Long-Term Impacts of Reading Recovery Through 3rd and 4th Grade: A Regression Discontinuity Study

Henry May, Aly Blakeney, Pragya Shrestha, Mia Mazal, and Nicole Kennedy Center for Research in Education and Social Policy (CRESP), University of Delaware

May 23, 2023

Author Note

Henry May https://orcid.org/0000-0002-5965-3933

Aly Blakeney is now an Instructor of Economics in the Department of Humanities & Social Sciences at the Hotchkiss School.

This work has been supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A170171. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. The SAS code used to conduct statistical analyses and produce data visualizations for this study are openly available in the Open Science Framework (OSF) at http://doi.org/10.17605/osf.io/jc9bm.

We have no conflicts of interest to disclose.

The authors wish to thank Prof. Sean Reardon and three anonymous reviewers for their detailed and helpful feedback on this manuscript.

Correspondence concerning this article should be addressed to Henry May, 125 Academy Street, Rm. 107 Pearson Hall, University of Delaware, Newark, DE 19716

Abstract

To estimate the long-term effects of the Reading Recovery® intervention, a regression discontinuity design (RD) was implemented in a randomly selected sample of Reading Recovery schools during each year of the federally-funded i3 Scale-Up external evaluation (2011-2015) and also in one additional cohort during the 2016-17 school year. Long-term outcomes were measured by collecting scale scores on state achievement tests in reading or English language arts (ELA) that were administered to students during 3rd and 4th grades (2013-2019). We were able to record and link 3rd grade state test score data in reading/ELA for 9,879 students, and 4th grade state test data for 6,345 students. Results suggest that the long-term impact of Reading Recovery on students' reading/ELA test scores in 3rd and 4th grades is statistically significant and substantially negative; students who participated in Reading Recovery in first grade had 3rd and 4th grade state test scores in reading/ELA that were, on average, .19 to .43 standard deviations (about one-half to one full grade level) below the state test scores of similar students who did not participate in Reading Recovery.

Keywords: early reading, intervention, instruction, regression discontinuity, impact evaluation, literacy education, elementary school students, reading difficulties

Long-Term Impacts of Reading Recovery Through 3rd and 4th Grade: A Regression Discontinuity Study

Early literacy instruction continues to be a topic of critical and contentious debate, and the adoption of evidence-based strategies and interventions in mainstream practice has been anything but straightforward (Goldenberg & Goldenberg, 2022). Teachers in early elementary grades are expected to engage students in various activities that build their ability to decode, or sound-out, words (code-focused instruction) and other activities that build their ability to understand and draw inferences from what they've read (meaning-focused instruction) while matching the balance of these two sets of strategies to the current skills of each student (Connor, Morrison, & Katch, 2004). Recommendations from the What Works Clearinghouse (WWC) Practice Guide on early literacy (Forman et al, 2016) focus on four key elements: (1) academic language to support drawing inferences, understanding sequences and relationships, and building vocabulary; (2) phonemic awareness to understand letter-sound relationships; (3) decoding to learn how to sound out words; and (4) reading connected text to support accuracy, fluency, and comprehension. The activities teachers use to support these skills vary tremendously, and often involve grouping students based on current skills. Students in early grades who are just learning to read are better served by more teacher-led activities (Connor, Morrison, & Katch, 2004) such as reading a book together (with the teacher guiding students in decoding and comprehension as the book is read), or activities in which the teacher engages the students in understanding how letters and specific combinations of letters produce sounds to form words (i.e., phonics instruction). As students build their decoding skills, they engage more frequently in student-led activities. For example, small groups of children discussing answers to a reading comprehension worksheet or more independent activities like sustained silent reading. Conversely, students who

struggle when learning to read often participate in supplemental reading interventions like Reading Recovery. This study aims to contribute to the evidence base regarding the effectiveness of Reading Recovery.

Notably, this study coincides with a recent and very public debate on popular practices in early literacy instruction that has drawn attention to the "three-cueing" strategies used in several widely-implemented early literacy curricula, including Reading Recovery (Schwartz, 2020). Ken Goodman first proposed three-cueing in his 1967 article entitled, "Reading: A psycholinguistic guessing game," arguing that word identification is informed by multiple cues that assist a reader in "guessing" a word. Within modern curriculum the three cues are meaning, structure, and visual (MSV). These correspond to questions like: Does the guessed word make sense in the sentence? Is it grammatically correct? Does it match what you see (i.e., the letters and pictures)? But many literacy experts say that research has proven this theory of reading to be false, that children using this approach produce too many incorrect guesses that undermine effective reading, and that teaching this approach may be harmful because it reinforces ineffective strategies used by struggling readers (Hanford, 2019; Seidenberg, 2017; Kilpatrick, 2015). Yet, 75% of early elementary grades teachers report using the three-cueing system in their instructional approach to early reading (Kurtz et al., 2020, p. 11), and many teachers and education leaders are now appalled to learn that a foundational theory behind the curricula and instructional strategies they have been using for decades was discredited years ago (Hanford, 2019; Goldenberg & Goldenberg, 2022). Reading Recovery, with its use of cueing strategies and running records for miscue analysis, is one of the programs that have been targeted by critics of the three-cueing approach.

The Reading Recovery Program

Reading Recovery is a widely implemented teacher-led, one-to-one, short-term early literacy intervention targeting first-grade students experiencing difficulty learning to read and write. Created in New Zealand in the 1970s, the number of schools implementing the program has increased, decreased, and increased again during its 35-year history in the United States. Many proponents of Reading Recovery argue that it is one of the most effective early reading interventions available (RRCNA, n.d.), while critics of Reading Recovery argue that it uses outdated strategies that are not supported by the current research base (Wrightslaw, n.d.).

Children participating in Reading Recovery receive daily one-on-one instruction from a specially trained Reading Recovery teacher who follows a sequenced intervention protocol that begins with re-reading familiar books and a running record, then word or letter work on the wallboard, story composition, assembling a cut-up sentence, and finally previewing and reading a new book. Lessons last 30 minutes per day, and continue for 12 to 20 weeks. Information from each lesson, including miscue analysis of the running record to identify any substitutions, omissions, insertions, or repetitions of words that a student makes while reading aloud, are used by the teacher to inform strategies in the next lesson and assess and monitor the students' progress and instructional level. Once a student is assessed to read within the average range of their class and continue independent literacy learning in the regular classroom, their lessons are discontinued, and a new student enters the program. The program requires that students participate in Reading Recovery lessons at times that do not conflict with the regular literacy block or other important instructional time; thus, Reading Recovery is a supplement to regular classroom instruction.

Reading Recovery has been the subject of many research studies. Results from short-term impact evaluations of Reading Recovery to date have been largely positive, suggesting substantial gains on first-grade reading outcomes (see D'Agostino & Murphy, 2004; D'Agostino & Harmey, 2016; and U.S. Department of Education, Institute of Education Sciences, 2013 for reviews). There have also been a few studies of long-term impacts of Reading Recovery, but their designs do not support strong causal inference. A series of matched-comparison group studies from the UK found positive long-term impacts of Reading Recovery at ages 7, 9, 11 and 16 years old (Burroughs-Lange & Douëtil, 2007; Hurry & Sylva, 2007; Holliman & Hurry, 2013; Tanner et al., 2010). However, these studies matched Reading Recovery students to comparison group students from different school districts that did not offer Reading Recovery, yet their multilevel models treated the intervention indicator as a school-level variable instead of a district-level variable, and no data were presented to confirm the comparability of the schools or districts. A more recent study by D'Agostino, Lose, and Kelly (2017) using propensity score matching compared Reading Recovery students to other students from the same sample of schools. Results from that study were mixed for long-term impacts on students' 3rd and 4th grade scores on the Michigan state reading test. While significant positive impacts were observed for students identified as "most eligible" (ES = +.35) or "moderately eligible" (ES = +.34) for Reading Recovery, significant negative impacts (ES = –.51) were observed for students identified as "least eligible" (D'Agostino, Lose, & Kelly, 2017, p. 124). Notably, this study used a propensity score matching model with only three predictor variables and it achieved baseline balance on the pretest for only the "most eligible" subgroup. As such, there are still no studies of the long-term effects of Reading Recovery that could meet What Works Clearinghouse standards for rigor (WWC, 2022).

The most rigorous study of the impacts of Reading Recovery was a large-scale multi-site randomized controlled trial (RCT) funded by a $55M Investing in Innovation (i3) Scale-Up grant and conducted from the 2011-12 school year through 2014-15. This study confirmed evidence of large short-term impacts (ES>+.50) of Reading Recovery on students' learning in first grade (see Sirinides, Gray, & May, 2018; May, Gray, Gillespie, et al., 2015 & 2013; and May, Sirinides, Gray, & Goldsworthy, 2016). However, because of the delayed treatment design under the RCT, a separate and parallel regression discontinuity (RD) study was conducted to estimate the sustained effects of Reading Recovery beyond the first half of first grade.

Methods

Study Population

A random sample of schools participating in and funded by the i3 grant (i.e., i3 schools), along with a separate sample of non-i3 schools, implemented an RD design to enable estimation of long-term impacts on student reading achievement through 3rd and 4th grades. The study presented in this paper includes four i3 cohorts of students who were in 1st grade in i3 schools during the 2011-12 through 2014-15 school years, as well as one additional cohort of 1st graders in any Reading Recovery school implementing the RD design during the 2016-17 school year. Adding an extra cohort to this study allowed collection of 3rd grade test scores for this cohort in fall 2019, only a few months after students had completed their 3rd grade state tests in spring 2019, with the intent of increasing the proportion of students for whom 3rd grade scores could be obtained, as well as expanding the generalizability of the results to non-i3 schools.

Research Design

This study employs a multi-site RD design. Each participating school carried out the research procedures described below, including administration of a pre-intervention assessment,

assignment of students to the intervention based on pre-assessment scores, and collection of outcomes data at the middle and end of first grade and again at the end of 3rd grade.

The What Works Clearinghouse (WWC) has published standards for RD studies (WWC, 2022) that include five elements: (1) integrity of the forcing variable, (2) attrition, (3) continuity of the outcome-forcing variable relationship, (4) functional form and bandwidth for statistical models estimating impacts, and (5) fuzzy regression discontinuity. The results reported in the sections that follow address each of these standards.

Pre-Intervention Assessment (RD Forcing Variables)

The instrument used to assess students' reading performance at the beginning, middle, and end of first grade was the OS (Clay, 2005). The OS is the primary screening, diagnostic, and monitoring instrument for Reading Recovery. It is a one-to-one, teacher-administered, standardized assessment that includes six subscales: Letter Identification, Concepts about Print, Ohio Word Test, Writing Vocabulary, Hearing and Recording Sounds in Words, and Text Reading Level. Reported test-retest and internal consistency reliability estimates for the OS range from moderate to high on the individual OS Tasks (Clay, 2005). Measures of the inter-assessor reliability of the Text Reading and Writing Vocabulary tasks yielded coefficients of .92 and .87 (Denton, Ciancio, & Fetcher, 2006). In addition, several studies have assessed the construct and criterion validity of the OS tasks using the sub-tests of various norm-referenced tests, including the Iowa Tests of Basic Skills (ITBS). Across these studies, researchers have found that scores can be validly interpreted for the following purposes: (a) identification of at-risk students (Gomez, Rogers, Wang, & Schultz, 2005), (b) measurement of early reading constructs (Tang & Bellenge, 2007; Gomez, Gibson, Tang, Doyle, & Kelly, 2007), and (c) prediction of the attainment of performance benchmarks (Denton et al., 2006).

Selection of Schools

During the summer before the start of each school year (i.e., 2011-2014), each new school joining the Reading Recovery i3 Scale-Up was randomly selected to participate in one of three research designs during their first year. The three research designs included (1) the RCT, (2) the RD design, and (3) an internal study used by Reading Recovery to monitor and update national norms for the OS assessment. Each school was then rotated through these three research designs based on a predetermined order (e.g., RD design in 2011-12, RCT in 2012-13, OS norming in 2013-14, etc.).

Non-i3 schools are schools that participated in Reading Recovery, but did not receive any funding from the i3 grant. Each school year, the Reading Recovery International Data and Evaluation Center (IDEC) assigns about half of Reading Recovery schools to implement an RD design and collect comparison group data. A simple odd/even algorithm is used to determine if a school is collecting RD data. For example, for the 2011-12 school year, if the school's unique IDEC identification number is odd, then it collected RD data.

Collection of Long-Term Outcomes Data

The long-term outcome measures for this study include Reading/ELA scores from state tests administered to students in 3rd and 4th grades, as well as additional data documenting individual students' participation in reading interventions during 2nd, 3rd, and 4th grades. Collection of these data began in October 2017 and concluded in January 2020. Reading Recovery teachers, teacher leaders, or school administrators were asked to submit data for individual students through a dynamic survey built using the Qualtrics platform. Some schools were excluded from long-term data collection due to insufficient sample sizes within schools (i.e., fewer than seven students participating in Reading Recovery, or fewer than three

comparison group students), and some students were excluded due to missing pre-intervention scores. The proportion of schools and students excluded from long-term data collection are presented in the section on sample attrition.

Assignment of Students

In each school participating in the RD study, Reading Recovery teachers were instructed to administer the OS at the beginning of the school year to all students who were selected for Reading Recovery screening, plus four to twelve additional students who were among the lowest performers in reading, but who were not previously selected for Reading Recovery screening. Typically, this resulted in pre-intervention testing of eight to 25 students with the OS at the beginning of first grade in each school.

Teachers were instructed to then rank order the students by their OS scores and assign those students with the lowest scores to Reading Recovery, working from lower to higher scores until all open Reading Recovery slots were filled. This typically resulted in four to twelve students assigned to Reading Recovery during the first half of the school year and an additional four to fourteen students in each school assigned to a waitlist.

As the first wave of Reading Recovery students completed their intervention approximately midway through first grade, Reading Recovery teachers were instructed to test all previously assessed students again using the OS. In addition, any new students who had been selected for Reading Recovery screening midway through first grade were also assessed with the OS. Those students who had not previously participated in Reading Recovery were then rank ordered based on their OS scores, and those students with the lowest scores were assigned to Reading Recovery in order of lower to higher scores until all open Reading Recovery slots were

filled. This typically resulted in four to twelve students assigned to Reading Recovery and one to seven students assigned to a waitlist in each school during the second half of the school year.

Because the RD design in the initial study involved two distinct waves with separate assignment processes, each wave is treated as a separate RD design. More specifically, the first wave allows us to use the OS scores from the beginning of first grade as the forcing variable in an RD design analysis to examine impacts of receiving Reading Recovery in the fall of first grade. The second wave of the RD uses the OS scores from the middle of first grade as the forcing variable to examine impacts of receiving Reading Recovery in the spring.

Although Reading Recovery requires that students with the lowest OS scores be selected for the intervention, there is no predetermined formula for creating a composite OS score to assign students to the intervention. As such, Reading Recovery teachers are expected to use the information across the subscales to rank the students according to the following process:

> *Select the lowest-achieving students first. Generally speaking, these are the children with the most stanine 1s, 2s, and 3s on the six tasks of the Observation Survey. But also take into account the students' responses and raw scores on the individual tasks. Where children's profiles are similar, select the student with the least evidence of problem-solving activity (monitoring, self-correcting, initiating solving). Also, ask the classroom teacher who is gaining the least from classroom instruction (North American Trainers Group, 2015).*

This suggests two things: (a) the relative weight of each subscale for determining assignment to Reading Recovery is likely to vary from one school to another given differences in level and variation of OS subscale scores in different schools, and (b) there is the potential for inclusion of

information beyond the OS subscales in determining assignment when students have similar OS score profiles.

To deal with the first issue, we used a statistical model to estimate OS subscale score weights separately for each school (see next section). To deal with the second issue, we used a "fuzzy" RD design (Trochim, 1984), which estimates impacts given noncompliance with pure OS-based assignment (Bloom, 2012). The next section describes how each school's compliance with the RD design was determined and how their subscale weights, OS composite scores, and cutscores were derived from the data.

Determining Cutscores and Composite OS Scores

To derive each non-i3 school's OS composite scores and ranking of students for selection into Reading Recovery, we utilized a strategy described by Reardon and Robinson (2012) called "binding score" RD, in which we estimated a separate logistic regression model for each school that predicted assignment to Reading Recovery based on the smallest subset of OS subscale scores that would yield perfect (or best possible) prediction. More specifically, we estimated all possible logistic regression models using all possible combinations of one to six subscales, and then identified the model with the fewest predictors that produced a concordance index of 1.00 (i.e., perfect prediction) and quasi-complete or complete separation in the data (Allison, 2012). Schools for which perfect prediction was not possible using all six OS subscales were deemed to have implemented an imperfect RD, which is reflected as noncompliance in the fuzzy RD analyses. Logistic regression models were run and composite scores calculated for assignment to Reading Recovery separately at the beginning and at the middle of first grade (i.e., separately for each wave of the RD).

The final model for each school was used to calculate a predicted value for each student (p_{ik}) representing a composite score where each contributing subscale was weighted by its regression coefficient for school *k*. These scores were then normalized using Blom's (1958) method via PROC RANK in SAS 9.4 and rescaled to create final pretest OS composite scores ($Pretest_{ik}$) on a standard normal (i.e., z-score) scale with lower scores associated with assignment to Reading Recovery. The cutscore for each school ($Cutscore_k$) was then set as a random value from a uniform distribution ranging between the $Pretest_{ik}$ score for the student with the highest score in school *k* who was assigned to Reading Recovery and the student with the lowest $Pretest_{ik}$ score in school *k* who was not assigned to Reading Recovery. This allows the cutscore to be placed where all slots for participation in Reading Recovery in a given school are filled. Each student's composite OS score was then centered around the cutscore for that student's school (i.e., $Pretest_{ik} - Cutscore_k$).

Integrity of the Forcing Variable and Cutscores

OS testing was completed and scores recorded before the cutscore was determined for each school. This precluded teachers from manipulating students' "true" OS scores as a means of influencing treatment assignment (WWC, 2022; Schochet, et al., 2010). One could argue that manipulation could occur as a result of differentially weighting OS subscales in order to influence assignment to Reading Recovery. For example, a teacher could choose to emphasize one subtest more in order to justify assignment of a particular student, or they could justify assigning a specific student based on a teacher's recommendation in addition to OS scores. Unfortunately, there is no way to test directly for such manipulation. We pooled data across cohorts and assessed the continuity of the forcing variable through a McCrary test (2008) and histograms depicting the distribution of each OS composite forcing variable. As an indirect

method for assessing possible manipulation of the subscale weighting as well as the possibility of model overfitting due to the small sample sizes within schools, we also used as the RD forcing variable the OS Total Score derived independently of the i3 project by IDEC using a partial-credit Rasch model (D'Agostino, Rodgers, & Mauck, 2018). This approach uses consistent subscale weights in every school, although it produces less-perfect prediction of participation in Reading Recovery. This imperfect prediction is handled as non-compliance with treatment assignment under a fuzzy RD design described in the next section.

Fuzzy RD Assignment

The RD design does not require perfect compliance with the cutoff-based assignment result. Instead, it is sufficient to confirm a substantial and sharp difference in participation rates above and below the cutscore (Bloom, 2012). This constitutes a "Fuzzy" RD design in which students below the cutscore who did not participate in Reading Recovery are called "no-shows," and students above the cutscore who participated in Reading Recovery are called "crossovers."

Statistical Models of Impacts

Program impacts under the RD were estimated by comparing performance of students below and above derived cutoff scores for Reading Recovery eligibility in each school. Two sets of models were estimated. The first set produces "intent-to-treat" (ITT) estimates, which reflect the impact of scoring below the OS cutscore, regardless of whether a student did, or did not, end up participating in Reading Recovery. Therefore, the ITT estimates should not be interpreted as estimating the impact of Reading Recovery since the presence of no-shows and crossovers in the data will attenuate the ITT estimate toward zero. The second set produces effect of "treatment-on-the-treated" (TOT) estimates, which reflect the impact of participating in Reading Recovery after accounting for "no-shows" below the cutscore and "crossovers" above the cutscore. We

calculate TOT estimates via instrumental variables (IV) analysis (WWC, 2022; Angrist, Imbens, & Rubin, 1996).

The ITT analyses were conducted via multilevel statistical modeling, where the difference in performance of students above and below the cutscore was estimated, with students nested within each participating school. This multilevel model included the centered pretest assignment variable as a covariate at the student level, a parameter for the discontinuity associated with assignment to Reading Recovery (i.e., a 0/1 indicator for above/below the cutscore), and a random effect for overall school performance (i.e., a random school intercept) to account for the multilevel design. Initial models also included a random effect for the pretest slope (i.e., a random school slope), and a random effect for the impact of Reading Recovery (i.e., a random treatment effect across schools); however, the covariance estimates for these random effects were inestimable or not significantly greater than zero in nearly all models. Impact estimates did not differ substantially depending on whether these additional random effects were included or excluded; therefore, these random slope parameters were dropped from the final models. Because there is variability in schools' cutoff values on the raw OS scale (e.g., the lowest eight students in one school have OS scores that are higher or lower than the lowest eight students in another school), and the forcing variable is centered separately in each school, the generalizability of results beyond students near a single cutoff score is enhanced. Models for ITT impacts were estimated using PROC MIXED in SAS 9.4 via Restricted Maximum Likelihood (REML) with degrees of freedom based on within- and between-cluster sample sizes.

In accordance with WWC standards for RD studies (WWC, 2022), model fit and potential misspecification was assessed graphically via scatterplots and spline curves and also by testing for an interaction between pretest scores and the treatment assignment variable.

Assumptions of linearity in the RD analyses were further assessed by testing polynomial parameters and by imposing various restrictions on the bandwidth around the cutscore. More specifically, analyses were restricted to include progressively smaller subsets of students whose pretest scores fell within ±3.0 down to ±0.3 standard deviations of the cutscore. Lastly, robustness of the RD was assessed by arbitrarily shifting the cutscore up or down by ±0.5 standard deviations to confirm the absence of a discontinuity where none would be expected.

The TOT impact analyses were conducted using Instrumental Variables (IV) models with random effects for schools via generalized three-stage least squares (G3SLS) estimation using the approach of Balestra and Varadharajan–Krishnakumar (1987) as implemented in the *R* package *plm* (Croissant & Millo, 2008). The model uses the binary treatment assignment indicator (i.e., 1 = below vs. 0 = above the cutscore) as an instrumental variable predicting participation in Reading Recovery, with the OS test scores (i.e., the forcing variable) as an additional covariate. Impacts on long-term outcomes (i.e., 3rd and 4th grade test scores) are estimated using the exogenous component of participating in Reading Recovery (i.e., the component of participation in Reading Recovery that is associated with scoring above or below the RD cutscore, but is uncorrelated with endogenous confounding factors) and the OS test scores (i.e., the forcing variable) as an additional covariate.

In addition, we explored whether there might be evidence of a moderation effect of baseline performance, where the long-term impacts of Reading Recovery might be different for students who have exceptionally low or high pre-intervention scores. Because each school can be considered a mini-RD study with a school-specific cutscore, we included a standardized school-level mean of the OS forcing variable and its interaction with the cutscore indicator. Significance of the interaction would suggest that the long-term impact of Reading Recovery depends on the

school-level mean score of students tested with the OS in that school. For example, Reading Recovery might have larger long-term impacts for students who had the lowest OS scores at the beginning of first grade. We also tested for moderation based on schools' i3 status to determine whether the impacts were significantly different for i3 and non-i3 schools.

Lastly, we used the RD design to examine whether there were any discontinuities in (a) the rates of missing data for students just above and below the cutscores, and (b) the rates of follow-up intervention during 2nd and 3rd grades. Given the binary outcomes for these analyses, we used linear probability models with random effects for schools to estimate the ITT impacts of assignment to Reading Recovery on long-term data availability and follow-up intervention. A discontinuity in missing data rates would confirm the possibility of bias due to attrition, whereas a lack of a discontinuity in missing data rates would lessen the possibility of bias due to attrition. Similarly, a discontinuity in 2nd and 3rd grade intervention rates could help to explain long-term outcomes. For example, higher rates of intervention for students above the cutscore could be taken to suggest that the control group received extra support allowing them to catch up (and even pass) the Reading Recovery group. We also estimated instrumental variables probit regression models to produce TOT impacts of receiving Reading Recovery on long-term data availability and follow-up intervention, and results aligned with those from the linear probability models.

Replicating the Short-Term Impacts from the RCT

Before examining the long-term impacts of Reading Recovery using the RD design, it is helpful to confirm the validity of the RD design by comparing short-term impact estimates on 1st grade outcomes between the RD and RCT designs. Results from such a series of analyses are presented in May & Blakeney (2022) and confirm that for both i3 and non-i3 schools, the short-

term RD impacts of Reading Recovery in 1st grade are large and positive, and very similar to those produced from the original i3 RCT. Additional results presented here confirm that the short-term results are also replicated with the final analytic sample used to estimate the long-term impacts. This result helps to confirm the RD design used here as a valid method for estimating causal impacts of Reading Recovery.

Results

The sections that follow document several aspects of the RD study analyses including (a) attrition due to missing data, (b) visualization of the integrity of the forcing variables, (c) compliance with the forcing variable assignment, (d) estimates of the long-term impacts of Reading Recovery, including sensitivity and moderation analyses.

Attrition

Table 1 shows analytic sample sizes and attrition rates between 73 and 83 percent by treatment and control groups for the 31,881 students who had Fall OS scores and were eligible to be included in the wave 1 RD and the 20,187 students who had Mid-Year OS scores and were eligible to be included in the wave 2 RD. Across RD waves (i.e., Fall and Mid-Year) and treatment/control groups, the pre-intervention OS scores of students for whom we were able to collect 3rd and 4th grade test scores were consistently higher than those for students with missing 3rd and 4th grade test scores. However, the differences between students with and without long term outcomes was consistent between treatment and control groups (i.e., there was no significant differential attrition) across all wave 1 comparisons, and differences were statistically significant but small for the wave 2 comparisons. More specifically, the treatment students with 3rd grade scores has Mid-Year OS scores that were 4.47 points higher than those of treatment students who are missing 3rd grade data, while the difference in the control group was only 3.38

points. Also, the treatment students with 4th grade scores has Mid-Year OS scores that were 6.72 points higher than those of treatment students who are missing 4th grade data, while the difference in the control group was only 4.21 points. This suggests a differential attrition where the analysis sample for the treatment group had somewhat higher pre-intervention performance by approximately 1.1 points in 3rd grade and 2.5 points for 4th grade, which corresponds to effect sizes of +.02 and +.06 standard deviations on the OS scale.

School-level attrition was 64 percent for 3rd grade outcomes (N=728 from the full sample of 2,029) and 75 percent for 4th grade outcomes (N=503 from the full sample of 2,029). This suggests that most of the attrition was due to entire schools dropping out of the study or not participating in our process for collecting long term follow-up data. However, there was no significant difference (Welch's $t(977.39)=-1.64$, $p=.1023$) in the school-mean fall pretest scores between schools that remained in the RD analytic sample ($M=386.3$, $SD=24.1$) and schools that were dropped ($M=388.4$, $SD=27.8$).

We were able to acquire school demographic data for up to 1,664 (82%) of the 2,029 schools in the full RD sample, and analyses suggest demographic differences between participating and non-participating schools. Specifically, the sample of schools that remained in the RD study had a lower percentage of students eligible for free or reduced-price lunch (48% vs. 53%, $t(1,662)=-3.69$, $p<.001$), they had a higher percentage of white students (68% vs. 62%, $t(1,662)=4.23$, $p<.001$), the schools were less likely to have school-wide Title I eligibility (73% vs. 80%, $\chi^2(1, 1,200)= 9.16$, $p<.01$), and they were less likely to be in an urban locale (20% vs. 27%) and more likely to be in rural (48% vs. 45%) or suburban (32% vs. 28%) locales ($\chi^2(2, 1,664)= 9.74$, $p<.01$).

RD analyses using a multilevel linear probability model with RD bandwidths from ±3.0 down to ±0.3 standard deviations confirmed that availability of 3rd and 4th grade test scores was unrelated to assignment status. Only three models out of the forty models estimated showed significant differences in long-term outcomes data availability for students above and below the RD cutscore, and all of the estimated differences were 2.9 percentage points or smaller. This finding is also evident in the plots shown in Figure 1 where there are no apparent discontinuities in data availability rates just above and below the RD cut scores.

Integrity of the Forcing Variables

Figure 2 shows histograms of the two forcing variables. The graphs suggest excellent continuity in the forcing variables near the cutscore for both waves, and the McCrary tests confirmed this for both wave 1 (p = .506) and wave 2 (p = .627). Each graph is trimodal, suggesting three potential populations of students: (1) those with very low OS scores who had a near 100 percent chance of participating in Reading Recovery, (2) those with OS scores above and below the cutscore whose chances of participating in Reading Recovery were driven largely by their position above or below the cutscore, and (3) those with higher OS scores who had a near zero percent chance of participating in Reading Recovery. Additionally, the balancing score models produced perfect prediction of participation in Reading Recovery in 75 percent of schools for RD wave 1 and in 81 percent of schools for RD wave 2.

Compliance with Forcing Variable Assignment

Figure 3 shows a visualization of the participation rates in Reading Recovery and the two forcing variables across the RD waves, with substantial and sharp discontinuities confirming the appropriateness of a Fuzzy RD analysis. The top two graphs show that the Fall OS Composite score was a very accurate predictor of participation in Reading Recovery during fall of first

grade, with a substantial and sharp difference in participation rates above and below the cutscore. More than 85% of students who scored below the fall OS cutscore participated in Reading Recovery (i.e., less than 15% were "no-shows"), while fewer than 15% of students who scored above the fall OS cutscore participated in Reading Recovery (i.e., less than 15% were "crossovers"). This relationship is observed for both the full wave 1 sample (top left graph) and for the subset of the wave 1 sample for whom we were able to obtain long-term outcomes (top right graph). The middle two graphs show similar findings for the Mid-Year OS Composite score as a predictor of participation in Reading Recovery during spring of first grade using both the full sample and the subsample with long-term outcomes data.

The bottom two graphs in Figure 3 show that the Fall OS Composite score remains a very good predictor of participation in Reading Recovery at any time during first grade. Despite the much higher rate of crossovers above the cutscore, there is still a substantial and sharp difference in participation rates above and below the cutscore. More than 90% of students who scored below the fall OS cutscore participated in Reading Recovery in either fall or spring, while between 30% and 65% of students who scored above the fall OS cutscore participated in Reading Recovery in either fall or spring (i.e., up to 65% were "crossovers"). A similar relationship is observed for both the full wave 1 sample (bottom left graph) and for the subset of the wave 1 sample for whom we were able to obtain long-term outcomes (bottom right graph).

Short-Term Impact Estimates: Replication of RCT Results

Because the RD design was conducted in a parallel set of schools sampled randomly from the population of schools participating each year of the i3 scaleup of Reading Recovery, we can assess the validity of the RD design by comparing the estimates of short-term impacts in 1st grade based on RD schools and students to the impact estimates from the original RCT study

(ES= +.89 *SD*; Sirinides et al., 2018). A separate paper (May & Blakeney, 2022) presents results from such comparative analyses confirming that the RD design replicates the results from the RCT using the full sample of participating schools and students. However, the analyses presented in that paper did not account for attrition of students with missing 3rd and 4th grade test scores, which could induce bias in long-term impact estimates. Therefore, we implement here the same strategy and analyses from the May and Blakeney (2022) paper to compare short-term impacts from the original RCT and the RD sample after including only the students for whom we were able to obtain 3rd or 4th grade test scores. Results confirmed that, for the sample of students for whom we were able to obtain 3rd grade test scores (N=7,781), short-term impacts on 1st grade OS scores were statistically significant ($p<.001$), large and positive (IV-based TOT ES = +.77 *SD*), and that for the sample of students for whom we were able to obtain 4th grade test scores (N=4,917), short-term impacts on 1st grade OS scores were statistically significant ($p<.001$), large and positive (IV-based TOT ES = +.68 *SD*). Figure 4 shows RD scatterplots confirming the positive short-term impact for students below the RD cutscore.

Long-Term Impact Estimates

Results from the intent to treat (ITT) and treatment-on-the-treated (TOT) analyses for long-term impacts are shown in Table 2 by grade and RD wave. For both 3rd grade and 4th grade outcomes, and across both waves, the long-term ITT impact estimates are significant and negative. The raw impact estimates, which do not account for no-shows and crossovers, range from -.11 to -.16 standard deviations. This suggests that students with pre-intervention OS scores just below the RD cutscore had 3rd and 4th grade state test scores in reading/ELA that were .11 to .16 standard deviations below the state test scores of students who had pre-intervention OS scores just above the RD cutscore.

The TOT impact estimates, which account for which students actually received the treatment, are also negative and statistically significant for most bandwidths. Using an optimal bandwidth for the RD analysis, estimated long-term impacts of Reading Recovery on 3rd grade test scores were -.24 standard deviations (p = .144) and -.19 standard deviations (p = .003) for the fall and spring waves, respectively, and impacts on 4th grade test scores were -.30 standard deviations (p = .033) and -.43 standard deviations (p < .001) for the fall and spring waves, respectively.[1] Based on this IV analysis correcting for the presence of no-shows and crossovers, the TOT impact estimates show that students with pre-intervention OS scores just below the RD cutscore who participated in Reading Recovery in first grade had 3rd and 4th grade state test scores in reading/ELA that were .19 to .43 standard deviations below the state test scores of students who had pre-intervention OS scores just above the RD cutscore and did not participate in Reading Recovery.

Results from the treatment-on-the-treated (TOT) impact analyses via generalized 3-stage least squares (G3SLS) using progressively restricted bandwidths are shown in Figure 5 with TOT impact estimates (as solid black dots) and 95% confidence limits (as whisker lines) from G3SLS IV models with bandwidths (on the horizontal axis) as large as ±3.0 standard deviations down to as small as ±0.3 standard deviations, in 0.1 increments. Additionally, the right vertical

[1] For all of the G3SLS models, the relationship between the forcing variable and the cutscore indicator was highly significant (p<.001), which confirms a strong relationship between the instrumental variable (i.e., the cutscore indictor) and treatment receipt. This suggests that the instrument in each G3SLS IV model is not “weak” and satisfies the WWC standard for sufficient instrument strength (WWC, 2022).

axis of each graph shows the analytic sample size for each estimate (symbolized as a gray diamond). Results confirm that the TOT point estimates for both 3rd and 4th grade state tests scores and across both RD waves are consistently negative, although the standard errors of the estimates grow considerably and eventually become non-significant as sample sizes drop below 25% of the full sample.

Figure 6 shows scatterplots of the ITT RD analysis (i.e., without correcting for no-shows and crossovers) with a spline regression by grade and RD wave. A small negative discontinuity is evident below the cutscore (i.e., the red line on the left meets the cutscore midline just below where the red line on the right side meets the cutscore midline) in all four of the plots. This is consistent with the ITT estimates between -.11 and -.16 standard deviations.

As another robustness check, we also re-estimated the G3SLS IV models using the OS Rasch Total Score (D'Agostino, Rodgers, & Mauck, 2018) as the forcing variable, with school-specific cutscores set via the same method as the binding score models. This approach uses consistent weights of the six OS subscales in each school, thus avoiding the need for a statistical model to derive the forcing variable. At the student level, prediction of participation in Reading Recovery was nearly as accurate as the "binding score" approach. Of the students who scored below the fall OS cutscore, 76% participated in Reading Recovery in fall and 92% participated in Reading Recovery in either fall or spring. Of the students who scored above the fall OS cutscore, 25% participated in Reading Recovery in fall and 56% of students participated in either fall or spring. Thus, RD wave 1 included 8% "no-shows" and 56% "crossovers." For RD wave 2, 85% of students who scored below the mid-year OS cutscore participated in Reading Recovery in spring, while 32% of students who scored above the mid-year OS cutscore participated in Reading Recovery in spring. Thus, RD wave 2 included 15% "no-shows" and 32% "crossovers."

However, the number of schools with perfect prediction of participation was considerably lower than that from the binding score models, with only 13% of schools in wave 1 and 22% of schools in wave 2 with perfect prediction of Reading Recovery participation based on the OS Rasch Total Score.

Despite the lower accuracy of prediction based on the OS Rasch Total Score, results from the RD impact analyses using the OS Rasch Total Score were fairly consistent with results from the binding score approach presented earlier. For both 3rd grade and 4th grade outcomes, and across both waves, the TOT impact estimates were slightly smaller in magnitude, but were again negative and statistically significant across the range of bandwidths. Using an optimal bandwidth for the RD analysis with the OS Rasch Total Score, estimated long-term impacts of Reading Recovery on 3rd grade test scores were -.26 standard deviations for the fall wave (bandwidth= ±0.55 *SD*, $p<.001$) and -.14 standard deviations for the spring wave (bandwidth= ±.51 *SD*, $p=.002$). Impacts on 4th grade test scores were -.29 standard deviations for the fall wave (bandwidth= ±.58 *SD*, $p=.002$) and -.17 standard deviations for the spring wave (bandwidth= ±.70 *SD*, $p<.001$).

Sensitivity Analyses

Following the WWC standards for RD, we conducted several sensitivity analyses to examine whether the ITT and TOT impact estimates changed substantively when the analytic models included additional parameters. The first sensitivity analysis added an interaction between the cutscore indicator and the forcing variable to allow the relationship between pre-intervention scores and the outcome variable to differ on each side of the cutscore. There was no substantive change (e.g., >30% magnitude) either direction in any of the impact estimates.

The second sensitivity analysis added polynomial terms for the forcing variable (quadratic, cubic, and quartic) to allow the relationship between pre-intervention scores and the outcome variable to be curvilinear. Across all polynomial models, the impact estimates remained negative and grew to be very close to the impact estimates from the TOT analyses with restricted bandwidths.

The third sensitivity analysis involved a shifted cutscore test, where the cutscore was shifted up and down by 0.5 standard deviations, which reflects a point at which no discontinuity in participation or outcomes would be expected. Results from this analysis confirmed that there was no significant discontinuity in the regression model away from the original cutscore.

In sum, the RD analyses passed all sensitivity checks, suggesting that the RD results are unlikely to be biased due to model misspecification.

Moderation Analyses

In addition to the overall impact analyses, we explored whether there may be evidence of a moderation effect of i3 status or a moderation effect of baseline performance, where the long-term impacts of Reading Recovery might be different for students who have exceptionally low or high pre-intervention scores. We found no significant moderation ($p > .10$) of i3 status on 3rd grade impacts (recall that 4th grade data were not available for the cohort including non-i3 schools), indicating that impacts were not significantly different between i3 and non-i3 schools. For the baseline performance moderation analyses across the two RD waves and the two grades, three of the moderation analyses did not produce a significant moderation interaction ($p>.10$). However, the analysis based on RD wave 2 for 4th grade impacts showed a significant negative interaction ($b = –0.0747$, $p < .05$), suggesting that the long-term impacts of Reading Recovery may be less negative in schools where students begin first grade with exceptionally low OS

scores. For example, in a school where baseline OS scores are three standard deviations below average, the long-term impact estimate in that school would be calculated as −.1589 + [(−0.0747) × (−2.0)] = −0.0095, which is less negative, but still not positive.

Impacts on Subsequent Intervention in Grades 2, 3, and 4

To explore the possibility that follow-up intervention during grades 2-4 (or a lack thereof) might explain the results from the main impact analyses, we estimated multilevel linear probability models to explore potential differences in grade 2-4 intervention rates between Reading Recovery and control students. Analyses using Fall RD wave 1 data with varied bandwidths from ±.3.0 down to ±0.30 standard deviations (see the top two graphs in Figure 7) revealed no significant differences in subsequent intervention rates during grades 2-3 or grades 2-4. That is, RR students just below the cutscore were just as likely to receive additional intervention in grades 2-3 or grades 2-4 than control group students just above the cutscore. Analyses using Spring RD wave 2 data with varied bandwidths (see the bottom two graphs in Figure 7) revealed significant differences in subsequent intervention rates during grades 2-3 ($p < .01$) as well as grades 2-4 ($p < .01$). Analysis of the Spring RD data using the instrumental variables model described earlier suggest that Reading Recovery students just below the cutscore were predicted to receive additional intervention at least once during grades 2-3 about 60% of the time and in grades 2-4 about 61% of the time, while control group students just above the cutscore who did not receive Reading Recovery were predicted to receive additional intervention at least once during grades 2-3 only 47% of the time and in grades 2-4 about 52% of the time. These results were consistent using bandwidths ranging from ±0.3 down to ±3.0 standard deviations.

Discussion

Results suggest that the long-term impact of Reading Recovery on students' reading/ELA test scores in 3rd and 4th grades is statistically significant and substantially negative, and this finding is consistent across analytic methods and robust to sensitivity checks. Using growth-based benchmarks from Bloom et al. (2008, p. 305) showing that annual growth in 3rd and 4th grades is approximately .40 standard deviations, the TOT effect estimates of -.19 to -.43 standard deviations suggest that students who participated in Reading Recovery scored about one-half to one full grade level below the state test scores of similar students who did not participate in Reading Recovery.

Study Limitations

This study, like any, has several limitations. These include (a) the lack of a single explicit forcing variable and cutscore, which can mask the influence of additional information influencing treatment assignment, (b) very high attrition, and (c) the blending of long-term outcomes data from many different state tests that are not psychometrically equated. Perhaps the most salient of these limitations is the very high rate of missing data and attrition. Losing over 75% of the original sample during long-term follow-up is quite high. However, we saw little evidence of differential attrition, and there were no RD effects for missing data rates. Therefore, the high attrition is unlikely to have induced bias in the results. Moreover, the final analytic sample sizes were large enough to yield high statistical power and precision. But it is important to note that the analytic sample in this study may not be representative of the population schools implementing Reading Recovery, nor is sample is not representative of Title 1 schools nationally. There is also the issue of having a cutscore defined during the analysis phase, given that assignment to the intervention was based on test score rankings and dependent on the

number of Reading Recovery slots available in each school. This raises questions about whether there might be manipulation of the assignment mechanism near the cut score, or that the cut score derived through analysis is not an accurate proxy for the true unobserved cutscore, and this might induce bias in the RD analysis. However, the short-term impacts of the RCT were replicated with the RD using both the full sample, as well as the reduced sample based on only those students for whom 3rd or 4th grade scores were obtained. If there were bias associated with the balancing score method or the approach to estimating the cutscore in each school, that bias should be apparent in outcomes in all grades. Yet the first grade outcomes from the RD mirror the results from the large-scale RCT and appear to be relatively unbiased. Therefore, despite the limitations of this study, the results should be taken seriously as feasibly valid causal estimates of the long-term impacts of Reading Recovery.

One additional caveat is that the Regression Discontinuity design provides a "local average treatment effect" (Imbens & Angrist, 1994; Imbens & Lemieux, 2008) that represents the impact of the treatment near the RD cutscore. Given that D'Agostino, Lose, and Kelly (2017) found significant negative impacts of Reading Recovery on 4th grade state test scores in reading (i.e., albeit, with a non-equivalent comparison group), it is possible that Reading Recovery's negative impacts occur only for students near the RD cutscore; that is, students who have the highest pre-intervention reading scores among those selected to participate in Reading Recovery. However, given that the RD design in this study used school-specific cutscores that varied substantially, the negative impacts would need to be present for a wide range of pretest scores.

Conclusions

Acknowledging the methodological limitations, the negative findings from this study have at least two plausible explanations. Given the fervent disagreement in the early literacy

community about Reading Recovery and the theory and practices that it embraces, one of these two hypothetical explanations will likely be espoused and repeated selectively by those who advocate for Reading Recovery and the other by those who are critical of Reading Recovery. As such, it is important to acknowledge these competing hypotheses here. As independent evaluators, we do not have conclusive empirical data to determine which is correct; however, some of our results suggest which of these hypotheses is more plausible.

One hypothesis is that Reading Recovery produces large impacts on early literacy measures (i.e., in first grade), but these gains are lost when students do not receive sufficient intervention in later grades. Approximately half of the students in this RD study who participated in Reading Recovery during first grade went on to receive additional reading intervention in 2nd, 3rd, and 4th grades. It is certainly plausible to expect that many factors which lead to a student having exceptionally low reading skills at the beginning of first grade (e.g., low socioeconomic status, minimal exposure to books, dysfunctional home environment, low parental engagement) would persist even after school-based intervention. As such, it is not implausible to expect that gains achieved though participating in Reading Recovery might quickly disappear, and students may once again fall behind their peers, if they do not receive adequate support and engagement in reading in later grades. However, the analyses of RD effects on subsequent intervention rates in grades 2 through 4 presented in this paper show that students just below the RD cut score who participated in Reading Recovery were actually more likely to receive additional intervention at least once during grades 2-4 than were control group students just above the RD cut score. This suggests that the negative impact estimates from the RD analyses of state test scores cannot be attributed to inadequate rates of subsequent intervention since Reading Recovery students were found to actually be more likely to receive additional intervention. In other words, the argument

that the control group students caught up to and passed the Reading Recovery students due to higher rates of intervention in subsequent grades does not hold true based on the observed data. Furthermore, the results presented here show that the majority of Reading Recovery students actually did receive additional intervention at least once during grades 2 through 4, so the argument that they did not get enough follow-up intervention may be hard to justify. In other words, the Reading Recovery students actually got significantly more follow-up intervention than the control students, so a lack of follow-up intervention does not explain why students who participated in Reading Recovery actually fell significantly behind the control students by 3rd grade.

Another hypothesis is that the practices and strategies within Reading Recovery may produce large impacts on early literacy measures (i.e., reading first-grade books), but those practices and strategies do not translate to skills needed for continued success in later grades. While Reading Recovery includes "letter and word work" to help build phonemic awareness, students in Reading Recovery are expected to build phonics and decoding skills as part of the intervention's full set of activities and strategies. More specifically, Reading Recovery does not include isolated, systematic phonics instruction that is advocated by some literacy experts. As such, the program lacks a scope and sequence of sound-spelling patterns to be decoded, encoded, and read aloud in books. Given that much research, as reflected in the IES Practice Guide (Forman et al., 2016), concludes that systematic and explicit phonics instruction is essential for building decoding skills and mastering the morphophonemic aspects of the English language, then it is plausible that students who participate in Reading Recovery may not develop sufficient decoding and orthographic mapping skills that lead to success with larger words in 2nd and 3rd grade texts. Furthermore, this hypothesis may explain why, despite significant additional

intervention in 2nd through 4th grades, students who participated in Reading Recovery fall behind in later grades if the strategies they are taught actually hinder reading complex words.

One thing we can conclude with certainty at this point is that more research is needed to better understand why and when reading interventions lead to long-term impacts. The fact that Reading Recovery is one of the only interventions that has been subjected to multiple large-scale experimental and quasi-experimental studies of both short-term and long-term effects should be applauded. Too many popular reading programs and interventions have minimal to no evidence of short-term nor long-term effectiveness in school-based research that employs a rigorous causal design. The research and practice communities should work together to change that.

References

Allison, P.D. (2012). *Logistic Regression Using SAS: Theory and Application, Second Edition*. Cary, NC: SAS Institute Inc.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association, 91*(434), 444-455.

Balestra, P., & Varadharajan-Krishnakumar, J. (1987). Full information estimations of a system of simultaneous equations with error component structure. *Econometric Theory, 3*(2), 223-246.

Baenen, N., Bernhold, A., Dulaney, C., & Bankes, K. (1997). Reading Recovery: Long-term progress after three cohorts. *Journal of Education for Students Placed at Risk, 2*, 161-181.

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*. New York: John Wiley & Sons, Inc.

Bloom, H. S. (2012). Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness, 5*(1), 43-82.

Blom, G. 1958. *Statistical Estimates and Transformed Beta Variables*. New York: John Wiley & Sons, Inc.

Burroughs-Lange, S., & Douëtil, J. (2007). Literacy Progress of Young Children from Poor Urban Settings: A Reading Recovery Comparison Study. *Literacy Teaching and Learning, 12*(1), 19-46.

Clay, M. (2005). *An Observation Survey of Early Literacy Achievement*. Portsmouth, NH: Heinemann.

Connor, C. M., Morrison, F. J., & Katch, L. E. (2004). Beyond the reading wars: Exploring the effect of child-instruction interactions on growth in early reading. *Scientific Studies of Reading, 8*(4), 305-336.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software, 27*(2).

D'Agostino, J. V., & Harmey, S. J. (2016). An international meta-analysis of Reading Recovery. *Journal of Education for Students Placed at Risk, 21*(1), 29-46.

D'Agostino, Lose, & Kelly, 2017) Examining the sustained effects of Reading Recovery, *Journal of Education for Students Placed at Risk, 22*(2), 116–127.

D'Agostino, J. V., Rodgers, E., & Mauck, S. (2018). Addressing Inadequacies of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly, 53*(1), 51–69. http://www.jstor.org/stable/26622558

Denton, C.A., Ciancio, D.J., & Fletcher, J.M. (2006). Validity, reliability, and utility of the Observation Survey of Early Literacy Achievement. *Reading Research Quarterly, 41*(1), 8-

34.

Dimmery, D. [computer software] (2016). Package '*rdd*'. Manual for the statistical software 'R'.

Dong, N. and Maynard, R. A. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness, 6*(1), 24-67.

Foorman, B., Beyler, N., Borradaile, K., Coyne, M., Denton, C. A., Dimino, J., Furgeson, J., Hayes, L., Henke, J., Justice, L., Keating, B., Lewis, W., Sattar, S., Streke, A., Wagner, R., & Wissel, S. (2016). Foundational skills to support reading for understanding in kindergarten through 3rd grade (NCEE 2016-4008). Washington, DC: National Center for Education Evaluation and Regional Assistance (NCEE), Institute of Education Sciences, U.S. Department of Education. Available from: http://whatworks.ed.gov.

Goldberg, M., & Goldenberg, C. (2022). Lessons Learned? Reading Wars, Reading First, and a Way Forward. *The Reading Teacher, 75*(5), 621-630.

Gómez-Bellengé, F., Gibson, S., Tang, M., Doyle, M., & Kelly, P. (2007). *Assessment and Identification of First Grade Students at Risk: Correlating the Dynamic Indicator of Basic Early Literacy Skills and An Observation Survey of Early Literacy Achievement*. Retrieved from https://www.idecweb.us/

Gomez-Bellenge, F., Rodgers, E., Wang, C., & Schulz, M. (2005). *Examination of the validity of the Observation Survey with a comparison to ITBS*. Retrieved from https://www.idecweb.us/

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Literacy Research and Instruction, 6*(4), 126-135.

Hanford, E. (2019, August 22). What the words say. APM Reports. Available at https://www.apmreports.org/episode/2019/08/22/whats-wrong-how-schools-teach-reading

Holliman, A. J., & Hurry, J. (2013). The effects of Reading Recovery on children's literacy progress and special educational needs status: a three-year follow-up study. *Educational Psychology, 33*(6), 719-733.

Hurry, J., Fridkin, L., & Holliman, A. J. (2022). Reading intervention at age 6: Long-term effects of Reading Recovery in the UK on qualifications and support at age 16. *British Educational Research Journal, 48*(1), 5-21.

Hurry, J., & Sylva, K. (2007). Long-term outcomes of early reading intervention. *Journal of Research in Reading, 30*(3), 227-248.

Imbens, G. W. & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica, 62* (2), 467.

Imbens, G. & Kalyanaraman, K. (2009) *Optimal Bandwidth Choice for the regression discontinuity estimator*. NBER Working Paper Series. 14726. http://www.nber.org/papers/w14726

Imbens, G. W., & Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of Econometrics, 142*(2), 615-635.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*, 437–447.

Kilpatrick, D. (2015). *Essentials of assessing, preventing, and overcoming reading difficulties*. Wiley.

Kurtz, H., Lloyd, S., Harwin, A., Chen, V., & Furuya, Y. (2020). *Early Reading Instruction: Results of a National Survey*. Bethesda, MD: Editorial Projects in Education.

May, H. (2023, March 22). *An Efficacy Follow-Up Study of the Long-Term Effects of Reading Recovery*. Public-release data and analysis code available at http://www.osf.io/jc9bm

May, H., Gray, A., Gillespie, J., Sirinides, P., Sam, C., Goldsworthy, H., Armijo, M., Tognatta, N. (2013). *Evaluation of the i3 scale-up of Reading Recovery: Year one report, 2011-12*. Philadelphia, PA: Consortium for Policy Research in Education (CPRE); Newark, DE: Center for Research in Education and Social Policy (CRESP).

May, H., Gray, A., Gillespie, J., Sirinides, P., Sam, C., Goldsworthy, H., Armijo, M., Tognatta, N. (2015). Year one results from the multi-site randomized evaluation of the i3 scale-up of Reading Recovery. *American Education Research Journal, 52*(3), 547-581.

May, H., Goldsworthy, H., Armijo, M., Gray, A., Sirinides, P., Blalock, T., Anderson-Clark, H., Schiera, A., Blackman, H., Gillespie, J., & Sam, C. (2014). *Evaluation of the i3 Scale-up of Reading Recovery: Year Two Report, 2012-13*. Philadelphia, PA: Consortium for Policy Research in Education (CPRE); Newark, DE: Center for Research in Education and Social Policy (CRESP).

May, H., Perez-Johnson, I., Haimson, J., Sattar, S., & Gleason, P. (2009). *Using state tests in education experiments: A discussion of the issues* (NCEE 2009-013). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

May, H., Sirinides, P., Gray, A., & Goldsworthy, H. (2016) *Reading Recovery: An Evaluation of the Four-Year i3 Scale-Up*. Philadelphia, PA: Consortium for Policy Research in Education (CPRE); Newark, DE: Center for Research in Education and Social Policy (CRESP).

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698-714.

North American Trainers Group (2015). *Rationales and Guidelines for Selecting the Lowest-Achieving First-Grade Students for Reading Recovery*. Worthington, OH: Reading Recovery

Council of North America.

Reading Recovery Council of North America. (2012). *Standards and Guidelines of Reading Recovery in the United States* (6th ed.). Worthington, OH: Author.

Reading Recovery Council of North America (RRCNA). (n.d.) Reading Recovery achieves strong results and ranks highest in evidence of effectiveness. Retrieved from https://readingrecovery.org/reading-recovery/research-evaluation/what-works-clearinghouse/

Reardon, S. F., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness, 5*(1), 83-104.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J.R., Porter, J., Smith, J. (2010). *Standards for Regression Discontinuity Designs*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf

Schwartz, S. (2020). Is this the end of 'three cueing'? *Education Week*. Available from https://www.edweek.org/teaching-learning/is-this-the-end-of-three-cueing/2020/12

Seidenberg, M. (2017). *Language at the Speed of Sight: How we Read, Why so Many Can't, and what can be done about it*. Basic Books.

Sirinides, P., Gray, A., & May, H. (2018). The Impacts of Reading Recovery at Scale: Results From the 4-Year i3 External Evaluation. *Educational Evaluation and Policy Analysis.* https://doi.org/10.3102/0162373718764828

Tang, M., & Gómez-Bellengé, F. X. (2007, April). *Dimensionality and concurrent validity of the Observation Survey of Early Literacy Achievement*. Paper presented at the annual meeting of the American Educational Research Association in Chicago, IL.

Tanner, E., Brown, A., Day, N., Kotecha, M., Low, N. & Morrell, G. et al (2010) *Evaluation of Every Child a Reader (ECaR)*. Department for Education Research Report DFE-RR14.

Trochim, William (1984). *Research Design for Program Evaluation: The Regression Discontinuity Approach*. Beverly Hills: Sage.

U.S. Department of Education, Institute of Education Sciences, What Works Clearinghouse. (2013, July). Beginning reading intervention report: Reading Recovery®. Available from http://whatworks.ed.gov

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). Available from https://ies.ed.gov/ncee/wwc/Handbooks.

Wrightslaw. (n.d.). Experts say Reading Recovery is not effective, leaves too many children behind: An open letter from reading researchers. Retrieved from http://www.wrightslaw.com/info/read.rr.ltr.experts.htm

Table 1. Analysis of Overall and Differential Attrition for 3rd and 4th Grade Impacts

		Treatment Group		Control Group		*p-value* for Differential Attrition
		Has 3rd Grade Test Scores	Missing 3rd Grade Test Scores	Has 3rd Grade Test Scores	Missing 3rd Grade Test Scores	
Fall OS Score	*Mean*	380.89	377.92	414.44	410.66	.6249
	SD	37.98	40.28	37.11	40.87	
	N	5,815	16,578	2,536	6,952	
Percent Attrition			74.03%		73.27%	.1579
Mid-Year OS Score	*Mean*	468.44	463.97	499.24	495.86	.0367
	SD	34.62	38.85	37.05	40.84	
	N	2,949	8,144	2,470	6,624	
Percent Attrition			73.42%		72.83%	.3578
		Treatment Group		Control Group		*p-value* for Differential Attrition
		Has 4th Grade Test Scores	Missing 4th Grade Test Scores	Has 4th Grade Test Scores	Missing 4th Grade Test Scores	
Fall OS Score	*Mean*	381.53	378.12	414.56	411.08	.6576
	SD	38.32	39.97	36.89	40.51	
	N	3,735	18,658	1,624	7,864	
Percent Attrition			83.32%		82.88%	.3400
Mid-Year OS Score	*Mean*	470.76	464.04	500.27	496.06	.0231
	SD	33.32	38.57	35.52	40.68	
	N	1,852	9,241	1,562	7,532	
Percent Attrition			83.30%		82.82%	.3644

Table 2. Intent to Treat (ITT) and Treatment on the Treated (TOT) Analysis of Long-Term Impacts of Reading Recovery by Grade and RD Wave

3rd Grade Impacts, Wave 1 RD Fixed-Effects Parameters	ITT Impact Estimate	ITT Std. Error	TOT Impact Estimate	TOT Std. Error
Intercept	-0.5286***	0.0246	-0.4087***	0.1250
Fall OS Score (cutscore-centered)	0.1072***	0.0167	0.1417~	0.0768
Impact Estimate	-0.1139***	0.0303	-0.2354	0.1611
Optimal Bandwidth			±.7648	
3rd Grade Impacts, Wave 2 RD Fixed-Effects Parameters	**ITT Impact Estimate**	**ITT Std. Error**	**TOT Impact Estimate**	**TOT Std. Error**
Intercept	-0.3965***	0.0261	-0.3721***	0.0387
Fall OS Score (cutscore-centered)	0.1164***	0.0191	0.0682	0.0581
Impact Estimate	-0.1215***	0.0331	-0.1929**	0.0641
Optimal Bandwidth			±.8967	
4th Grade Impacts, Wave 1 RD Fixed-Effects Parameters	**ITT Impact Estimate**	**ITT Std. Error**	**TOT Impact Estimate**	**TOT Std. Error**
Intercept	-0.4644***	0.0302	-0.2951**	0.1083
Fall OS Score (cutscore-centered)	0.0937***	0.0216	0.0725	0.0467
Impact Estimate	-0.1142**	0.0392	-0.3049*	0.1429
Optimal Bandwidth			±1.3133	
4th Grade Impacts, Wave 2 RD Fixed-Effects Parameters	**ITT Impact Estimate**	**ITT Std. Error**	**TOT Impact Estimate**	**TOT Std. Error**
Intercept	-0.3083***	0.0323	-0.1799**	0.0564
Fall OS Score (cutscore-centered)	0.0791***	0.0243	-0.2145*	0.1035
Impact Estimate	-0.1634***	0.0443	-0.4258***	0.0987
Optimal Bandwidth			±.7301	

Note: $^{***}p<.001$, $^{**}p<.010$, $^{*}p<.05$, $^{\sim}p<.10$; Optimal bandwidth was determined using the method described by Imbens & Kalyanaraman (2009) as implemented in the *R* package *rdd* (Dimmery, 2016).

Figure 1. Long-Term Outcomes Data Availability by Grade Level and RD Forcing Variable

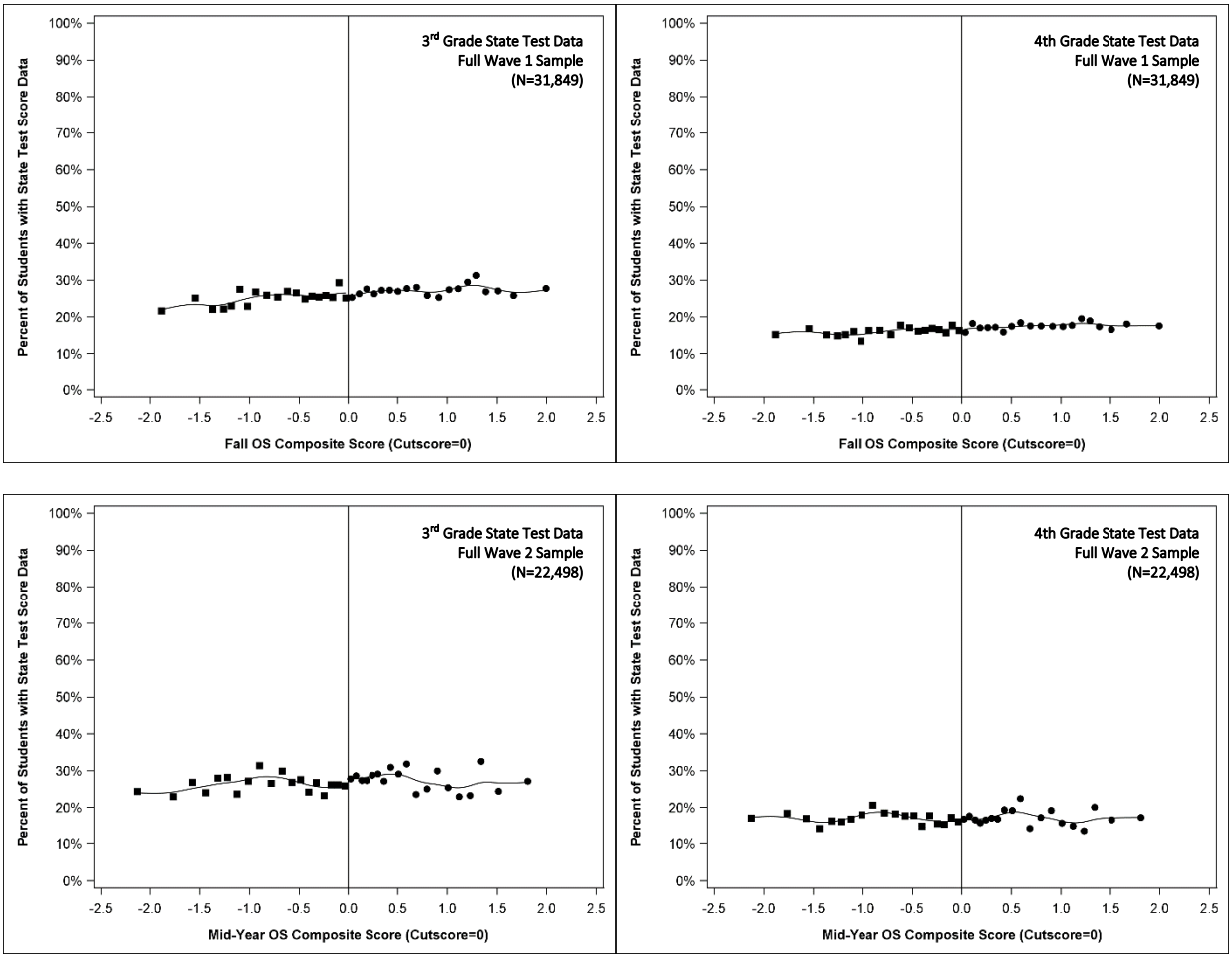

Figure 2. Histograms of Forcing Variables by RD Wave

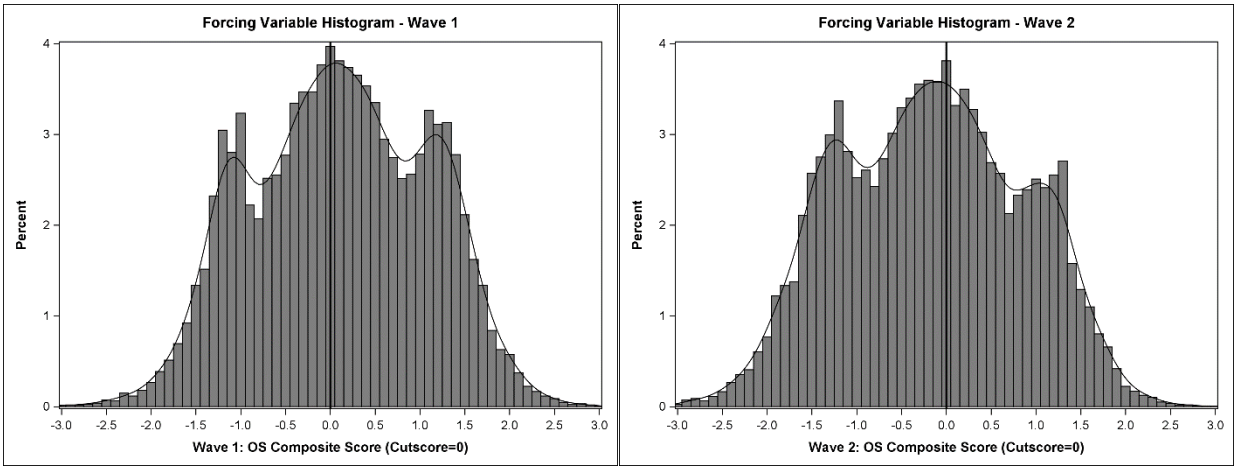

Figure 3. Reading Recovery Participation Rates by Forcing Variable and RD Wave

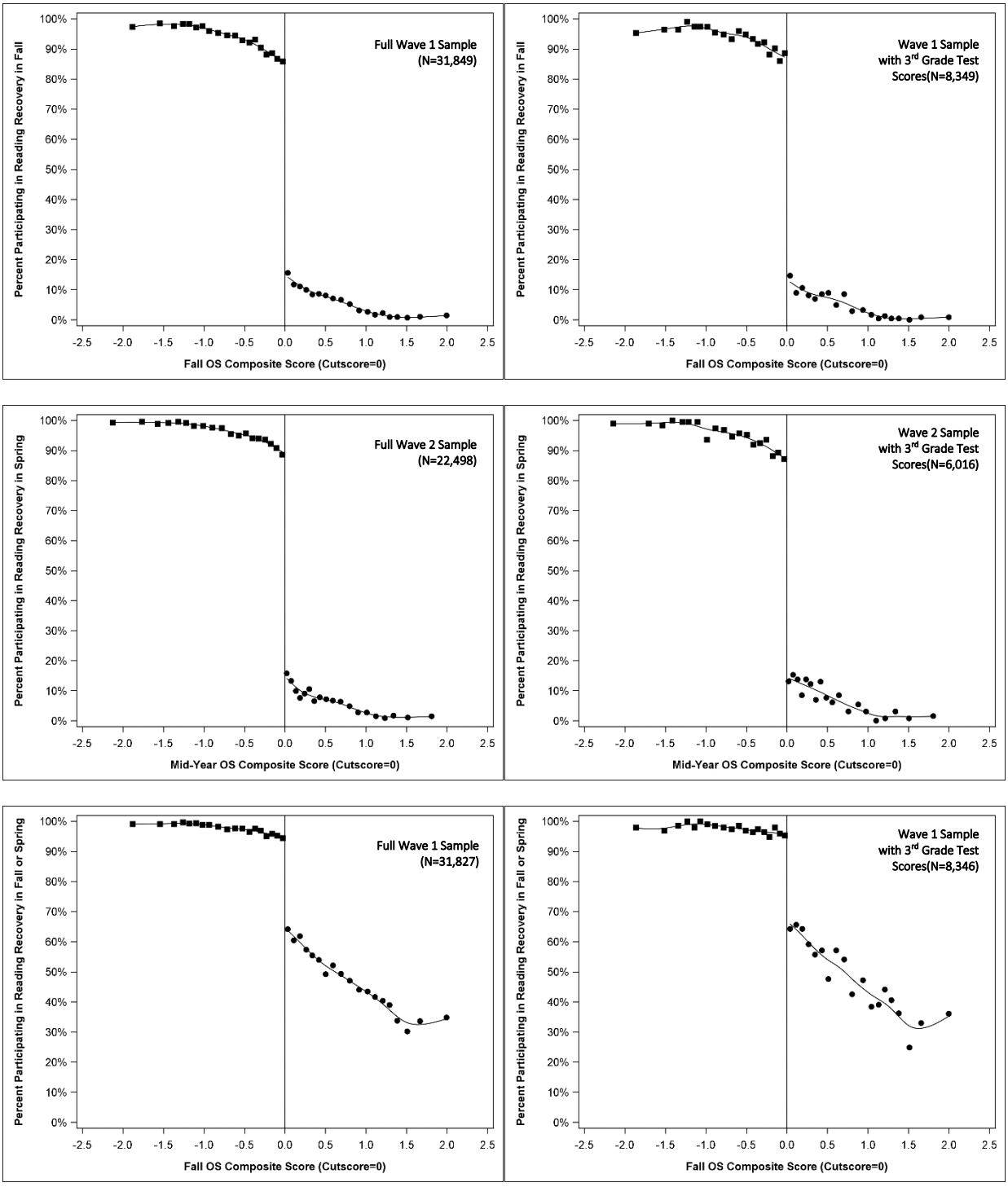

Figure 4. Short-Term Impacts (1st Grades) ITT Regression Discontinuity Scatterplots

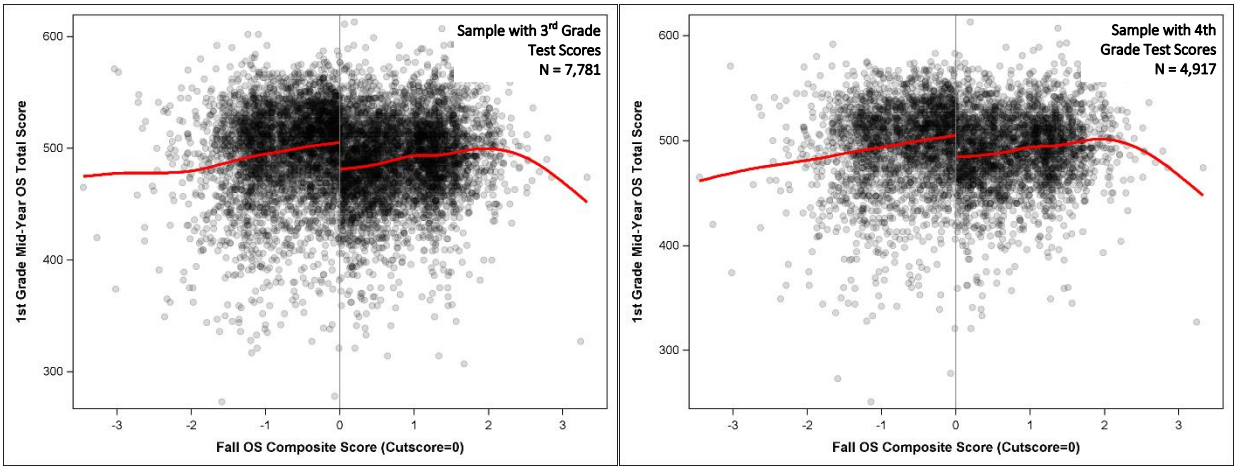

Figure 5. Treatment-on-the-Treated (TOT) Instrumental Variables Analysis of Long-Term Impacts Under Restricted Bandwidths from ±3.0 to ±0.3 standard deviations.

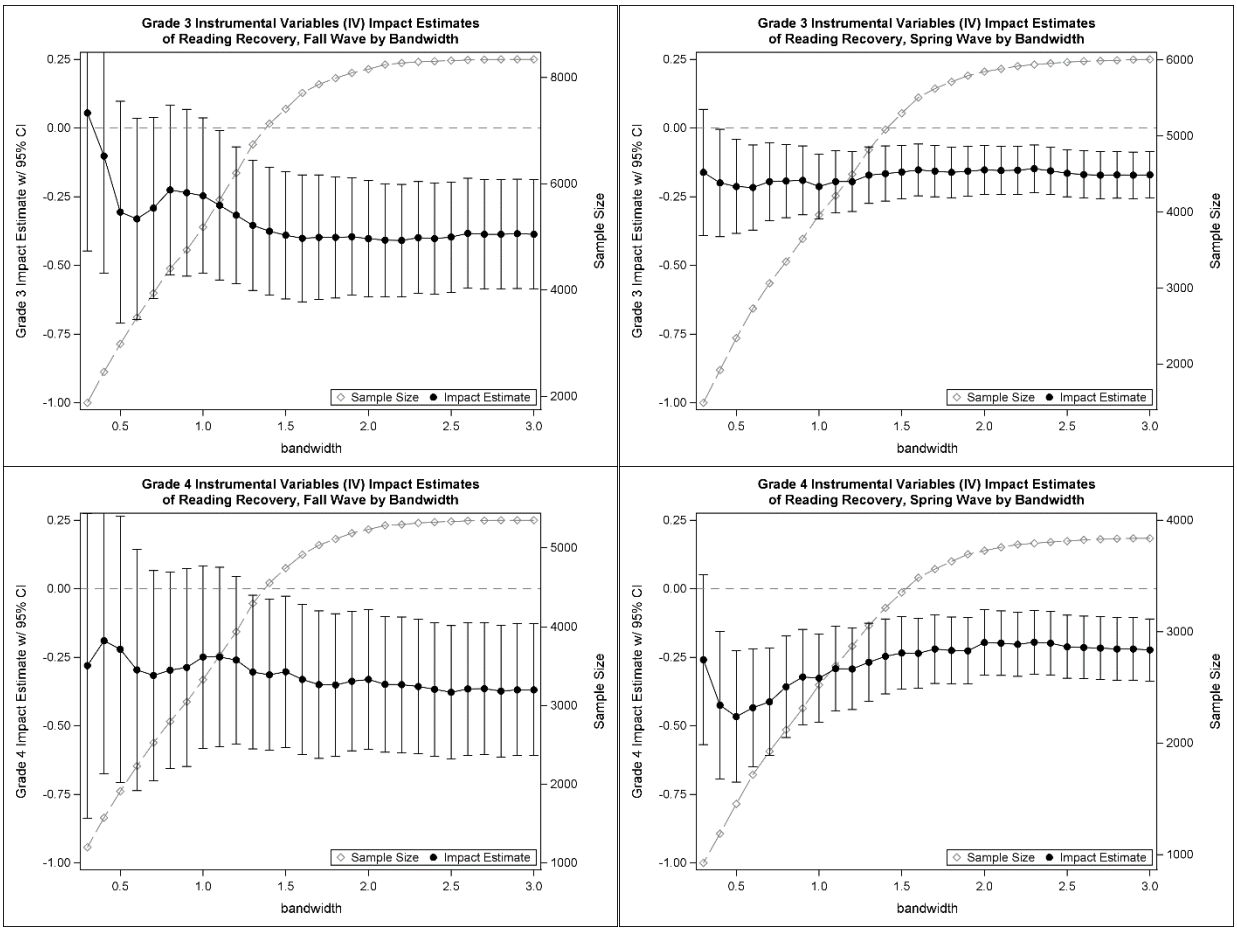

Figure 6. Long-Term Impact (3rd and 4th Grades) ITT Regression Discontinuity Scatterplots

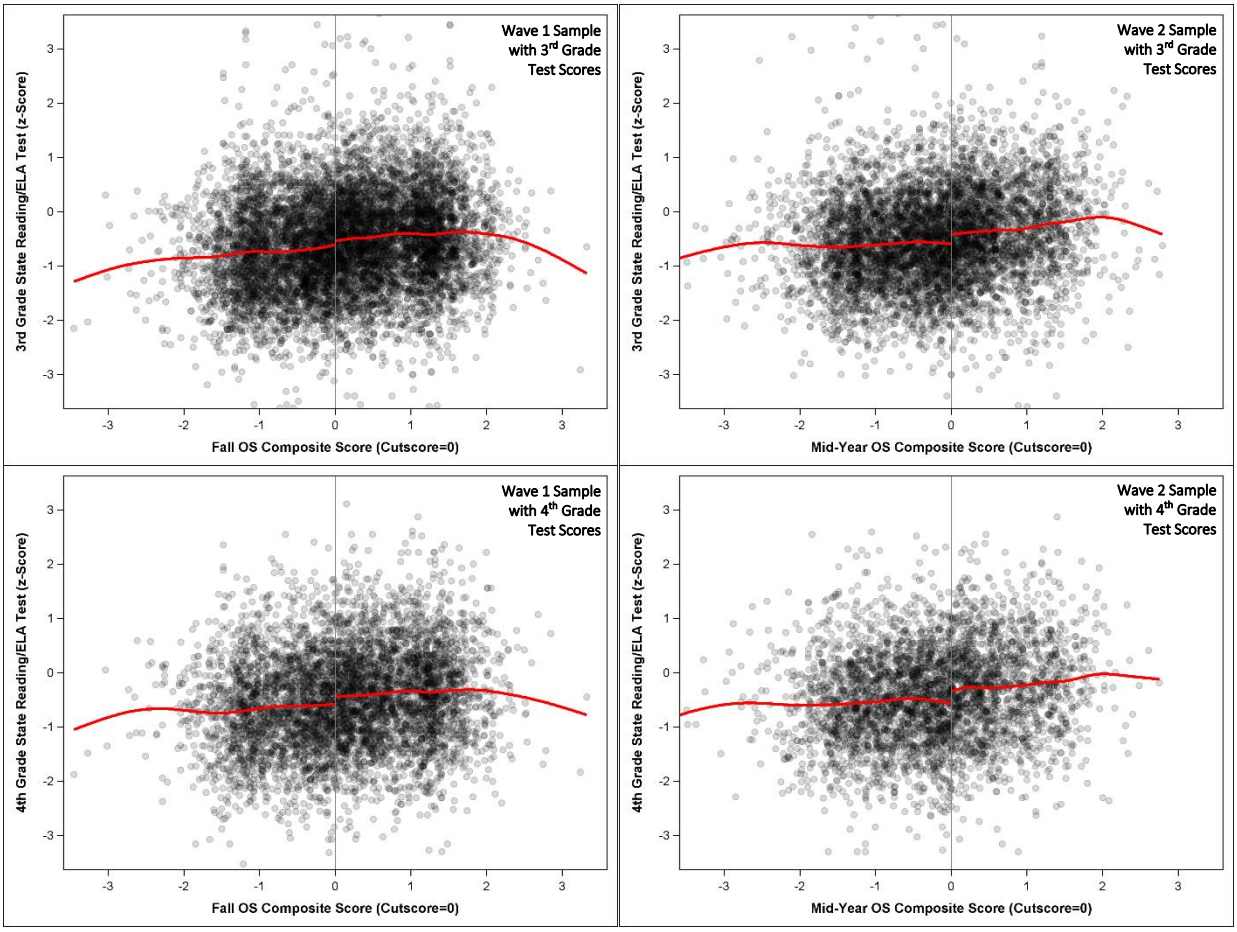
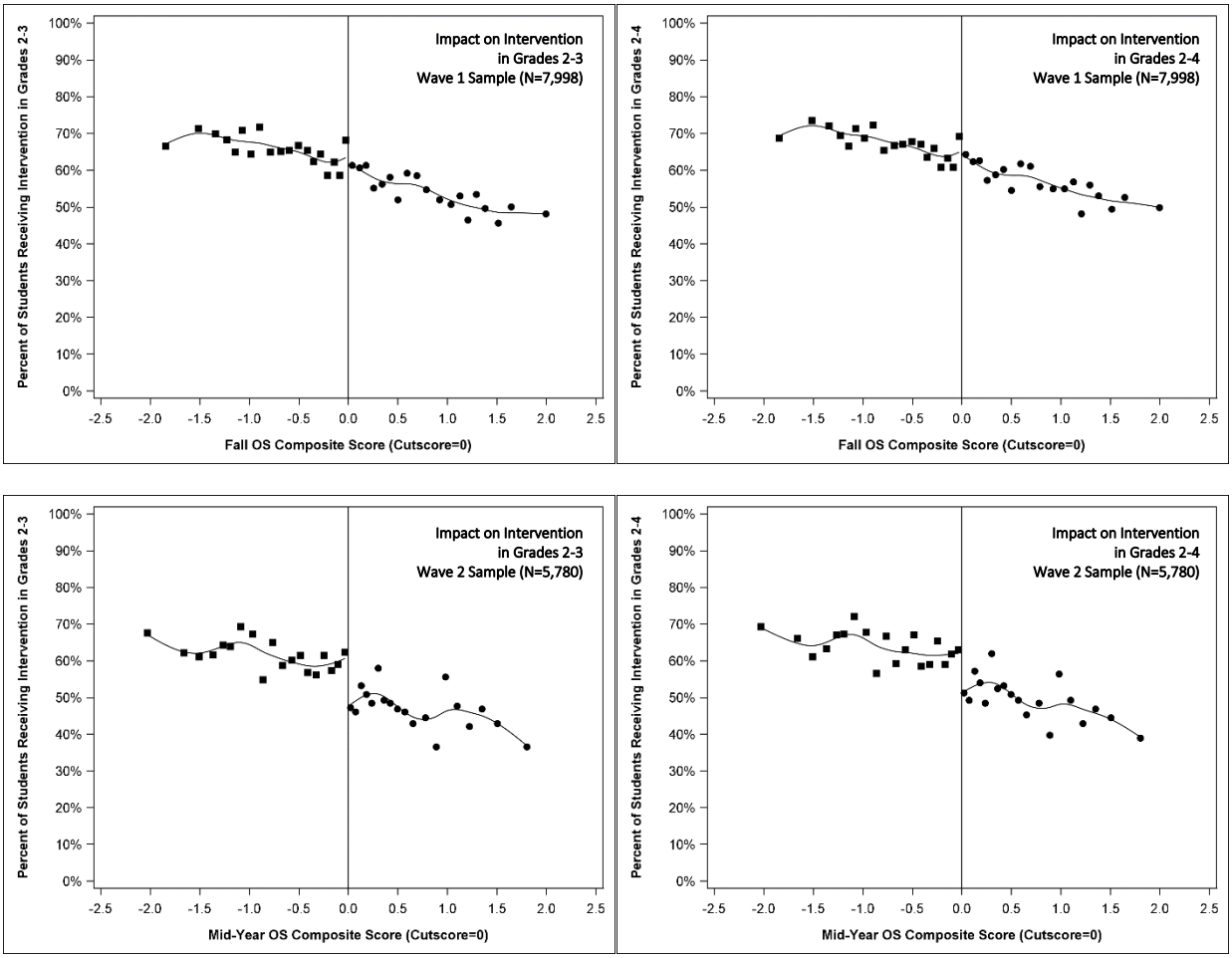
Figure 7. Intervention Rates in Grades 2-3 and 2-4 by RD Wave and Forcing Variable