Abstract Title Page

Not included in page count.

Title:

Comparing a Program Implemented Under the Constraints of an RCT and in the Wild

Authors and Affiliations:

Denis Newman

Valeriy Lazarev

Jenna Zacamy

Empirical Education Inc.

Abstract Body

Background / Context:

This paper examines the *ecological validity* of an RCT conducted in 42 high schools to measure the impact of a content literacy program. We interpret this term in a specific way: an experiment's ecological validity can be threatened if the controls or processes put in place to assure internal or external validity result in a deviation from how the program would be implemented outside of an experiment. Gibson (1966) provides the classic example in the study of visual perception where he compares the laboratory apparatus in which the subject is stationary with how the senses are used in the wild where the subject is moving his eyes and body. In the same spirit, Cole and his colleagues (Cole, Hood, & McDermott, 1978; Newman, Griffin, & Cole, 1989) questioned the ecological validity of laboratory-based cognitive psychology as a study of the thinking and problem-solving occurring in classrooms.

Shadish, Cook, and Campbell (2002) consider ecological validity as a kind of external validity, but it may be more useful to view it more specifically in terms of the effect of the experimental design itself. Nobody will argue that an RCT is not superior to the comparison of two intact groups in terms of internal validity. But it is possible that the RCT imposes constraints and involves processes that do not occur in the wild. Since these constraints may be salient to the practitioner and stakeholders of the evaluation, it is important to recognize them in the explanation of rigorous design otherwise the value of the resulting evidence may be lost in translation.

The research reported here, takes advantage of an unusual situation where researchers were able to collect comparable data from teachers and principals in the treatment group of an RCT and from teachers and principals involved in an implementation of the same program outside of the RCT but in very comparable schools. The comparison of these contexts provides a demonstration of the differences in participant engagement and an approach to estimating the likely impact of the program when implemented in the wild.

Purpose / Objective / Research Question / Focus of Study:

The evaluation of a program to improve academic literacy, which was awarded an i3 validation grant in 2010, consisted of an RCT with 42 schools in PA and CA (final report, in preparation) and a formative study of an additional 239 schools in four states, that included PA (Zacamy, Newman, Lazarev, & Lin, 2015). The formative study, which we called the Scale-up study, was conducted primarily to help the developers improve the program and support processes.

The analysis reported here addresses the question: In the context of projects funded by programs such as i3, where a rigorous evaluation accompanies the scale-up of an innovative program, is it possible to provide systematic evidence that an RCT is under or over estimating the impact that the program has when implemented outside of the trial. A common approach to scale-up research is to consider the problems of implementation when a program is evaluated in a large scale trial. We are asking a different question: are the processes involved in recruiting, training, supporting, randomizing, and measuring for the RCT having a positive or negative impact on the outcomes. Can we identify mechanisms by which those impacts may be occurring? Can we measure the impact, for example, through quasi-experimental methods?

We can compare teachers and schools in the RCT's treatment group to other participants implementing the same program under ordinary conditions of recruitment and program implementation. Where an RCT makes use of state administered tests as a student outcome measure, the same measure can be obtained from the scale-up sites and a quasi-experiment is feasible. In our case, however, the RCT used a researcher-administered outcome measure. Where both groups are surveyed, as in the current study, differences in levels of implementation or engagement can provide a contrast with the implication of potential differential impacts. This study takes several additional steps in identifying productive mediating processes in the RCT and measuring the presence of those in the comparable scale-up schools. Our goal is to approximate the potential positive or negative impacts of the ecological invalidity of our RCT. Our goal is also to illustrate an approach to improving the validity of rigorous tests of program effectiveness.

Setting / Population / Participants / Subjects:

The i3 was conducted in high schools in five states but for greater comparability in the study reported here we focused on one state where we analyzed data from 11 of the 22 treatment schools participating in an RCT and 31 of the 239 schools implementing the same intervention in the scale-up portion of the i3 project. Recruiting for RCT and Scale-up differed in that the target number of schools was determined through a power analysis, while on the scale-up side, the target number of schools and teachers was set in the proposal in terms of the number that would be reached in the project. Recruiting was conducted over four years with four cohorts of schools and teachers although recruiting for the RCT was completed in the first two years so as to allow teachers in the treatment schools to be in the program for two years. In Scale-up, new schools were added in each cohort and in many cases, new teachers were trained in schools that had joined in the prior cohort. A constraint on the RCT schools: teachers could not have participated in program training prior to the project, whereas Scale-up schools included those where teachers may have already been familiar.

Intervention / Program / Practice:

While the program as designed was the same in the RCT and Scale-up, and for this study in the one state, teachers from both contexts were provided PD in the same summer institute, there were differences in recruitment, implementation, and support.

The basic program is an instructional framework that helps teachers support discipline-specific literacy and learning in content areas. At the center of the program is "metacognitive conversation" carried on both internally through reading and reasoning routines and externally, as teacher and students talk about reading. This takes place through extensive reading including increased in-class opportunities for students to practice reading academic texts. The framework targets learning dispositions as well as literacy skills and knowledge. The professional development is designed to change teachers' understanding of their role in adolescent literacy development and to build capacity for literacy instruction in the academic disciplines. Each teacher was offered 10 days (65 hours) of subject-specific professional development over twelve months. For both contexts, the program was modified for the i3 in order to make it more scaleable. This included a focus on developing a community of teachers in the high school by including teachers of ELA, biology and history. The developers also developed a cadre of "teacher leaders" at each school site who were expected to convene monthly team meetings to at their schools to support implementation.

Research Design:

Study is a comparison of schools from the RCT and Scale-up contexts where the outcomes are teacher self-report. We do not have a baseline for the outcome measure but have documented commonalities between the two groups. In order to increase comparability all schools were from the same state and all teachers were trained together in the same summer workshops. We address three sets of questions

1. Descriptive comparison of the two contexts. Were there significant differences in school characteristics between the treatment group of the RCT and schools participating in Scale-up?

2. Characteristics associated with successful program adoption – gain or loss in the number of participating teachers in the school. Is there a difference between RCT and Scale-up on those characteristics? Characteristics of interest are based on Zacamy, Newman, Lazarev, & Lin (2015) who undertook the analysis of characteristics associated with successful program adoption in the context of the Scale-up study. The metric was the gain or loss (GL) in the number of participating teachers by school. Covariates included schools characteristics (such as school demographics, average teacher experience, and teacher turnover), school-average teacher survey responses, and principal survey responses. The study identified a small number of variables that had a robust positive association with successful program adoption within a school. These variables include teachers' attendance of program meetings, levels of responsibility for and commitment to the success of the program at the school level.

3. Characteristics associated with better student outcomes in RCT. Are these characteristics also associated with successful program adoption? Is there a difference between RCT and Scale-up on those characteristics? These questions call for a correlational study whereby the strength of associations between student outcomes and characteristics of participating teachers and fidelity of implementation is established. To perform the analysis at the teacher level, a teacher-level aggregate student outcomes measure is required. We calculated teacher value added and used it as the outcome measure in this analysis.

Data Collection and Analysis:

The data used were 1) NCES data on demographics and other characteristics of the schools to establish baselines. 2) Tracking participation of teachers. Because the RCT had specific concerns with attrition, this was more carefully tracked in that context. In Scale-up, researchers tracked the numbers of schools, teachers, and students served by this initiative. Data included teacher attendance at the program PD institute (using the attendance logs), which schools and teachers agreed to complete study surveys, and if/when schools or teachers were no longer participating in the program (either because they left the school or were no longer implementing the program). 3) Surveys of teachers and principals. A large number of the same questions were asked annually in both contexts.

In Scale-up, data were linked across years to track the expansion and participation of states, districts, and schools. The "participant tracker" was updated with information as researchers received it. The method for uncovering teachers or schools that were no longer participating in the program was primarily survey follow-up or other direct communication with teachers or administrators. The tracker served the important function of tracking which district, schools, and teachers were participating, but it was not initially designed as a formal data collection tool for research purposes. It does, however, allow us to understand the processes of "attrition" or expansion beyond what is possible with only the survey data.

Answering all questions involved performing t-test for the differences between the two groups. In addition, the third question involved two intermediate steps. First, for the RCT we need to construct the teacher-level student outcome measure – teacher value-added. We used a conventional approach, in which value-added is calculated as teacher fixed effects, T, included in the student-level regression of posttest on pretest and student characteristics: $Y_t = \alpha Y_{t-1} + T + \beta X + \varepsilon$. A study-administered student assessment developed by ETS was used for both pretest and posttest.

Second, we performed linear regression analysis, with the teacher value added as the outcome and a variety of teacher survey items as covariates: $T = \beta Z + \varepsilon$. Exploratory nature of the analysis and the large number of covariates included a priori involved stepwise model selection based on the maximization of adjusted R-squared.

Findings / Results:
We did not find any differences in school characteristics significant at 0.05 level. Comparative descriptive statistics for RCT and Scale-up schools is presented in Table 1 in Appendix B.

Our analysis of the correlates of student outcomes (teacher value added) shows that only teachers' confidence in their ability to implement the program in the classroom has a significant (at 0.05 level) positive association with student outcomes and several others are marginally significant (see Table 2, Appendix B). RCT and Scale-up groups of schools do not differ on that characteristic (Table 3, Appendix B) suggesting that the program implemented in the wild has the same potential to affect student outcomes positively.

At the same time, we see that the set of variables associated with student outcomes does not overlap with those identified in our earlier study set of determinants of successful adoption: reported by teachers attendance of program meetings, levels of responsibility for and commitment to the success of the program at the school level. These two characteristics have higher values in Scale-up schools where teachers volunteered to participate in the study (Table 3, last two rows).

Conclusions:
Our study of comparable schools implementing the same program in different contexts highlights characteristics that are often not attended to in rigorous effectiveness research but are pertinent to understanding the effectiveness and scalability of the program. In the current study, a direct comparison of school achievement was not possible since the outcome measure used in the RCT was researcher-administered. However, we are able to look at teacher characteristics associated with higher student achievement in the RCT and we can look at teacher (and principal) characteristics associated with school-level gains in program participation in Scale-up. We found that RCT schools had lower levels of program-related characteristics associated with scale-up. Scale-up schools had similar levels of program-related teacher characteristics associated with greater achievement. But schools that were successful in growing participating internally were not more likely to have the characteristics associated with achievement gains. This suggests that program effectiveness in a controlled study may not be indicative of the program's prospects of wider adoption and, in fact, such studies may not create the best conditions for future program sustainability.

Appendices

Not included in page count.

Appendix A. References

Cole, M., Hood, L., & McDermott, R. P. (1978). *Ecological Niche-Picking: Ecological Invalidity as an Axiom of Experimental Cognitive Psychology*. Unpublished manuscript. Retrieved September 27, 2015 from http://www.academia.edu/11407068/Ecological_Niche_Picking_Ecological_Invalidity_as_an_Axiom_of_Cognitive_Psychology

Gibson, J.J. (1966). *The Senses Considered as Perceptual Systems*. Boston: Houghton Mifflin.

Zacamy, J., Newman, D., Lazarev, V., & Lin, L. (2015). *School Processes That Can Drive Scaling-up of an Innovation, or Contribute to its Abandonment*. Paper to be presented at the national conference of The National Center on Scaling Up Effective Schools in October 2015.

Newman, D., Griffin, P., & Cole, M. (1989). *The construction zone: Working for cognitive change in school.* New York: Cambridge University Press.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton Mifflin.

Appendix B. Tables and Figures

Table 1. School characteristics in RCT and Scale-up schools.

Variable	RCT sites (treatment group)	Scale-up sites
Number of districts	11	31
Number of schools	12	35
Average school size	988	1196
% FRPL	34.54%	31.79%
%ELL	2.21%	1.27%
%Minority	18.03%	17.73%
Number of teachers	64	231
Average years teaching experience	10.09	10.90

Note: None of the differences (t-test) are significant at 0.05 level.

Table 2. Association between teacher value-added and program characteristics

	Estimate	Std. Error	p value
Constant	-1.606	1.054	0.136
Confidence level	0.223	0.091	0.019
Preparation (days total)	0.194	0.123	0.124
Fidelity of Implementation (score)	-0.618	0.422	0.151
Effectiveness	-0.126	0.094	0.185
Teaching Experience (years)	0.021	0.011	0.067

Table 3. Differences between RCT and Scale-up schools on potential determinants of program effectiveness and scale up.

	SU	RCT	p value
Confidence level	3.65	3.64	0.97
Preparation (days total)	8.93	9.27	0.161
Fidelity of Implementation (score)	0.68	0.82	0.012
Effectiveness	3.97	3.57	0.0001
Teaching Experience (years)	10.9	10.1	0.415
Level of responsibility for program success	3.45	3.28	0.03
Level of commitment to program success	4.14	2.43	<0.001