REL West Toolkit Efficacy Evaluation

SUPPORTING STATEMENT
FOR PAPERWORK REDUCTION ACT SUBMISSION

PART B: Collection of Information Employing Statistical Methods

April 2023

Submitted to:
Institute of Education Sciences
U.S. Department of Education

Submitted by:
RAND Corporation
1200 South Hayes Street
Arlington, VA 22202
(703) 413-1100

Tracking and OMB Number: (XX) XXXX-XXXX
Revised: 03/22/2023

Overview

The U.S. Department of Education (ED), through its Institute of Education Sciences (IES), requests clearance for the recruitment materials and data collection protocols under the OMB clearance agreement (OMB Number (XX) XXXX-XXXX) for activities related to the Regional Educational Laboratory West Program (RELWest).

Elementary-grade students in U.S. public schools continue to struggle with reading comprehension, with only 35 percent of 4th-grade students performing at or above proficient on NAEP scores in reading (Hussar et al., 2020). To address this problem in earlier grades, when schools begin reading comprehension instruction, the REL West toolkit development team is developing a toolkit to support teachers in implementing evidence-based instructional strategies to improve reading comprehension among students in grades K–3. The toolkit is based on the *Improving Reading Comprehension in Kindergarten Through 3rd Grade* IES practice guide (Shanahan et al., 2010) and is being developed in collaboration with state and district partners in Arizona.

The REL West toolkit evaluation team is requesting clearance to conduct an independent evaluation that will assess the efficacy and cost-effectiveness of the school-based professional development resources included in the toolkit. The evaluation will also assess how teachers and facilitators implement the toolkit to provide context for the efficacy findings and guidance to improve the toolkit and its future use. The evaluation will take place in 70 schools across six districts in Arizona and focus on K–3 reading comprehension for all students.

B1. Respondent Universe and Sample Design

The evaluation will employ a school-level, cluster-randomized controlled design, and take place in six districts in the state of Arizona. The evaluation will examine the impact of the toolkit on student reading comprehension for all students in grades K–3 in study schools. In addition, evaluating the toolkit's impact on Hispanic/Latinx students—50 percent of the student population in the state's 25 largest districts—provides an opportunity to assess how this toolkit may improve early literacy in this historically underserved population.

For recruitment purposes, the toolkit evaluation team will prioritize school districts that have at least 10 elementary schools, with at least 50 percent of Hispanic/Latinx students in the school. Using 2018 Common Core of Data (CCD) and publicly available Arizona Department of Education (ADE) data, the team has identified 25 districts that meet these criteria; the team anticipates needing 6 districts to participate in the study. The team will restrict the universe of schools to public, non-charter schools with at least two full-time equivalent (FTE) teachers and at least two classes in each grade, grades K–3. The school must also have a staff member who can serve as the toolkit facilitator in the school; we anticipate this will be a site-based coach or teacher leader, but the exact title may differ by district and school. Within schools, the sample will include all K–3 regular classroom teachers ($N \approx 720$) and their students ($N \approx 9{,}000$);

principals and facilitators in 70 schools; and 12 district administrators and professional development coordinators.

Table B.1 shows the sample sizes and expected response rates for each level of data collection.

Level of Sample	Sample Size	Response rate
District	12	100%
School/School Leader	70	100%
Teacher	720	85%
Student	9,000	85%

B2. Information Collection Procedures

a. Notification of the Sample and Recruitment

The Arizona Department of Education (ADE) will help connect the evaluation team to districts, and the team will leverage REL West's established relationships throughout the state to set up the first meetings. ADE will make initial contact with district leaders, through phone calls or emails, to ask them to look at the study communications. Researchers will follow up within a day with informational materials and schedule a time to meet. In the district call, the toolkit and the study will be described to district leaders and their questions will be answered. District leaders will also be asked to describe district-specific benefits and challenges of participating in the study. If district leaders are interested in participating, the evaluation team will ask for their help contacting schools and their ideas for how the study might be a fit for their schools.

Upon district agreement, the team will reach out to school principals. District leaders will be asked by the evaluation team to hold information meetings for principals. The researchers will then email each school an information package and schedule a school-specific conversation with the principal. If the school principal is interested, staff Q&A meetings by school will be held, and informational webinars across schools, to provide information directly to teachers and facilitators and to hear their thoughts.

Researchers on the team will ask school principals, facilitators, and teachers to review and sign a brief consent statement prior to random assignment indicating that they understand the intervention and the study and will participate to the best of their ability, regardless of the condition to which they are assigned. This is a non-binding agreement. Schools in the study will be included if the principal, the facilitator, and at least one-half of the teachers in each grade make this commitment, and if the district does not require active consent from students/parents for participation in the study. Recruitment materials that will be used for this study are included in Appendix A.

Primary data collection recruitment will occur in schools that consent to be part of the study. Each round of data collection will include an initial outreach and three follow-up emails for participants who have not completed the consent form. Data collection communication email texts are included in Appendix B.

Active consent for the this research is not required by the district. Also, the IRB has determined that passive consent is acceptable and the research is exempt, because it involves established educational settings and involves normal educational practices (i.e., professional development) that are not likely to adversely impacts students' opportunities to learn required educational content. Also, requiring active consent increases burden on parents and could jeopardize sample size (and by extension, the study's ability to detect significant effects if the intervention has a truly positive impact on student outcomes).

b. Statistical Methodology for Stratification and Sample Selection

The evaluation team will recruit districts in Arizona that meet the criteria of having an average of at least 50% Hispanic/Latinx students in each school until 6 districts have agreed to participate in the study. Within each district, the recruitment will target all elementary schools that have at least two full-time equivalent (FTE) teachers and at least two classes in each grade, grades K–3. Schools will be recruited until 70 schools have agreed to participate. All teachers who teach reading instruction to students in grades K-3 will be recruited for the study, with the expected sample size of 720 teachers.

Because the toolkit, centering on teacher knowledge and use of the five practice guide recommendations for improving students' reading comprehension, is designed to be used as part of a school's approach to improve literacy instruction by all K–3 teachers within a school, the evaluation team proposes using the school as the unit of assignment. This level of assignment has multiple methodological benefits, including removing the within-school threat of diffusion of the toolkit use and crossovers. Random assignment will take place in April 2024—immediately after recruitment closes and administrative data baseline student achievement scores and characteristics has been collected.

The study will use a rerandomization process (Morgan & Rubin, 2012) to assign schools to treatment or control group status. Rerandomization is an iterative process that seeks to find the optimized allocation to treatment and control conditions that balances baseline characteristics, based on a pre-specified criterion that defines the level of imbalance. It relies on the availability of covariate information at the design stage of the experiment. The benefit of rerandomization is that it improves covariate balance and leads to more precise estimates of the treatment effect (Morgan & Rubin, 2012). For this study, the toolkit evaluation team will identify covariates to use for rerandomization and will set an empirical threshold for the group differences for acceptable randomization prior to assigning units to conditions. The rerandomization procedure will then continue to rerandomize until the empirical threshold for acceptable differences is met. Potential covariates to be included in this model include school average reading comprehension scores by grade level (measured at baseline before random assignment), school average demographic composition (e.g., percent of Hispanic/Latinx and Black students), school average special education students, school average English learner students, and school average teacher first-year retention rates. Before proceeding with rerandomization, the study team will conduct simulations to examine the extent to which a simple random assignment is likely to result in imbalanced covariates using the baseline data gathered from school districts. If based on 1,000 simulations of simple randomization we find that at least 1 of the covariates has a difference between treatment and comparison groups of 0.20 standard deviations, then we will proceed with rerandomization.

The timing of random assignment will take place to allow for facilitator training starting in June 2024, and teacher training in August 2024, and is therefore embedded in regular summer training activities. Plans for random assignment will be communicated with district and school officials early in the recruitment process to ensure buy-in and randomization assignments will be carefully documented. To maintain the integrity of the random assignment, all analysis of data will account for these procedures, as described below.

c. Estimation Procedures

Because outcomes are measured at the student or teacher level and each is nested within schools, the evaluation team will use hierarchical linear models to compare student and teacher outcomes at schools randomly assigned to the treatment group to student and teacher outcomes at schools randomly assigned to the control group (business as usual). Error terms within a group will not be statistically independent because of nesting in schools. Hierarchical linear modeling accounts for the statistical dependence of the error terms and, with other things equal, produces unbiased estimates of the impacts and standard errors (Raudenbush & Bryk, 2002; Shadish et al., 2002). The evaluation team will estimate the Intent-To-Treat (ITT) effect for RQ1 and RQ2 in a two-level model, with students at Level 1 and schools at Level 2. Because the school is the level of assignment, the treatment indicator is at Level 2. The team will estimate the effect of the toolkit on teachers who used it for RQ3 with a Complier Average Causal Effect (CACE) analysis. A two-level model to estimate the effect of the toolkit on teacher knowledge and teacher practices (RQ4) will be used, with teachers at Level 1 and schools at Level 2, and again with the treatment indicator at Level 2. For binary teacher outcomes, a hierarchical generalized linear model with a logistic link function will be used.

Continuous Student Outcomes**.** Student assessment outcomes will be modeled with a two-level hierarchical model. Because schools are randomly assigned to the treatment group, the treatment effect is included in Level 2 (i.e., the school level) of the model. The team will calculate the ITT effect of the toolkit, relative to business as usual, on student outcomes. The two-level model for continuous student outcomes is given by:

Level 1 (students): $$Y_{ij}=\pi_{0j}+\pi_{01j}\,PriorAch_{ij}+\sum_{p=2}^{P}\pi_{pj}\left(a_{pij}-\bar{a}_{p.j}\right)+\varepsilon_{ij} \tag{1a}$$

Level 2 (schools): $$\pi_{0j}=\beta_{00}+\beta_{01}Treat_{j}+\sum_{p=2}^{P}\beta_{0p}X_{pj}+\sum_{q=P+1}^{Q}\beta_{0q}D_{qj}+\omega_{0j} \tag{1b}$$

$$\pi_{pj}=\beta_{p0}\text{ , }p = 1, . . ., P \tag{1c}$$

where Y_{ij} is the student assessment for student *i* in school *j* at follow-up period, and $PriorAch_{ij}$is the student's baseline score on an assessment of the same subject for student *i* in school *j* in the previous time period. The term a_{pij} are student variables that influence the outcome of interest measured at baseline, such as FRPL eligibility, race, ethnicity, gender, English learner status, IEP status, age, and an indicator for student grade level, and $\bar{a}_{p.j}$ is the school-level average of the student variable. The term $Treat_j$ is an indicator variable taking a value of 1 for treatment-group schools and 0 for control-group schools. The term X_{pj} are school-level covariates (*p = 2, . . ., P*), including outcome measures collected at baseline and school characteristics. School characteristics will include school-level average reading achievement scores from the prior year (based on benchmark or standardized state assessments), the demographic composition of the school, school enrollment, and any other relevant measures available from data sources measured at baseline (e.g., principal experience or school professional development opportunities). D_{pj} are district fixed effect indicators, which will be entered as K -1 dummy variables, where K is the number of districts. The error term in Equation 1a, ε_{ij}, is assumed to be distributed $N(0,\sigma^2)$, and the error term in Equation 1b, ω_{0j}, is assumed to be distributed $N(0,\tau^2)$. Covariates in the Level 1 model are centered at the school mean, so student-level parameters are estimated using within-school variation.

Students in grades K–2 in 2023/24 will take a benchmark assessment, administered by the reading teacher per state requirements, and standardized scale scores from this assessment will be considered as the baseline.

- For students who are in grades 1 and 2, the baseline assessment will be the January 2024 assessment, with the September 2024 or May 2023 assessment as a backup if the January 2024 assessment is missing.
- For students who start Kindergarten in fall 2024 or students who join the sample in grades 1, 2, and 3 in the fall of 2024, their September 2024 benchmark assessment will be considered as the baseline measure, using the scale score, if available, or the proficiency rate, if necessary.

The evaluation team will standardize assessment scores within grade level and type of assessment. The parameters of interest in this model, as well as the ones described below, will be estimated using standard methods and statistical software. The previously described models will be estimated jointly for all grade levels for all students, and then again jointly for all grade levels for only Hispanic/Latinx students. In exploratory analyses, researchers will estimate the models separately by grade level for all students and for Hispanic/Latinx students.

Complier Average Causal Effect Estimate (CACE). Because use of the toolkit is optional, a difference between random assignment to the toolkit and the actual use of the toolkit (uptake) may be observed. To examine the effect of students' receipt of instruction using the toolkit on reading achievement for all students (RQ3), the team will estimate a two-stage least squares (2SLS) model:

$$Z_{ips} = \phi\, Treat_s + \tilde{\beta}\left(X \text{¿¿} ips - \overline{X}_s\right) + \mu_{ips} \text{¿} \qquad (2)$$

$$Y_{ips} = \delta\, \widehat{Z_{ips}} + \beta\left(X \text{¿¿} ips - \overline{X}_s\right) + \epsilon_{ips} \text{¿} \qquad (3)$$

In this model, Z_{ips} is a binary indicator of whether student *i* with teacher *p* in school *s* received instruction using the toolkit; X_{ips} represents student characteristics measured at baseline, including prior year test scores on the same subject test, FRPL eligibility, race, ethnicity, gender, English learner status, IEP status, age, and an indicator for student grade level; $Treat_s$ is an indicator for the school being assigned to the treatment group, and $\widehat{Z_{ips}}$ is the receipt of instruction using the toolkit, calculated from equation (2). The coefficient of interest is δ, which is the CACE estimate. The indicator for take-up (Z_{ips}) will be defined as 1 if the student attended school with their teacher for at least 85 percent of the school year combined with if the teacher met a pre-set threshold on a measure of implementation, described in more detail below in the implementation section. Specifically, as indicated in Table 5 below, there are teacher-level measures of implementation, with a 0–1 score for each measure. For the CACE analysis, the evaluation team will average the scores from the 5 measures and create the indicator Z_{ps} as equal to 1 if the average of the five scores is at least 0.8 and the student attended the school with this teacher for at least 85 percent of the school year. Our estimation method is aligned with What Works Clearinghouse (WWC) standards on individual RCT designs, and follows the strategy proposed in Schochet and Chiang (2009), which raises concerns about CACE estimation in the cluster RCT setting. In particular, the covariates in equation 3 will be the same as the covariates used in equation 2. Researchers on the team will report the F-statistic from the first stage regression to demonstrate that the instrument had sufficient strength. Attrition levels for the treatment and comparison groups will be reported, and baseline equivalence of the analysis sample, if necessary due to high attrition. We will also explore the level of non-compliance and consider using maximum likelihood estimators developed for this context (Jo et al., 2008).

***Continuous Teacher Outcomes*.** ITT impacts on continuous teacher outcomes will also be modeled using the two-level hierarchical model in Equation 1, with teachers nested within schools and the treatment variable modeled at the school level. This model is similar to the student model, except that teacher characteristics will be included at Level 1. These characteristics include teacher experience, teacher demographics, and teacher access to professional development. Outcomes for these models include the percentage of correct responses on the teacher pedagogical content knowledge assessment from the teacher survey and an index of teacher practices from the teacher logs (RQ4). These outcome measures will be constructed according to directions by the measure developers.

***Binary Teacher Outcomes*.** Some teacher-level outcomes may be coded as indicators (e.g., an indicator for whether the teacher uses a particular practice for RQ4). For teacher dichotomous outcomes, the evaluation team will use a logit link function to transform the dependent variable into the odds of achieving an outcome, such that Y_{ijt} is replaced with $\log\left(\frac{\varphi_{ij}}{1-\varphi_{ij}}\right)$ in Equation 1a, where φ_{ij} is the probability of success. The covariates and the parameters of the binary model are defined similarly to the continuous teacher outcome model described in the previous section. The β_{01} coefficient measures the impact of the toolkit on the log-odds of success for the teacher outcome.

***Mediator Analysis*.** To conduct a mediator analysis (RQ5), the team will follow Baron and Kenny (1986) and examine whether there is a direct relationship between the intervention (the toolkit) and the mediator (e.g., teacher knowledge). If there is a statistically significant relationship, the mediator will be included as one of the covariates in the main impact regression, and the team will examine the extent to which this changes the relationship between the effect of the toolkit on student reading test scores for each grade level, separately for the full sample and for Hispanic/Latinx students. For example, this post random assignment variable in the model will be included, and if the estimated treatment coefficient decreases in absolute magnitude, then that provides evidence that this variable is among the pathways through which the impact occurs, and therefore the variable is a mediator.

Sensitivity Analyses. Sensitivity analyses will be conducted to test the extent to which the evaluation team's estimates are driven by model assumptions. The results of these analyses will be summarized in a report, and more detailed results will be displayed in appendices. In particular, the research team will examine whether the inclusion of school indicators (fixed effects), as opposed to school characteristics, changes the coefficient estimates for the treatment effect. The team will also estimate models where district fixed effects are replaced with continuous district characteristics.

Strategies for Correcting for Multiple Hypothesis Testing. The proposed analysis does not require correction for multiple hypothesis testing, as it includes only one confirmatory analysis comparison within the reading comprehension domain (treatment vs. control outcome for all students).

Implementation Fidelity. To assess fidelity at the program level (6a), the evaluation team will construct quantitative indicators of the extent to which the intended activities were carried out and individual teachers and facilitators participated in the intervention. These analyses will culminate in a program-level rating of whether the program was implemented with fidelity overall, by facilitators, and by teachers. This evaluation will examine implementation of the key school-based components of the toolkit, which are hypothesized as the primary mechanism for improving teacher knowledge, instruction, and student reading comprehension outcomes. The key components of the toolkit intervention—integrating both activities and materials—are planning, learning, and institutionalizing. Because these are three distinct phases of implementation, the evaluation team will look at implementation for each of these three components, constructing separate measures for planning, learning, and institutionalizing.

Implementation Treatment Contrast. To discern treatment contrasts between intervention and control group teachers (6b), the research team will conduct descriptive analysis of use of toolkit-like activities. The evaluation team will conduct descriptive analysis of teacher survey items on professional learning experiences. Items are asked of toolkit teachers about the toolkit and, separately, about non-toolkit activities. Parallel items are asked of control group teachers about any relevant professional learning activities. The structure facilitates analysis of treatment contrast for RQ6b by allowing the analysts to combine toolkit and non-toolkit activities for treatment teachers in contrast to all relevant activities for control teachers.

Implementation Challenges. To identify challenges for completing toolkit activities (6c), the evaluation team will collect and analyze data from teachers, school leaders and facilitators, and district leaders. The evaluation team will report the average percent of teachers in the sample who disagree strongly or somewhat that leaders encourage participation in the toolkit professional development. The threshold for identifying a factor (school leader support in this example) as a challenge is that more than 50 percent of treatment teachers in the sample disagree with this statement. Similarly, the school leader and facilitator survey asks questions about potential challenges for toolkit implementation, and these will be summarizes. The district interview protocol has some closed-ended questions about potential implementation challenges, which will be analyzed as the questions above. In addition, the district interview has five open-ended questions about alignment with district policies, supports, challenges, and facilitators. The evaluation team will analyze these responses qualitatively, coding each interview in Dedoose and developing analytic summaries for the most commonly occurring challenges and supports.

d. Degree of Accuracy Needed

The evaluation team used NCES's Elementary/Secondary Information System to identify Arizona elementary schools and district sizes. The team used the PowerUp! tool to calculate the number of schools required for the study to have an 80 percent chance of detecting, as statistically significant, an effect size (ES) of 0.15 standard deviations in student achievement scores (Dong & Maynard, 2013). This effect size is conservative compared to those found in prior studies of lower elementary school reading interventions (Gersten et al., 2010: ES = 0.77; Knezek & Christensen, 2007: ES = 0.30; Vernon-Feagans et al., 2012: ES = 0.25), and is supported by a meta-analysis (Didion et al., 2020) of studies of the impact of professional development on student achievement (i.e., for 21 RCT studies reviewed, Hedges' g = 0.18, p < .05, and a 95% confidence interval of [0.07, 0.29]). The analysis for RQ2 assumes a two-level model (with students nested in schools). The research team assumes a school-level ICC of 0.15, and the team also assumes the proportion of variance in outcome between schools and students (R2) to be 0.7 for reading achievement (Hedges and Hedberg,2013). Using data on Arizona elementary schools from the 2018 CCD, the evaluation team assumes a harmonic mean of 200 eligible students (grades K–3) and 100 eligible Hispanic/Latinx students per school. Using schools as the unit of assignment, assumptions about the number of Hispanic/Latinx students per school, and a target effect size that has substantive importance and is reasonable to expect, the team anticipates needing a sample of 70 schools across six districts.

e. Unusual Problems Requiring Specialized Sampling Procedures

There are no unusual problems requiring specialized sampling procedures.

f. Use of Periodic (less than annual) Data Collection to Reduce Burden

This project will collect data one time for recruitment. Outcome data will need to be collected more frequently than annually because the evaluation is occurring within one school year, and some measures will need to be assessed in September and June of the same year. A longer period between data collection would make it difficult for the study team to meet the requirements for the efficacy study (by preventing baseline and follow up data collection in the timeframe necessary for the evaluation).

B3. Methods for Maximizing the Response Rate

The evaluation team is committed to obtaining complete data for this evaluation. A large share of the impact analysis question for the evaluation relies heavily on administrative data. ED's contractor anticipates a 100-percent response rate from Arizona districts on teacher and principal measures in the administrative data, and an 85-percent response rate from Arizona districts on students measures in the administrative data. A key to achieving complete administrative data is tracking the data components from each district with e-mail and telephone contact to the appropriate parties to resolve issues of missing or delayed data files. All administrative data files will be reviewed for consistency and completeness. If a data file has too many missing values or if an instrument in the implementation study has too few items completed to be counted as a response, the evaluation team will seek to obtain more complete responses by e-mail or phone.

Based on its prior experience with administering surveys to principals and teachers in a variety of schools, districts, and states, the evaluation team expects the response rate for the baseline survey to be at 85 percent for those principals/facilitators and teachers who consent to participate in the study. We will contact non-responding school staff members up to four times to encourage participation. Three follow-up email reminders will be sent to individual respondents in the event that responses are not obtained for Web-based surveys. The evaluation team will consider other modes of follow-up including reminder letters and reminder phone calls if response rates are below expectation.

In addition, a number of steps will be taken to maximize response rates. For example, sampled respondents will receive advance communications explaining the study, introducing the REL West, provide an assurance of confidentiality, and encourage them to participate as a way to help refine the toolkit. Respondents also will be given a contact number to reach the evaluation team with questions.

Finally, respondents will receive an incentive for participating in the study: $30 per teacher survey, $50 per school leader or facilitator survey, and $75 per two-week round of logs, to be provided at the end of each round of data collection. A maximum of $210 in gift cards will be provided to each participant as compensation. In addition, to ensure responses from control school, schools randomly assigned to this group will receive $2,500 to be used on activities unrelated to the intervention.

The evaluation team has multiple strategies to deal with missing data due to non-response. Prior to starting the analyses, the evaluation team will examine the extent of missing data overall and by treatment group. Starting from the WWC options for dealing with missing data, researchers on the team will use appropriate analytic methods to account for missing data and will consider options such as complete case analyses with regression adjustment, maximum likelihood methods, or non-response weights. Implementation of the approach will follow requirements such as using one of the WWC acceptable approaches and assessing the analysis sample for low attrition before applying the acceptable missing data approach. The most recent statistical literature will also be considered to examine other additional methods. If such methods are necessary, results using data not adjusted for missingness will also be included in an appendix for the report.

B4. Test of Procedures

The student outcome measures used to analyze the impact of the toolkit will be one of four benchmark reading assessments that were selected by ADE, all of which have demonstrated validity and reliability, and will not require pretesting. Teacher survey instruments and teacher practice logs will use valid and reliable measures, where feasible, but will include new items. These instruments will be pilot tested in Fall 2023, during the usability testing of the toolkit. We expect only minor changes based on the pilot test, given that many of the scales in the instruments we will pilot have been validated externally. District interview protocols will also be pilot tested, but with less than 9 respondents. The instruments are included in Appendix C.

B5. Methods to Minimizing Burden on Small Entities

The individuals consulted on the statistical aspects of the design include:

Herb Turner, President and Principal Scientist, Analytica; (215) 808-8808; herb@analytica-inc.com

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 5,* 1173–1182.

Didion, L., Toste, J. R., & Filderman, M. J. (2020). Teacher professional development and student reading achievement: A meta-analytic review of the effects. *Journal of Research on Educational Effectiveness, 13*(1), 29–66. doi:10.1080/19345747.2019.1670884

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness, 6*(1), 24–67. doi:10.1080/19345747.2012.673143

Gersten, R., Dimino, J., Jayanthi, M., Kim, J. S., & Santoro, L. E. (2010). Teacher study group: Impact of the professional development model on reading instruction and student outcomes in first grade classrooms. *American Educational Research Journal, 47*(3), 694–739.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445-489.

Hussar, B., Zhang, J., Hein, S., Wang, K., Roberts, A., Cui, J., & Smith, M. (2020). *The conditions of education 2020*. Washington, DC: Retrieved from https://nces.ed.gov/pubs2020/2020144.pdf

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods, 13*(1), 1–18. doi: 10.1037/1082-989X.13.1.1

Knezek, G., & Christensen, R. (2007). Effect of technology-based programs on first- and second-grade reading achievement. *Computers in the Schools, 24*(3-4), 23–41. doi:10.1300/J025v24n03_03

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, *40*(2), 1263–1282.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data\ analysis methods* (Vol. 1). Sage Publications.

Schochet, P. Z., & Chiang, H. (2009). *Technical methods report: Estimation and identification of the complier average causal effect parameter in education RCTs* (NCEE 2009-4040). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://ies.ed.gov/ncee/pdf/20094040.pdf

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

Shanahan, T., Callison, K., Carriere, C., Duke, N. K., Pearson, P. D., Schatschneider, C., & Torgesen, J. (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide* (NCEE 2010-4038). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. Retrieved from whatworks.ed.gov/publications/practiceguides

Vernon-Feagans, L., Kainz, K., Amendum, S., Ginsberg, M., Wood, T., & Bock, A. (2012). Targeted reading intervention: A coaching model to help classroom teachers with struggling readers. *Learning Disability Quarterly, 35*(2), 102–114. doi:10.1177/0731948711434048

Appendix A – Recruitment Materials

- Appendix A1 – Preliminary Letters to District Leaders, School Leaders, and Teachers
- Appendix A2 – District Recruitment Emails and Phone Call Talking Points
- Appendix A3 – School Leader Recruitment Emails, Phone Call Talking Points, and Agendas for Information Session and Webinar
- Appendix A4 – Teacher Recruitment Emails
- Appendix A5 – Frequently Asked Questions

Appendix B – Data Collection Communication Materials

- Appendix B1 – District Leader Interview Email & Follow-Ups
- Appendix B2 – School Leader Data Collection Email & Follow-Ups
- Appendix B3 – Teacher Data Collection Email & Follow-Ups

Appendix C – Primary Data Collection Instruments

- Appendix C1 – Teacher Surveys
- Appendix C2 – Teacher Log
- Appendix C3 – School Leader and Facilitator Survey
- Appendix C4 – District Interview Protocol