data where students nested within the classrooms, as presented in the equation below:

\[ Y_{ij} = \beta_{0ij} + \beta_1(ELL)_{ij} + \beta_2(FALL\ \text{STUDENT\ SCORE})_{ij} + \sum_{k=3}^{n} \beta_k (\text{STUDENT\ COVARIATES})_{kij} + \varepsilon_{ij} \]

\[ \beta_{0ij} = \gamma_{00} + \gamma_{01}(WG)_{ij} + \gamma_{02}(FALL\ \text{CLASSROOM\ MEAN})_{ij} + \gamma_{03}(GRADE)_{ij} + \sum_{n=3}^{n} \gamma_{on} (\text{SCHOOL} -\ \text{PAIR\ FIXED\ EFFECTS})_{nij} + \mu_{0j} \]

\[ \beta_{1ij} = \gamma_{10} + \gamma_{11}(WG)_{ij} \]

**Results:**

In Year 1 of Word Generation RCT trial, we found no significant variation in the WG impact by ELL status. However, the exploratory analyses suggested that fourth and fifth grade ELLs in treatment group learned more vocabulary (coefficient=6.87, \(p<.05\)) than non-ELL students in the treatment group.

In RCT Year 2, we found that ELLs had significantly higher gain than their non-ELL classmates in social perspective articulation (coefficient=.28, \(p<.05\)) and academic language (coefficient=.22, \(p<.01\)) skills. This suggests that WG intervention helps ELLs catch up their peers in these skills (Figure 1). Specifically, exploratory analyses find that fourth and fifth grade ELL students had higher increase in academic language (coefficient=.38, \(p<.01\)) compared to non-ELL students in WG classrooms; and sixth and seventh grade ELL student increased more in social perspective articulation (coefficient=.55, \(p<.01\)) and positioning (coefficient=.34, \(p<.05\)) in WG classrooms compared to non-ELL sixth and seventh graders.

**Conclusions:**

These results provide first evidence that the WG intervention can provide a boost in critical academic skills for current-ELL students, who may otherwise experience more challenges in learning due to their limited English proficiency.

**Reference**


Figure 1. RCT year 2 WG impact variation by ELL status on social perspective taking articulation skills (left) and academic language (right).
Paper 3

Title: Varying Indices of Argumentative Writing Quality and Vocabulary Learning in Word Generation

Authors and Affiliations:
Lisa Hsin, University of Alabama
Robert Selman, Harvard University
Paola Uccelli, Harvard University
Suzanne Donovan, Strategic Education Research Partnership (SERP)
Catherine Snow, Harvard University

Background
Writing persuasive arguments is a novel challenge posed to U.S. students by the Common Core State Standards, but it is only recently that attentions have focused on students’ approaches to these kinds of texts (cf. Cutler & Graham, 2008). Much work in this domain seeks to foster argumentative writing skills through engagement in argumentative dialogue, which at its best makes external the questions children should learn to ask themselves about the soundness of their own reasoning (Nussbaum & Edwards, 2011). Students’ skill at dialogic argumentation improves with experience with collaborative reasoning-style classroom discussions (e.g., Chinn et al., 2001), meant to encourage reflective judgment, but it may do so at different rates depending on student characteristics, e.g. English proficiency and gender. The Word Generation curriculum (WG) employs and richly supports classroom-based dialogic reasoning and argumentation with the primary goal of fostering deep reading comprehension. But given evidence of transfer between high-quality classroom discussion and argumentative writing skill (Reznitskaya et al. 2001; Kuhn et al., 2016), WG may enhance the quality of (pre-)adolescents’ argumentative writing as well. This study explores the impacts of WG on argumentative writing with respect to varying indices of quality, including mechanics, word choice and language complexity, and incorporation of perspective-taking.

Research Question
Does the persuasive writing of students exposed to WG differ from that of students in the control condition? Do WG impacts on persuasive writing vary by student characteristics?

Intervention
The data discussed here were collected in the context of the Catalyzing Comprehension through Discussion and Debate (CCDD) project, which developed and tested the impacts of the Word Generation curriculum, designed to foster deep reading comprehension by supporting the development of three hypothesized predictors: academic language, perspective-taking, and complex reasoning (Authors, 2016). Indeed, curriculum impact analyses have shown that its adoption leads to the learning of taught vocabulary words, but perhaps more importantly also to improvement in academic language skills, perspective taking, and deep reading comprehension as well (Authors, under review). (see symposium justification for further details)

Design & Setting
The current study focuses on the second year of the RCT trial, when the ‘taking a stand’ essay was administered in both treatment schools and in control schools (n=21 schools: a small number of schools in each condition declined to administer this measure, but virtually all randomization pairs were preserved). The data discussed here therefore represent the only extended writing samples from the CCDD project that can be used to analyze the impact of the WG curriculum.

Data Collection & Analysis
The measure was designed to be appealing and engaging to students, and conditions were created that were highly conducive to drawing out students’ best work. Students wrote a persuasive essay in response to a prompt presenting the motivation for a principal’s decision to revoke students’ access to iPads in school. Participants were instructed to “write an article for the school newspaper that argues for or against allowing iPads at your school, mak[ing] sure to give specific reasons to support your position and to convince the people who read the; article to agree with you; explain[ing] how the Principal’s decision can impact you and others, [and] discuss[ing] other things that the school community could do to solve the iPad problem.”

The analysis included the full set of iPad essay data (n=3,546) along several different lines. We used CLAN (MacWhinney, 2000) to investigate writing mechanics (length, spelling, capitalization, contraction conventions), then turned to indices of writing complexity (e.g., degree of lexical diversity present) and, finally, the use of the WG target academic language words. Given previous research indicating important changes in perspective-taking skill in WG students, we are also hand-coding (blind to condition) the essays for the exercise of that skill, according to the Social Perspective-Taking Acts Measure codebook (CCDD, 2012). The preliminary analysis includes data from a targeted subsample of 20 students per grade (grades 4-7, n=80) and the coding and analysis of the full sample will be completed by February. The schools selected for the current study had comparable numbers of ELLs and of students receiving free/reduced-price lunches. Here we focused on including as many ELLs per grade as possible, also matching the number of male and female students per grade between the treatment and the control school. We report the results of a series of exploratory regression model analyses below.

Findings
Mechanics (full sample). Significant variation was found in the length of students’ essays, as measured in words, but it did not differ systematically by condition. Controlling for essay length, we found that essays from the WG treatment group (WGT) contained significantly fewer errors in spelling ($b=-0.39, p<.001$), and significantly more contractions ($b=0.78, p<.001$), than essays from the control group (WGC). The two sets of essays did not differ in their over- or under-use of capitalization. ELLs made more spelling errors than non-ELLs, and a marginally significant interaction between condition and ELL status emerged (WGC ELLs produced fewer spelling errors than WGT ELLs).

Complexity (full sample). We operationalized morphosyntactic complexity by computing the mean utterance length per essay and counting the varieties of morphemes present in each essay. These analyses revealed that WGT students wrote significantly longer sentences ($b=0.27, p<.04$) and used more varied morphemes ($b=0.14, p<.01$) than WGC students. Critically, WGT students used more of the taught WG vocabulary words than WGC students ($b=0.21, p<.001$).
Perspective-Taking (targeted subsample). WGT students acknowledged significantly more actors in their essays than did WGC students ($b=0.83, p<.04$), and they articulated the perspectives of those actors significantly more often than did the WGC students ($b=1.02, p<.01$). Girls articulated more perspectives than boys ($b=0.87, p<.01$). We found no difference in ELLs and non-ELL students’ use of perspective-taking in their essays.

Conclusions
This study shows that WGT students’ persuasive writing shows more linguistic and perspectival sophistication than that of WGC students, even for ELLs who on many measures are shown to lag behind their English-proficient peers. Contributions of classroom discussion quality and the relationship between these outcomes and students’ reading skill will also be discussed.

References