Examining Impact of a Supplemental Vocabulary Curriculum on Upper Elementary Students' Vocabulary Acquisition Considerate of Generalizability

Bryan J. Matlen, Gary Weiser, Chun-Wei (Kevin) Huang, & Linlin Li

WestEd STEM Program

Word count: 1,000 (excluding tables and figures)

Purpose. This presentation describes the impact of *PROGRAM* [name blinded] on students' test scores related to vocabulary and reading comprehension, and how those findings may or may not generalize to different regions across the U.S. In keeping with the conference theme, we will discuss the practical significance of our findings.

Background. Vocabulary is a central component of the reading process (Baumann et al., 2003; Graves & Silverman, 2010; Hiebert et al., 2017). Among early learners, building a strong vocabulary entails learning a large number of words. A recent study of the vocabulary of elementary level core reading programs (Graves, 2015) revealed over 17,000 different words. Direct instruction is not a sufficient mechanism for learning so many words. Rather, students need to become adept at deciphering the meaning of unknown words independently. *PROGRAM* is a supplemental curriculum that explicitly attends to helping students develop skills needed to make sense of new words.

The present study assessed the impact of *PROGRAM* on fourth-grade students language achievement, addressing two research questions:

1. What is the impact of *PROGRAM* on fourth-grade students' vocabulary development and reading comprehension?
2. To what extent might findings generalize to other fourth-grade classrooms across the U.S.?

Design. To address these questions, we conducted a study with a true, group-randomized, experimental design with two cohorts. This study took place in diverse public elementary schools throughout California:109 fourth-grade classrooms were randomized to a treatment (n=57) or control (n=52) condition–assignment was blocked by percent of students enrolled in the Free/Reduced Lunch program (n=24, m=4.54 classrooms per block). Sample numbers and characteristics are presented in Tables 1 and 2, respectively. Attrition rates are presented in Table 3.

Intervention. The treatment classes implemented *PROGRAM* while control classrooms implemented their usual English Language Arts curricula. *PROGRAM* provides 15 weeks of whole-class instruction; 22 remedial, web-based lessons for students needing more practice; three web-based lessons on Spanish cognates; and three web-based lessons on idioms. Whole-class instruction includes four main instructional sections about how to derive the meaning of unknown words: examining word parts; using context clues; searching the dictionary; and combining the strategies together. *PROGRAM* is typically delivered three days a week for about 30 minutes per day.

Student Outcome Measures

(1) The *PROGRAM* Test was created by the developer. It assesses student knowledge of prefixes, suffixes, context cues, and asks them to employ these strategies to decipher the meaning of new words. Cronbach's alpha for the entire instrument ranged from 0.875 at pre-test to 0.921 at post-test.

(2) The VASE Assessment (Scott, Flinspach, Vevea, & Castaneda, 2012) measures students' capacity with grade-appropriate vocabulary in math, science, social studies, and language arts and can be used to evaluate their vocabulary growth over the school year. Results indicated that the VASE Assessment has good convergent and construct validity. The internal reliability coefficient was 0.95.

(3) The GMRT (MacGinitie, MacGinitie, Maria, & Dreyer, 2002) includes two subtests—vocabulary and comprehension. Using Kuder-Richardson Formula 20, we found reliability coefficients of .80 and .90 for the vocabulary and comprehension subtests, respectively.

Results:

To address the first research question, we analyzed student outcomes using three-level hierarchical linear models. The models accounted for the nesting of students within classrooms and classrooms within blocks, controlling for student-level covariates while estimating the average impact of the classroom's treatment assignment on student outcomes (see Table 4). The missing-indicator method (White & Thompson, 2005) was used to account for missing values on the covariates (not the outcome variables) in the impact analysis models.

Results from the *PROGRAM* test serve as a proximal measure of vocabulary development. After controlling for student baseline characteristics, we found that using *PROGRAM* corresponded to a statistically significant increase in students' post-test scores relative to the control group. The Hedges *g* value for this effect is 0.70 (95% CI=0.62-0.78) corresponding to an estimated 0.70 standard deviations higher performance on the PROGRAM test by treated students, on average, than the control students. The impact of *PROGRAM* impact on two distal measures of vocabulary knowledge and reading comprehension (VASE and GMRT, respectively) was minimal between the groups (Hedges *g*s=0.04 and -0.03, respectively). Those differences are not statistically significant at .05 level.

To address our second research question, we calculated a generalizability index based on the sample's school characteristics using the www.thegeneralizer.org (Tipton & Miller, 2015). *The generalizer* pulls publicly available data on all K-12 schools in the U.S., and – using researcher-specified moderators – estimates the generalizability of the findings (Tipton 2014). For our study, we defined the inference population as all U.S. public schools with 4th-grade classrooms (see Figure 1 for the list of moderators). The generalizability index was estimated at 0.67, indicating that the findings would generalize moderately well to the U.S. as a whole with covariate corrections. Among California (where the study took place), generalizability results were strong (>.90), approaching a level comparable to a simple random sample. Most other

states fell within levels of moderate-to-high generalizability (.05-.09), indicating that findings may generalize after correcting for covariates.

One limitation *of the generalizer* is that it is limited to comparisons of school-level characteristics while the study design occurred at the classroom level. Practical limitations (e.g., available data) inhibit perfect alignment of these factors. Nevertheless, the results of this analysis offer some value to practitioner-stakeholders who seek to assess whether their schools might benefit from *PROGRAM*.

Conclusions:

Our investigation of the *PROGRAM* showed students making substantial gains on proximal assessments but did not show differences on distal assessments. As we examine new strategies for making sense of the evidence collected about *PROGRAM* efficacy, we plan to use structural equation modeling to synthetize outcome measures (Pearl, 1995). This analysis will include data on the fidelity of *PROGRAM* implementation, providing a richer picture of the pathways by which *PROGRAM* impacts student outcomes. A second contribution is the assessment of the generalizability of these findings across the U.S. This analysis represents a pragmatic approach for addressing questions related to external validity and provides opportunities for stakeholders to assess the relevance of the findings to their local settings.

References:

August, D. L., & Shanahan, T. (2006). Developing literacy in a second language: Report of the National Literacy Panel. *Mahwah, NJ: Lawrence Erlbaum. Goldenberg, C.(2008). Teaching English Language Learners: What the research docs—and does not—say. American Educator*, *32*(2), 8-23.

Baumann, J. F., Kaméenui, E. J., & Ash, G. E. (2003). Research on vocabulary instruction: Voltaire redux. In J. Flood, D. Lapp, J. R. Squire, & J. M. Jensen (Eds.), *Handbook on research on teaching the English language arts* (2nd ed., pp. 752–785). Mahwah, NJ: Erlbaum.

Clearinghouse, W. W. (2017). Standards handbook (version 4.0). *Washington, DC: Institute of Education Sciences*

Goldenberg, C. (2013). Unlocking the Research on English Learners: What We Know—and Don't Yet Know--about Effective Instruction. *American Educator*, *37*(2), 4.

Graves, M. F. & Silverman, R. (2010). Interventions to enhance vocabulary development. Interventions to enhance vocabulary development. In R. Allington & A. McGill-Franzen (Eds.), *Handbook of reading disabilities research* (pp. 315–328). Mahwah, NY: Erlbaum.

Graves, M. F. (2016). *The vocabulary book: Learning and instruction*. Teachers College Press.

Hart, B. & Risley, T. R. (1995). *Meaningful differences in the everyday experiences of young American children*. Baltimore: P. H. Brookes.

Hiebert, E. H., Goodwin, A. P., & Cervetti, G. N. (2018). Core vocabulary: Its morphological content and presence in exemplar texts. *Reading Research Quarterly*, *53*(1), 29–49. https://doi.org/10.1002/rrq.183

MacGinitie, W.H., MacGinitie, R.K., Maria, K., & Dreyer, L.G. (2002). *Gates-MacGinitie Reading Test.* Rolling Meadows, IL: Riverside Publishing

National Governors Association Center for Best Practices, & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts: Reading foundational skills grade 4.* Retrieved from http://www.corestandards.org/ELA-Literacy/RF/4/#CCSS.ELA-Literacy.RF.4.3

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, *82*(4), 669–988.

Scott, J. A., Flinspach, S. L., Vevea, J. L., Citkowicz, M., & Castaneda, R. (2014). The VASE Assessment. Santa Cruz, CA: The Regents of the University of California.

Snow, C. E. & Kim, Y. (2007). Large problem spaces: The challenge of vocabulary for English language learners. In R. K. Wagner, A. E. Muse, & K. R. Tasnnenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 123-139). New York: Guilford Press.

Stahl, S. A. & Nagy, W. (2006). *Teaching word meanings*. Mahway, NJ: Erlbaum.

Tipton, E. (2014). How generalizable is your experiment? An index for comparing experimental samples and populations. *Journal of Educational and Behavioral Statistics*, *39*(6), 478-501.

Tipton, E. & Miller, K. (2015) Generalizer [Web-tool]. Retreived at http://www.generalizer.org.

White, I. R., & Thompson, S. G. (2005). Adjusting for partially missing baseline measurements in randomized trials. Statistics in Medicine, 24(77), 993–1007.

Wright, T. S., & Neuman, S. B. (2014). Paucity and disparity in kindergarten oral vocabulary instruction. *Journal of Literacy Research*, *46*(3), 330-357.

Table 1. Randomized and analytic sample *N*s.

	Blocks	*Schools*	*Teachers*	*Students*
	Randomized Sample			
Control	22	46	52	1296
Treatment	24	48	57	1514
	WLS Analytic Sample			
Control	22	45	51	1085
Treatment	22	46	55	1254
	VASE Analytic Sample			
Control	22	45	51	1088
Treatment	22	46	55	1245
	GMRT Analytic Sample			
Control	22	44	50	1019
Treatment	22	46	55	1237

Table 2. Teacher and student participant characteristics.

	Treatment		Control	
	M	*SD*	*M*	*SD*
		Teachers		
Years Teaching Exp	14.67	7.08	13.44	7.27
Advanced Degree	68%	--	68%	--
		Students		
SBAC Math	2417.39	76.51	2410.83	77.42
SBAC ELA	2404.70	85.26	2396.78	88.49
Female	49%	--	50%	--
FRL	81%	--	84%	--
ELL	34%	--	35%	--

Table 3. Attrition at the classroom and student levels. Student-level attrition is calculated based on classrooms that remained in the analytic sample.

Outcome	Treatment Attrition	Control Attrition	Total Attrition	Differential Attrition
	Classroom (cluster-level random assignment)			
WLS	3.51%	1.92%	2.75%	1.59%
VASE	3.51%	1.92%	2.75%	1.59%
GMRT	3.51%	3.85%	3.67%	-0.34%
	Student (sub-cluster level)			
WLS	14.15%	15.10%	14.59%	-0.95%
VASE	14.90%	14.87%	14.89%	0.03%
GMRT	15.45%	18.61%	16.91%	-3.16%

Table 4. HLM output: fixed and random effects. Missing indicator variables are omitted.

	PROGRAM	**VASE**	**GMRT**
	Coefficients (SE)		
Intercept	11.81 ***	65.87 ***	472.91 ***
	(0.67)	(2.88)	(2.60)
Treatment	5.24 ***	1.23	-1.18
	(0.45)	(2.07)	(1.44)
Pretest	11.02 ***	107.36 ***	84.60 ***
	(0.49)	(3.83)	(2.21)
Female	0.30	0.67	1.31
	(0.49)	(1.96)	(1.93)
ELL	-1.59 *	-9.41 ***	-7.17 **
	(0.63)	(2.53)	(2.53)
FRL	0.14	-6.09	-6.52 *
	(0.81)	(3.25)	(3.14)
SBAC ELA	10.75 ***	42.36 ***	44.07 ***
	(1.01)	(3.95)	(4.08)
SBAC Math	8.57 ***	12.16 **	15.35 ***
	(1.05)	(4.09)	(4.12)
CKTR	0.58	4.35	0.15
	(0.60)	(2.69)	(1.60)
	Standard Deviation		
Classroom	2.05	9.75	6.14
Block	0.57	1.97	2.40
Residual	4.46	17.63	17.08

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Figure 1. Generalizability index across all 50 U.S. states (provided by www.thegeneralizer.org). Moderators included school size, the school proportion of FRL and Hispanic students, and the district proportion of ELL students and students whose only language at home is English.