Tensions and Tradeoffs: Responding to Diverse Demands for Evidence - Case Study of an Underpowered, Randomized Control Trial

SREE 2019

Daniel F. McCaffrey

John Sabatini

Jill Burstein

Kelsey Dreier

Kietha Biggers

Education Testing Service

Princeton, NJ

Introduction

IES Goal 2 Development grants include the requirement to evaluate the promise of the intervention, with one option being to carry out an underpowered randomized control trial (RCT). We will describe such an evaluation conducted in a study of the Language Muse Activity Palette^{TM} (hereafter the *Palette*). We willf elaborate on the tensions and trade-offs that arise from design decisions, and how best to interpret outcomes in terms of promise and next steps for the design team (e.g., redesign of implementation, adapting to specific audiences, or planning for large-scale study).

Background

English language learners (ELLs) continue to be the fastest growing subpopulations in US schools, but they lag in literacy and subject-area achievement. Helping ELLs maintain adequate growth towards content and reading goals, is a struggle for both content-area teachers and EL specialists. The *Palette* was designed to aid both content area teachers and EL specialists to adapt curriculum for ELLs. To use *the Palette*, a teacher simply uploads a text into the web interface. Using NLP, linguistic features in that text are automatically analyzed across *vocabulary, sentence,* and *discourse* categories, and language activities are generated and recommended. Seventeen different activity types have been developed. Teacher scripts provide tips for using activities to foster student learning.

Setting and Participants

School Sites. The study was conducted in six middle and six high schools (grades 6-12). Four schools were in an urban area, one rural and seven suburban. The study took place in two cycles – six schools in 2016-17 (Year 1), then six in 2017-2018 (Year 2). We recruited 28 teachers from participating school sites. Three high school teachers dropped out in Year 2.

Research Design

The goal of the study was to test the impact of the *Palette* on student reading skills and develop other evidence of potential impact of the intervention. We focus on the test of impact.

Intervention

Treatment Group. Treatment teachers received a two day professional development workshop to learn *Palette* functionality, detailed planning for *Palette* implementation, and refreshing linguistic awareness. Teachers then had the opportunity to use the *Palette* in their classrooms for between 8 to 12 weeks.

Control Group. Control teachers conducted business-as-usual during the evaluation period.

Random Assignment of Teachers

We stratified the fourteen teachers in Year 1 into 5 blocks based on their school district. In Year 2, with fewer districts, we stratified by school type (middle or high school).

Instruments

*Training Evaluation, Professional Experiences, Teacher Reflection Logs, and Observation Protocol*s. Space precludes describing here, but relevant details will be shared in presentation, as appropriate.

Pre/Post Test: The RISE. The Reading Inventory and Scholastic Evaluation (RISE) is a Web-based, reading components assessment battery that measures language and reading skills including decoding/word recognition, vocabulary, morphology, sentence processing, efficiency of basic text reading, and reading comprehension (Sabatini et al., 2015). The RISE utilizes IRT-based, developmental scales by subtest. Multiple parallel forms are available, with grade level reliabilities for each subtest ranging from .83-.92.

Data Collection and Analysis

Participating teachers tested their students using the RISE prior to the intervention and at the end with an alternative form. Scores on the six subtests of the RISE pretest served as covariates for testing the intervention. Posttest scores were combined into two composites: WORD that heavily weighted lexical skills and ranged from 224 to 275, and COMP, which weighted comprehension and ranged from 224 to 267. We tested the effect of the *Palette* on the two posttest composites.

To estimate impact, we fit a linear regression to each of the composite posttest scores. The regression included an indicators for the student's teacher's experimental status (*Palette* or control) and indicators for the random-assignment strata. Per our original analysis plan, we also included pretest scores for each of the six RISE domains as covariates. The coefficient on the treatment indicator serves as the estimate of the treatment effects. We used a permutation test to the null hypothesis of zero treatment effect. We also ran joint tests of the null hypothesis of no effects on either test again using a permutation test.

Exploration of the data showed that as a result of heterogeneity in the sample, teacher dropout, student dropout, and bad luck with randomization, the control group had more middle school teachers and students whereas the treatment group had more high school teachers and students. Consequently, we ran additional tests of the intervention that controlled for student grade-level and interactions between grade-level and pretest scores.

Results

Following our planned protocol of controlling for the pretest and randomization strata, the estimated impact of the *Palette* on WORD was 1.03 scale score points (effect size, ES=0.12, p=0.33). The estimated impact on COMP was 3.14 (ES=3.8, p=0.006). The joint test of no impact on either scale had a p-value of 0.02.

We knew that the assignment of students and teachers was unbalanced and our results could be a type I error. We repeated the analyses adding indicator variables for the students' grade-level to the model and still found a significant impact on COMP (ES=0.27, p=0.03). We then added grade-level and pretest scores by grade-level interactions to the model. The estimated impact for the COMP composite now was 1.65 (p=0.26). The last model included many variables relative to the sample size of 305 students which can degrade the precision of the estimated impact and our confidence in the results.

Conclusions

The limited scope of the evaluation study meant that even small break downs, such as attrition of teachers, loss of outcome data, or uneven distribution of schools/teachers after randomization can create biases and threats to planned analyses. Multiple analysis strategies were used to appropriately address these potential biases and threats, but the result of implementing these statistical safeguards meant that the outcomes wavered from significant to non-significant. In the session, we will discuss how best to interpret 'promise' in making decisions about what is the best course of action for the research team for next steps.