

Note this Preregistration was accepted as a Stage 1 Registered Report at Exceptional Children.

**Examining differential intervention effects: Do Individualized Student Intervention effects vary by student abilities and characteristics?**

Wilhelmina van Dijk<sup>1</sup>, Christopher Schatschneider<sup>2</sup>, Stephanie Al Otaiba<sup>3</sup>, Holly B. Lane<sup>4</sup> and Sara A. Hart<sup>5</sup>

<sup>1</sup>Department of Special Education and Rehabilitation Counseling, Utah State University


<sup>2</sup>Department of Psychology, Florida State University

<sup>3</sup>Department of Teaching and Learning, Southern Methodist University


<sup>4</sup>Department of Special Education, School Psychology, and Early Childhood Studies, University of Florida

<sup>5</sup>Department of Psychology, Waterloo University

**Author Note**

Wilhelmina van Dijk  <https://orcid.org/0000-0001-9195-8772>

Christopher Schatschneider  <https://orcid.org/0000-0002-1700-7685>

Stephanie Al Otaiba  <https://orcid.org/0000-0001-7125-3791>

Holly B. Lane  <https://orcid.org/0000-0003-2663-8806>

Sara A. Hart  <https://orcid.org/0000-0001-9793-0420>

This work is supported by *Eunice Kennedy Shriver* National Institute of Child Health & Human Development Grants P50HD052120, R21HD072286 and R01HD095193. Views expressed herein are those of the authors and have neither been reviewed nor approved by the granting agencies.

**Abstract**

The Individualized Student Instruction (ISI) intervention was designed to help teachers increase their use of differentiated core reading instruction, to optimize student growth by providing appropriate amounts of code- and meaning focused instruction. Based on the results from original studies on ISI, it is still unclear if differentiated instruction can mitigate the influence of individual differences and if this is similar for all students. Using integrative data analytic techniques, we will combine data from six randomized control trials on the ISI intervention conducted in kindergarten and first grade, to obtain dataset with a total sample of 3,197. Conditional quantile regression models will be fit to gauge differential effects on word reading and vocabulary outcomes and the potential moderating effect of pre-intervention skills.

**Examining differential intervention effects: Do Individualized Student Intervention effects vary by student abilities and characteristics?**

Despite comprehensive efforts to increase the reading level of US students, only about one third of fourth graders reach proficiency (National Center for Education Statistics, 2019). Lack of reading proficiency is linked to poor academic, social, and economic outcomes (Sabatini, 2015). Providing high quality classroom instruction coupled with evidence-based supplementary intervention has been shown to reduce reading difficulties for many groups of students (e.g., Foorman & Moats, 2004; Foorman & Schatschneider, 2003). However, there are individual differences in how students respond to early instruction and intervention, with some students making greater gains and others benefitting less (Lovett et al., 2017; Pellegrini, 2001). Understanding the individual differences driving response to instruction and intervention can help researchers and practitioners make better predictions about which students may need extra help in reading, allowing more efficient allocation of precious time and resources (Al Otaiba & Fuchs, 2002; Lam & McMaster, 2014).

Reading proficiency is typically measured using assessments of reading comprehension. According to the simple view of reading, reading comprehension is the product of decoding and linguistic comprehension (Gough & Tunmer, 1986; Hoover & Gough, 1990). Skilled decoding includes multiple skills and mental processes, such as phonological decoding, orthographic mapping, and automatic recognition of words (Ehri et al., 2001; Lonigan et al., 2018; Nation, 2019). Linguistic comprehension refers to “the ability to take lexical information...and derive sentence and discourse interpretations” (Hoover & Gough, 1990, p. 131). This can include a wide range of skills and processes, such as vocabulary, syntax, inference-making, and listening comprehension (Catts et al., 2006; Oakhill & Cain, 2012; Ouellette & Beers, 2010). Numerous

studies have tested the simple view and supported the conclusion that decoding and linguistic comprehension are the primary contributors to reading comprehension across age and ability (e.g., Foorman et al., 2015; Kendeou et al., 2009; Lonigan et al., 2018). In fact, Lonigan et al. (2018) concluded that, across multiple models, 85% to 100% of reading comprehension variance was accounted for by latent measures of both decoding and linguistic comprehension combined.

The simple view holds as a viable explanation for the processes involved in developing reading comprehension, and it has also been used to explain reading difficulties. Gough and Tunmer (1986) suggested that the two components of the simple view can explain dyslexia (weak decoding, typical linguistic comprehension), hyperlexia (typical decoding, weak linguistic comprehension), and “garden variety” poor readers (weak in both areas). Its strength as a model for understanding reading difficulties has led to the development of many interventions for struggling readers that focus on word reading, vocabulary, listening or reading comprehension, or a combination of these. Intervention research has demonstrated that students’ word reading and comprehension skills increase when instruction and intervention are aligned and focused on developing young readers’ understanding of the alphabetic principle and decoding skills (Ehri et al., 2001; Vadasy et al., 2008), but the strength of the relation between decoding and comprehension weakens as students get older and the text they encounter becomes more complex (Catts, 2018; García & Cain, 2014). The relation between vocabulary and comprehension has also been well established (Bauman, 2009), but the strength of this relation increases as reading material becomes more challenging (Catts, 2018; Snow et al., 2007). However, the role of vocabulary and other types of linguistic-related instruction (i.e., syntax, morphology, pragmatics) in improving reading comprehension is complex (Snow, 2002), and individual differences appear to be a contributor to this complexity (Colenbrander et al., 2016). A comprehensive approach to

early reading instruction and intervention should include evidence-based practices for teaching decoding and vocabulary (National Institute of Child Health and Human Development (NICHD), 2000), but the focus of instruction should be based on an understanding of “the interactions between the skills that children bring to school and the instructional strategies they encounter in the classroom” (Connor, Morrison, & Katch, 2004, p. 332).

### **Individualized Student Instruction**

Individualized Student Instruction (ISI) is an intervention program originally developed by Carol Connor and her colleagues to help teachers increase their use of differentiated core reading instruction. Earlier studies indicated students’ skills on both word reading and vocabulary at the beginning of a year influenced how much growth students exhibit in word reading, vocabulary and reading comprehension; however, the growth also depended on the amount of time teachers spent on code focused or meaning focused instruction (Connor, Morrison, & Katch, 2004; Connor, Morrison, & Petrella, 2004). By helping teachers provide appropriate amounts of code- and meaning focused instruction, ISI meant to optimize students’ growth, not only in decoding and vocabulary skills but also in reading comprehension skills.

ISI has been described in detail previously (see for example, Al Otaiba et al., 2014; Al Otaiba et al., 2011; Connor et al., 2007; Connor, Morrison, et al., 2009; Connor, Morrison, Fishman, et al., 2011; Connor, Morrison, Schatschneider, et al., 2011) and we will provide a brief description. The ISI reading intervention consists of three main features, (a) a software program that supported data-based individualization and which recommended amounts of code- and meaning focused reading instruction for each student that were calculated based on student data at various points in the school year; (b) extensive professional development for teachers on how to use the software program and adapt instruction to meet students’ needs; and (c) coaching

for literacy instruction in the classroom through bi-weekly classroom-based observations and support as well as monthly meetings as communities of practice (Al Otaiba et al., 2011; Connor et al., 2013). Classroom instruction under ISI was conceptualized from the Simple View of Reading and supported teachers in providing the students with the appropriate amount of code- and meaning focused instruction in either teacher-directed small group settings or independent student centers. Activities and instruction followed core reading curricula that were adapted to meet the needs of the students, and were supplemented with other sources, such as activities from the Florida Center for Reading Research.

ISI was tested in several large scale RCTs and compared to business as usual (BAU) conditions where teachers were still expected to differentiate instruction. The results of the RCT studies largely suggest teachers increased their use of differentiating instruction and that the approach had a positive effect on students' reading skills compared to BAU instruction (Al Otaiba et al., 2014, 2016; Al Otaiba et al., 2011; Connor et al., 2007, 2013; Connor, Morrison, Fishman, et al., 2011; Connor, Morrison, Schatschneider, et al., 2011). In two of the studies, analyses showed that students performed better when they received amounts of code- and meaning focused instruction close to the amounts recommended by the software, but this was true in both intervention and control groups (Connor et al., 2007; Connor, Lara J., et al., 2009), suggesting that there was variation in the amount of differentiation in both ISI and BAU teachers and this variation affected student growth in similar ways.

In some of the published studies, the authors tested for individual differences based on pre-intervention skills, but results were variable. The interaction of pre-intervention skills and treatment effect in three studies (Connor, Morrison, Fishman, et al., 2011; Connor, Morrison, Schatschneider, et al., 2011; Connor, Piasta, et al., 2009) was not statistically significant.

Connor, Morrison, Schatschneider, et al. (2011) then compared of effect sizes for students scoring at the 25<sup>th</sup> and 75<sup>th</sup> percentile at pre-intervention showed slightly larger effects for students with lower pre-intervention scores. Authors noted the study was underpowered, which may have been the reason the interaction effect was not statistically significant. Connor, Piasta, et al. (2009) noted that students with lower fall word identification scores were less likely to receive appropriate amounts of meaning-focused instruction, presumably because teachers spend too much time on code-focused instruction. In a growth model, Al Otaiba et al. (2016) found that lower pre-intervention skills generally led to higher growth on both word reading and vocabulary outcomes; however, pre-intervention skill was not modeled as an interaction, but as a predictor in the growth model. A final set of three papers did not mention checking pre-intervention skills as moderators (i.e., Al Otaiba et al., 2014; Al Otaiba et al., 2011; Connor et al., 2013). Collectively, the results seem to suggest students with lower pre-intervention skills make more gains than students who start with more skills. It is thus possible ISI is effective at its goal: optimizing student growth through differentiated instruction. However, given the relatively small sample sizes for moderation analysis and the non-significant interaction results in the original studies, it is unclear if ISI truly mitigated the effects of pre-intervention skills and if there are differential outcomes for students across the distribution of their posttest scores.

The premise of ISI, to optimize students' growth by providing just the right amount of code- and meaning focused instruction, could lead to an additional potential moderating effect. Students with low preintervention word reading skills might show large growth in word reading, but low growth in vocabulary. While teachers are expected to provide both code- and meaning focused instruction, the ISI program would likely suggest teachers spend more time on code-focused instruction for students starting with low skills, especially when they have relatively

high vocabulary skills. However, students with both low preintervention word reading and vocabulary skills might make less growth in word reading, since they are likely to spend more balanced amounts of time on both code- and meaning focused instruction. In this last case, optimizing growth may lead to some growth in both areas, and not pronounced growth in one area.

### **Examining Differential Effects on Instruction**

Research on intervention effectiveness in reading aims to show participation in an intervention increases students' reading skills. After designing and running an experimental study, the statistical models most often used (i.e., ANOVAs or hierarchical linear models) to gauge the effect of interventions on student outcomes are based on the general linear model and evaluate the average treatment effect. These models come with an underlying assumption that the intervention is similarly effective for all students in the sample. In other words, the difference between the reading skills of students in the control and intervention groups is the same at the end of the intervention, no matter their final skills. Figure 1A illustrates this idea. In the figure, the dashed line represents the distribution of reading skills of a control group and solid line that of the intervention group. The distance between the two groups is 10 points, no matter how high or low the skills of the student. The estimation of the intervention effect at the mean (as is the default in general linear models) is thus representative of the effect for all participants.

As many researchers have noted, estimating the effect at the mean may conceal that an intervention, while on average effective, is not effective for specific students (e.g., Fuchs & Fuchs, 2019). Several other scenarios for an intervention effect are possible (see Porter, 2015 for an extensive explanation). First, an intervention may be more effective for students with lower skills and not so effective for students with higher skills when compared with their peers in a



control group. In this scenario, the intervention would help students with lower skills catch up faster. This scenario is depicted in Figure 1B. Second, the opposite scenario is also possible. An intervention may be more effective for students with higher skills. This possibility represents interventions that maintain a Matthew effect (Stanovich, 1986). Finally, we can imagine a scenario where an intervention is only effective for the average students. Both higher and lower achieving students do not benefit in this scenario. See Figure 1C for an illustration of this scenario. In each of these three scenarios, estimating the intervention effect with general linear models will not detect these differences. In fact, the estimates for the intervention effect in the four scenarios are the exact same: the intervention increases students' scores by 10 points.

Fortunately, a different type of model exists that can consider these possible differences in intervention effects. Quantile regression avoids the linear distributional assumption of the general linear model by estimating the intervention effect at different points along the distribution of the outcome variable (Koenker & Bassett Jr, 1978; Petscher & Logan, 2014; but see also Wenz, 2019). Whereas in ordinary least squares (OLS) based methods in the general linear model framework only one estimate for a relation is estimated (i.e., the average treatment effect), in quantile regression the parameters for this relation are estimated at pre-specified quantiles of the outcome distribution. A quantile is a cutpoint dividing a distribution in equal parts; a widely used quantile is the percentile, dividing a distribution in 100 equally sized parts. By estimating parameters at each pre-specified quantile, quantile regression can show how the intervention effect changes depending on a student's score on the outcome variable.

### **Previous Research on Differential Intervention Effects Using Quantile Regression**

The quantile regression approach has been used previously, albeit sparsely, to evaluate reading interventions and, in particular, to inform the important consideration for whom an

intervention is effective. For example, Wanzek et al. (2016) examined the effects of a Tier 2 literacy intervention. By first using linear mixed-models, they found an overall average effect of the intervention on one reading comprehension measure (i.e., Gates-MacGinitie Reading Tests; MacGinitie et al., 2000) but no effect on a second reading comprehension measure (i.e., Woodcock Johnson Passage Comprehension; Woodcock et al., 2007). The authors also employed quantile regression to determine if the intervention was more or less effective depending on students' reading comprehension scores post intervention. Outcomes from this analysis showed the intervention was effective for students with outcomes scores on the Gates-MacGinitie Reading Test between the 40<sup>th</sup> and 70<sup>th</sup> percentile, i.e., the middle of the distribution. The authors concluded the literacy intervention was least effective for students with low comprehension both at the start of the intervention and at the end of the intervention.

Similarly, Solari et al. (2018) also augmented the main effect analysis of a Tier 1 & 2 intervention with both an analysis of moderation of preintervention skills and a quantile regression analysis. The results from the main analysis showed the intervention had, *on average*, a statistically significant effect for almost all outcome variables, with pre-intervention word reading skills moderating the effect. The quantile regression analysis showed the intervention was statistically significant only for subsets of students. For decoding skills, the intervention was effective for those students at the 10<sup>th</sup> percentile; regarding word reading skills, students scoring between the 25<sup>th</sup> and 50<sup>th</sup> percentiles increased their skills statistically significantly. The oral reading fluency skills of students scoring above the 25<sup>th</sup> percentile increased significantly, as did the word reading fluency of students scoring above the 50<sup>th</sup> percentile. Finally, the intervention increased passage comprehension scores significantly only for students scoring at or above the 90<sup>th</sup> percentile. In contrast to the study by Wanzek et al. (2016), these results paint a less clear

picture of how individual differences influence intervention effects. However, combining results from both the quantile regression and moderator analysis, the results suggest interventions were more effective for students with low word reading skills prior to start of the intervention and at the end of the intervention. Students with low post intervention scores on either word reading, word fluency, or oral reading fluency, however, did not benefit as much as students at the highest end of the distribution. Due to the relatively small sample size and large number of parameters estimated, it is difficult to judge the generalizability of these results.

### **Previous Research on Differential Intervention Effects Using Moderator Analysis**

To understand which individual differences in students are related to lower response to interventions, most researchers have looked at moderator analyses in linear mixed methods frameworks (e.g., Coyne et al., 2019; Roberts et al., 2019). Moderator analyses show how effects of a treatment, and its statistical significance, are different along the distribution of the moderating variable. In other words, these analyses investigate if the effect of an intervention is dependent on individual differences in students' pre-intervention skill level. This technique has garnered considerable attention in recent years, including a special issue of *Exceptional Children* (Fuchs & Fuchs, 2019) completely dedicated to re-examining findings from randomized control trials (RCTs) to explore/identify pre-intervention characteristics that influenced treatment response of students. Beyond the articles in that special issue, there has been limited work done on moderator analysis of pre-intervention skills and effects of multicomponent reading approaches and the results of this work provide a blurred picture of the interaction between students' pre-intervention reading-related skills and multicomponent reading approaches.

### ***Word Reading Skills***

Word reading skills appear to interact with intervention effectiveness differently

depending on the grade level of students. For kindergarten students, it is unclear if pre-reading skills affect intervention outcomes. In a secondary analysis of a RCT of a code-based early reading intervention for kindergarten students, Hagan-Burke et al. (2013) found no significant interactions between pre-intervention alphabet knowledge (letter naming fluency and letter identification) or sound matching and the intervention. In a later study that included a second randomized control trial of the same intervention, however, Simmons et al. (2014) found students in the intervention condition who were better at sound matching pre-intervention scored higher on oral reading fluency, word identification, and passage comprehension post intervention.

In the early elementary grades, children with lower word reading skills generally benefit more from (or show stronger gains) in reading interventions. For example, Fuchs and colleagues (2019) found a multicomponent reading intervention was more effective for first grade students with lower word reading skills. These students gained more word reading, non-word reading, and reading comprehension skills than students who started the intervention with higher word reading skills. Similarly, Wolff (2016) showed students with lower decoding skills who had received a reading intervention in third grade, had greater gains in reading five years later.

### ***Vocabulary Skills***

Vocabulary seems to matter as a predictor of response to interventions primarily focused on increasing and strengthening students' vocabulary skills. For example, Coyne and colleagues (2019) showed that the effect of a vocabulary intervention on expressive vocabulary and listening comprehension was higher for students with higher vocabulary skills pre-intervention. In the case of comprehensive early reading approaches, however, the evidence that students' vocabulary skills are associated with response to intervention is mixed. Vadasy and colleagues

(2008), showed receptive vocabulary impacted growth on pseudoword and sight word reading from first through third grade, an indication that some of these code-based intervention outcomes could depend on children's vocabulary knowledge. Similarly, Lovett and colleagues (2017) found vocabulary interacted with intervention effects of a multicomponent reading intervention to increase students' reading comprehension outcomes. Students with higher vocabulary scores gained more in reading comprehension than students with lower scores. However, the interaction was not significant for any of the other student outcomes (e.g., word reading skills).

On the other hand, in the secondary analysis of the RCT by Hagan-Burke and colleagues (2013), there were no significant interactions between receptive vocabulary and intervention status on later decoding and phonemic awareness skills for kindergarten students. Extending this analysis to include a second randomized control trial, Simmons and colleagues (2014) showed that students' receptive vocabulary skills were predictive of post-intervention outcomes in the control condition, but not in the intervention condition.

### **Study Purpose and Research Questions**

For any instructional or intervention approach, it is important to understand for whom and under what conditions that instruction or intervention works. Unfortunately, results from general linear models may obfuscate contextual differences in interventions. In order for researchers and practitioners to make informed decision on which intervention to provide to whom, we need a better understanding if interventions have differential effects on students. Results from studies on ISI and other reading interventions (e.g., Al Otaiba et al., 2016; Fuchs et al., 2019; Simmons et al., 2014) hint at the presence of differential effects but are limited because they do not provide the specific range of children for whom the intervention was effective, or the size of the effect for each ability level. Other studies have shown different effects along the

outcome distributions but did not relate them to specific pre-intervention characteristics (e.g., Connor et al., 2007; Solari et al., 2018; Wanzek et al., 2016). While the outcomes from the analyses in these studies establish the presence of differential intervention effects due to certain student characteristics, they do not provide information on the specific range of students' skills that could yield more or less intervention effects, which is critical for understanding for whom certain interventions are likely beneficial.

The purpose of this study is to investigate if the effect of the ISI intervention is different for students with different post intervention skills, depending on their previous word reading and vocabulary skills. This approach expands on previous research by combining the quantile regression approach with a moderator analysis. Systematically examining the variation in student outcomes in ISI, specifically determining for which students the intervention was more, or less, effective and relating this to their pre-intervention characteristics, can inform our field regarding the effect of differentiated instruction and provide insight in which students likely will still need extra support to succeed in reading.

In this study, we will combine data from six different research projects on the ISI intervention. By using a combined data set, we are able to estimate effects across a larger sample of students, increasing the power of our final models to detect effects across the complete range of the distribution of the outcome scores.

In this study, we will examine the following research questions:

1. Do students who receive the ISI intervention outperform the comparison group on their (a) word reading skills and (b) vocabulary skills depending on varying points of the outcome distribution?

Following earlier results (Al Otaiba et al., 2016; Connor, Morrison, Schatschneider, et al.,

2011), we hypothesize the ISI intervention to be more effective for students in the lower quantiles in word reading and vocabulary.

2. To what extent do changes in (a) word reading skills and (b) vocabulary skills at varying points of the outcome distribution differ for students in treatment groups receiving the ISI intervention
  - i. based on pre-intervention word reading skills
  - ii. based on pre-intervention vocabulary skills

We hypothesize pre-intervention word reading skills will act as a moderator for outcomes on word reading and pre-intervention vocabulary skills moderate vocabulary outcomes. We expect the moderations to show lower effects on word reading outcomes for students with higher beginning word reading skills, and higher effects on vocabulary outcomes for students with higher beginning vocabulary outcomes. We hypothesize this moderation to be less strong at the higher quantiles. Given the nature of ISI, with recommended amounts of word reading and vocabulary instruction based on student skills, we hypothesize vocabulary outcomes at the lower quantiles are moderated by pre-intervention word reading skills, with lower word reading skills leading to less intervention effect. Similarly, lower initial vocabulary skills may moderate word reading outcomes at the lower quantiles. However, the effects will be much smaller in the ISI condition in comparison to the control condition.

## **Method**

### **Sample**

Data for this study comes from Project Kids (Daucourt et al., 2018). In this project, item level achievement and behavior data from eight independent data sets were pooled together with the intention of analyzing the data in innovative ways. Each of the original studies evaluated the

same comprehensive approach to early reading in the early elementary grades (K-3) in the same south-eastern state and spanned a complete academic year between 2005-2013. For this study, we will use data from six projects conducted in Kindergarten and Grade 1, including data on 3,197 participants clustered in 191 teachers. We provide short descriptions of each of the included datasets below. Because some of the studies were conducted in the same schools, care was taken to ensure that students who participated in more than one project, or in longitudinal studies, were represented only once in this analysis; hence the sample size of the original studies may differ from the sample sizes reported here. For a complete overview of Project KIDS data, sample, measures, interventions, and procedures see van Dijk et al. (2022).

### ***Project 1***

Data set 1 came from an iteration of the ISI applied with Kindergarten students and their teachers (Al Otaiba et al., 2011). The sample consists of 641 students in 44 classrooms; 362 students were in the treatment condition and received the ISI intervention and the 279 students in the control condition received typical classroom instruction (BAU).

### ***Project 2***

Data set 2 also came from a Kindergarten iteration of ISI (Al Otaiba et al., 2016). This sample consists of 514 students in 34 classrooms, 261 in the treatment condition (ISI) and 253 in the control condition (BAU).

### ***Project 3***

Data set 3 was taken from a study in which two types of response to intervention (RTI) models were compared (Al Otaiba et al., 2014). In the traditional model, students completed a cycle of classroom instruction only before being assessed and receiving supplemental intervention. In the dynamic model, students were immediately placed into intervention, if



pretest scores indicated at-risk status. In this study, 331 students in 34 first grade classrooms participated. All students received ISI instruction in Tier 1 and are considered as part of the treatment condition in this study.

#### ***Project 4***

Data set 4 came from an iteration of the ISI in first grade. Details about this study are described in Connor et al. (2007). The project consisted of 804 first grade students from 53 classrooms; 410 were in the treatment condition (ISI) and 394 in the control condition (BAU).

#### ***Project 5***

Data set 5 also included data from an ISI iteration conducted in first grade (see Connor, Morrison, Schatschneider, et al., 2011), and included 395 first graders from 26 teachers; 245 students were in the treatment condition (ISI) and 150 in the control condition (BAU).

#### ***Project 6***

Data set 6 included data from a three-year longitudinal study of the ISI intervention. Students in this sample were followed in first through third grade, and each year received either the ISI intervention or a math intervention (see Connor et al., 2013 for details). For the current study, we only used data from first grade. This included data on 512 first grade students, 279 of which were in the treatment condition (ISI) and 232 were in the control condition where they received the math intervention.

#### **Student characteristics**

Of the 3,197 students, 1,888 received ISI intervention. Table 1 shows the breakdown per project, and by Gender, Ethnicity, Race, LEP status, FRL eligibility, and ESE services.

Intervention and control groups are similar in make-up, with exception of gender ( $\chi^2 = 4.19$ ,  $df = 1$ ,  $p = 0.04^*$ ) and race ( $\chi^2 = 18.20$ ,  $df = 7$ ,  $p = 0.01^*$ ). In the final manuscript, we will also

include descriptive statistics (i.e., mean, standard deviation, and range) of pre-, and postintervention scores on the main variables of interest and show group differences using a two tailed t-test. Should groups be different, we will report Cohen's  $d$  as an indication of the size of the difference.

### **Procedures**

Each original study obtained IRB approval before project initiation. Project Kids obtained a separate IRB to procure, combine, and reuse the data of each of the original studies. During this component, data from the original studies were re-entered and checked. Students who had participated in more than one of the studies, were kept in the study in which they received intervention and removed from the other studies. In the current study, we will use moderated non-linear factor analysis (MNLFA), an integrative data analysis (IDA) technique, to generate scaled scores across the subset of projects and explore moderation on this combined data set.

### **Measures**

Each of the original RCTs used a large battery of cognitive ability and achievement measures to assess participants. These measures were administered by trained project personnel three times a year during the fall, winter, and spring. For this study, we will use scores on two subtests of the Woodcock Johnson- III (WJ-III; Woodcock et al., 2007, 2009) tests of achievement: *Letter word identification* (LWID) and *Picture Vocabulary* (PV) as representation of each of the components of the SVR (word recognition and language comprehension). Both subtests were administered in each of the original studies. We will use students' fall scores as pre-intervention predictors and students' spring scores as outcome variables.

#### ***WJ-III LWID***

The LWID subtest of the WJ-III (Woodcock et al., 2007, 2007) is a norm-referenced

standardized measure of word recognition. The test consists of 78 items, starting with letters and moving to increasingly more complex words, read in isolation. This subtest assesses students' ability to recognize words. The test is untimed. Test-retest reliability estimates of the norming sample range between .90 - .96 and split half reliability estimates range between .88 - .99 (McGrew et al., 2007). For the final report, we will include the internal consistency coefficient of our sample per project.

### ***WJ-III PV***

The PV subtest of the WJ-III (Woodcock et al., 2007, 2007) consists of 44 pictured objects and measures expressive vocabulary. During the initial items, students point to a picture corresponding to a vocabulary item and later move to naming the depicted objects. Test-retest reliability of the norming sample range between 0.70 – 0.81 and split half reliability ranges from .70 - .93 (McGrew et al., 2007). For the final report, we will include the internal consistency coefficient of our sample per project.

### **Data Preparation**

Data for the current project are a subset of a single, publicly available dataset (Hart et al., 2021). To ensure scores on the reading measures are unbiased and represent the same scale, we will perform several steps detailed below.

### ***Exclusion of participants***

Students with missing variables on both the pre- and post-intervention scores will be excluded from the analysis. Assuming the 4% of participants with missing pre-intervention scores on word reading also have missing post-intervention scores, this leaves a minimum sample size of 3,069 to be used for analysis for word reading. Three percent of participants have missing pre-intervention scores on vocabulary, under the same assumption, the minimum sample

size for analyses of vocabulary would be 3,101.

### ***Missing data***

Table 2 shows the distribution of missing data on the key variables of interest. The minimum expected sample sizes for each research question range from 2,578 to 2,717 participants (see Table 3 for minimum expected sample sizes per RQ). Since all assumptions on the patterns of missing data are untestable (Rhoads, 2012) and we have no reason to believe an unobserved variable is the cause of missingness, we are assuming data is missing at random (MAR). There are two ways to deal with MAR, multiple imputation (MI) (Rubin, 2004) and full information maximum likelihood estimation (see Rhoads, 2012). The *lqmm* package uses a ML estimator (i.e., asymmetric Laplace likelihood) with casewise deletion and is therefore not an appropriate way to handle our missing data. Instead, we will conduct a MI strategy (see Appendix for details).

### ***Data Integration***

To ensure data represent values on the same scale, we will use Moderated Non-Linear Factor Analysis (MNLFA; Curran et al., 2014), an IDA technique. Generally, IDA can be defined as “the analysis of a single data set that consists of two or more separate samples that have been pooled into one” (Curran & Hussong, 2009, p. 83). While to date not often used in educational science (e.g., Jansen et al., 2020), IDA has been used in health research, epidemiology, and developmental psychology (e.g., Daucourt et al., 2018; Hornburg et al., 2017; Leijten et al., 2018). However, IDA is an ideal methodology for pooling educational intervention studies together, because it capitalizes on between-study variability, for example variability that arises from differences in sampling techniques, timeframe in which a study was conducted, overall study design, and measurement (Curran & Hussong, 2009). Additional advantages of

IDA are 1) increased statistical power associated with the larger sample, 2) greater heterogeneity of the sample increasing generalizability to the population, 3) higher occurrence of low-base rate behaviors allowing for sub-group analysis in the pooled sample, and 4) stronger psychometric properties of measurement of a construct through pooling items (Curran & Hussong, 2009).

Specifically, MNLFA builds on the confirmatory factor analysis framework of creating a score on a latent factor that is representative of the construct of interest. To estimate scaled scores across the independent data samples, MNLFA tests for measurement invariance across potential influential covariates at both the factor and item level (Curran et al., 2014), using raw, individual level data. To create the scaled scores, we will perform the steps outlined by Curran et al. (2014) and Gottfredson et al. (2019) using the *amnlfa* package in R and Mplus.

To select items for both word reading and vocabulary that are representative items for each construct, we will retain items on each of the measures that have at least 5% coverage across the sample. For word reading, these are items 1-57 of the WJ-III LWID subtest, and for vocabulary these are items 1-39 of the WJ-III PV subtest. The second step involves identifying potential influential sources of variation between data samples. For this model, we will use project, gender, and age as potential sources of variation.

In the third step, we will use confirmatory factor analyses (CFA) to formally test if the retained items are representative of one factor. We will evaluate the CFA models by looking at Steiger-Lind Root Mean Square Error of Approximation (RMSEA) value, Bentler Comparative Fit Index (CFI) value, and the Standardized Root Mean Square Residual value (SRMR). An acceptable model will have an RMSEA value of  $\leq .05$ , CFI  $> .95$ , and SRMR  $\leq 0.10$ . Due to our sample size, we will disregard the model chi-square with its degree of freedom and  $p$ -value. If unidimensionality is not supported, we will cull items starting with those with the lowest factor

loadings.

Then, we will evaluate factor and items differences as a function of covariates using MNLFA. We will use a calibration sample consisting of one, randomly drawn, observation per participant (from either pre or postintervention scores). We will regress factor means and variances on project, gender, and age; all effects will be evaluated using  $\alpha = 0.10$  (Gottfredson et al., 2019). We will also test for effects on factor loadings; these effects will be evaluated using  $\alpha = 0.05$ . Finally, a complete model with all significant effects will be evaluated. In case of non-convergence, variance effects will be culled first. Using the parameter estimates from step 4, we will then generate factor scores for each of the participants; these scores represent the scaled scores that can be compared across projects.

### ***Statistical outliers***

The data sets have been cleaned thoroughly through Project Kids and it is highly unlikely that they contain invalid values on any of the variables of interest. Additionally, because QR does not make assumptions about the distribution of data and estimation is not influenced by outliers (Waldmann, 2018), all data will be included in the analyses as is.

### **Analytic Strategy**

Previous research has suggested there are individual differences in response to reading intervention depending on the students' posttest outcomes (e.g., Solari et al., 2018; Wanzek et al., 2016). We will use a conditional linear quantile mixed model regression approach (lqmm; Geraci & Bottai, 2014) to test if this hypothesis is supported by our data. LQMM estimates both fixed and random effects at each specified quantile along the posttest distribution, in contrast with traditional linear mixed models that estimate parameters conditional on the mean of the posttest distribution (Koenker & Bassett Jr, 1978). LQMM are an appropriate approach for

studies in which a different relation between the dependent and independent variable is expected at different points along the outcome distribution. Using traditional linear mixed models, datasets would need to be split according to the quantiles of interest and separate models run for each subset of data. This is problematic because samples will have a limited range through truncation and likely violate assumptions of distributions that are essential for linear mixed models. Additionally, subsetting the original data set leads to a loss of power, because each subset contains a fraction of the original sample size (Petscher & Logan, 2014). These problems can be avoided with *lqmm*, because it uses all available data to estimate parameters at each quantile and does not make assumptions about the shape of the distribution.

The choice for conditional instead of unconditional quantile regression is intentional. Unconditional quantile regression estimates parameters that are not conditioned on values of other variables in the models and therefore yield results that are generalizable across all quantiles (Firpo et al., 2009; Killewald & Bearak, 2014; Koenker, 2017). However, to be able to prevent the differences in research design of the original studies to mask any outcomes, we will estimate conditional quantile regression. More specifically, our conditional quantile regression model will provide estimates averaged across the projects and treatment status (i.e., within project effects), with estimates of the variability between projects and treatments status (i.e., between project effects). The results from our conditional quantile regression models can be interpreted as the effect of the ISI intervention (treatment) within each project. The alternative unconditional quantile regression would be interpreted as the effect of treatment across all quantiles. In our case, this interpretation is less desirable for several reasons. First, the RCTs were conducted in different classrooms. Even though we are using unbiased scaled factor scores based on MNLFA, the same factor score in first grade has a different skill connotation than in third grade. Second,

the original RCTs included different counterfactuals (e.g., business as usual control and alternate treatment control), and averaging across these control conditions would mask differences in effects of the treatment across conditions. Using conditional quantile regressions will place students along the outcome variable distribution relative to the students in their project (i.e., in the same grade compared to the same counterfactual).

### ***Model specification***

All models will be estimated using the *lqmm* package (Geraci & Bottai, 2014) in the R environment version 4.0.2 (R Core Team, 2020). For RQ1, we will regress posttest word reading and vocabulary on treatment in separate models, with the addition of both pre intervention scores of word reading and vocabulary as predictors and accounting for classroom clustering, allowing for random intercepts. We will adjust the standard errors using the procedures explained by Hedges (2005) and adopted in the WWC Procedures and Standards Handbook (v 3.0 p. 25) to account for clustering at the school level (the *lqmm* package currently only allows for one level of clustering). We will estimate parameters at each .10 quantile, resulting in estimates at 9 quantiles. Given our large sample size and the limited number of parameters to estimate, this number of quantiles is considered appropriate to present a highly specific set of the slope parameters (Petscher & Logan, 2014). We will add project as a fixed effect covariate to account for differences in grade level and counterfactual conditions. For RQ2, we will use the same general strategy, with the addition the interaction of word reading skills and vocabulary with treatment. The equation is represented as follows:

$$Y_{ijt} = \gamma_{00t} + \gamma_{10t}[Treatment] + \gamma_{20t}[Project] + \gamma_{30t}[WordReading_{ij}] + \gamma_{40t}[Vocab_{ij}] + \gamma_{50t}[Treat*Vocab_{ij}] + \gamma_{60t}[Treat*WordReading_{ij}] + \varepsilon_{ijt} + r_{0jt}$$

Where  $Y_{ijt}$  is the posttest score for student  $i$  in classroom  $j$ ;  $\gamma_{00t}$  is the conditional mean posttest



score for the control group;  $\gamma_{10_t[Treatment]}$  the effect of the intervention;  $\gamma_{20_t[Project]}$  the fixed effect of each project;  $y_{30_t}$  and  $y_{40_t}$  the effects of preintervention skill level;  $y_{50_t}$  and  $y_{60_t}$  the interaction of treatment with preintervention skill level.  $\varepsilon_{ij_t}$  and  $r_{0j_t}$  are the individual and classroom level residuals. Each of the parameters are estimated at the  $t^{th}$  percentile.

Example code for RQ2 is:

```
lqmm(fixed = Vocabulary (posttest) ~ treatment + project + WR(pre) + Vocab(pre)
+ treatment*WR(pre) + treatment*Vocab(pre), random = ~1, group = teacher, tau
= c(1:9/10).
```

In case of non-convergence of our model, we will first apply optimization methods to the estimation of the model. First, we will increase the maximum number of iterations, and then, if needed, decrease the tolerance level, and finally change optimization to the Nelder-Mead derivative free optimization. If models still fail to converge, we will estimate models without interaction first, and use those estimates as starting values for the full models.

### ***Effect size estimation***

We will use the parameter coefficients and their 95% confidence intervals to make inferences about the effect at each quantile. All  $p$ -values and confidence intervals will be based on two tailed tests. We will use graphs to visually inspect data as well as find the quantiles where an effect was present (i.e., the confidence intervals do not bound 0). We will then calculate Hedges'  $g$  to specify the size of the effect for those quantiles only following the procedures laid out in Wanzek et al. (2016). At each of the statistically significant quantiles, we will calculate

Hedges'  $g = \sqrt{\frac{F(n_1+n_2)(1-r^2)}{n_1n_2}}$  with  $r$  the correlation between pretest and posttest at that quantile,

$n$  the sample size of intervention (1) and control (2) groups, and  $F$  the squared  $t$ -test of coefficients at that quantile. Since quantile regression involves running multiple analyses on the

same sample within each model (the number depending on the width of the quantiles), we will use Benjamini and Hochbergs' Linear Step-up method (Benjamini & Hochberg, 1995) to control for the false discovery rate within each model.

### ***Power***

For RQ1, the main effect of treatment, we used the PowerUpr (Dong et al., 2016) tool to calculate the minimum detectable effect size (MDES) given a Type 1 error rate of 0.05, and power at 0.8. We assumed a 2-tailed test with 55% of units randomly assigned to the treatment at the student level. The MDES for a full sample mixed effect model with 3,069 units (i.e., the minimum sample size expected) in 191 clusters for the main treatment effect is 0.09 (95% CI [0.03, 0.16]). For RQ2, the interaction between treatment and pre-treatment, we calculated the MDES for a 2-level CRT with a continuous, level 1 moderator. Assuming Type 1 error rate of 0.05, power at 0.8, and a two-tailed test, the MDES for the treatment by previous performance interaction is 0.12 (95% CI [0.04, 0.21]). Other assumptions to calculate the MDES-difference were, an ICC of 0.1, proportion of variance in outcome explained by predictors 30%,  $\omega^2$  of 0.5 and 55% of units randomly assigned to treatment. Both power analyses show this study has ample power to detect small effects, both for main effects and interactions.

### **Open Science Practices**

The dataset for this study is publicly available on LDbase (Hart et al., 2020), as a delimited file (Hart et al., 2021). At the completion of the project, we will post the R scripts and outputs to a project page on LDbase. Of the available data, we will only use item level data from the LWID and PV subtests of the WJ-III used across the six projects, and student demographic data, along with classroom nesting information and treatment status.

### References

- Al Otaiba, S., Connor, C. M., Folsom, J. S., Wanzek, J., Greulich, L., Schatschneider, C., & Wagner, R. K. (2014). To wait in Tier 1 or intervene immediately: A randomized experiment examining first-grade response to intervention in reading. *Exceptional Children, 81*(1), 11–27. <https://doi.org/10.1177/0014402914532234>
- Al Otaiba, S., Folsom, J. S., Wanzek, J., Greulich, L., Waesche, J., Schatschneider, C., & Connor, C. M. (2016). Professional development to differentiate kindergarten Tier 1 instruction: Can already effective teachers improve student outcomes by differentiating Tier 1 instruction? *Reading & Writing Quarterly, 32*(5), 454–476. <https://doi.org/10.1080/10573569.2015.1021060>
- Al Otaiba, S., & Fuchs, D. (2002). Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial and Special Education, 23*(5), 300–316. <https://doi.org/10.1177/07419325020230050501>
- Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., Meadows, J., & Li, Z. (2011). Assessment data–informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *The Elementary School Journal, 111*(4), 535–560. <https://doi.org/10.1086/659031>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300.
- Buuren, S. van, & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Catts, H. W. (2018). The Simple View of Reading: Advancements and False Impressions.

*Remedial and Special Education, 39*(5), 317–323.

<https://doi.org/10.1177/0741932518767563>

Catts, H. W., Adlof, S. M., & Weismer, S. E. (2006). *Language deficits in poor comprehenders:*

*A case for the simple view of reading.*

Colenbrander, D., Kohnen, S., Smith-Lock, K., & Nickels, L. (2016). Individual differences in

the vocabulary skills of children with poor reading comprehension. *Learning and*

*Individual Differences, 50*, 210–220.

Connor, C. M., Lara J., J., Crowe, E. C., & Meadows, J. G. (2009). Instruction, student

engagement, and reading skill growth in Reading First classrooms. *The Elementary*

*School Journal, 109*(3), 221–250. <https://doi.org/10.1086/592305>

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C.

(2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade.

*Psychological Science, 24*(8), 1408–1419. <https://doi.org/10.1177/0956797612472204>

Connor, C. M., Morrison, F. J., Fishman, B., Giuliani, S., Luck, M., Underwood, P. S.,

Bayraktar, A., Crowe, E. C., & Schatschneider, C. (2011). Testing the impact of child

characteristics × instruction interactions on third graders' reading comprehension by differentiating literacy instruction. *Reading Research Quarterly, 46*(3), 189–221.

<https://doi.org/10.1598/RRQ.46.3.1>

Connor, C. M., Morrison, F. J., Fishman, B. J., Ponitz, C. C., Glasney, S., Underwood, P. S.,

Piasta, S. B., Crowe, E. C., & Schatschneider, C. (2009). The ISI Classroom Observation

- System: Examining the Literacy Instruction Provided to Individual Students. *Educational Researcher*, 38(2), 85–99. <https://doi.org/10.3102/0013189X09332373>
- Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C., & Underwood, P. (2007). Algorithm-guided individualized reading instruction. *Science*, 315(5811), 464–465. <https://doi.org/10.1126/science.1134513>
- 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. [https://doi.org/10.1207/s1532799xssr0804\\_1](https://doi.org/10.1207/s1532799xssr0804_1)
- Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction: Examining child x instruction interactions. *Journal of Educational Psychology*, 96(4), 682–698. <https://doi.org/10.1037/0022-0663.96.4.682>
- Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E., Crowe, E. C., & Fishman, B. (2011). Effective classroom instruction: Implications of child characteristics by reading instruction interactions on first graders' word reading achievement. *Journal of Research on Educational Effectiveness*, 4(3), 173–207. <https://doi.org/10.1080/19345747.2010.510179>
- Connor, C. M., Piasta, S. B., Fishman, B., Glasney, S., Schatschneider, C., Crowe, E., Underwood, P., & Morrison, F. J. (2009). Individualizing Student Instruction Precisely: Effects of Child × Instruction Interactions on First Graders' Literacy Development. *Child Development*, 80(1), 77–100. <https://doi.org/10.1111/j.1467-8624.2008.01247.x>
- Coyne, M. D., McCoach, D. B., Ware, S., Austin, C. R., Loftus-Rattan, S. M., & Baker, D. L. (2019). Racing against the vocabulary gap: Matthew effects in early vocabulary

- instruction and intervention. *Exceptional Children*, 85(2), 163–179.  
<https://doi.org/10.1177/0014402918789162>
- Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods*, 14(2), 81–100.  
<https://doi.org/10.1037/a0015914>
- Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., Sher, K., & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in Integrative Data Analysis. *Multivariate Behavioral Research*, 49(3), 214–231. <https://doi.org/10.1080/00273171.2014.889594>
- Daucourt, M. C., Schatschneider, C., Connor, C. M., Al Otaiba, S., & Hart, S. A. (2018). Inhibition, updating working memory, and shifting predict reading disability symptoms in a hybrid model: Project KIDS. *Frontiers in Psychology*, 9.  
<https://doi.org/10.3389/fpsyg.2018.00238>
- Dong, N., Kelcey, B., Spybrook, J., & Maynard, R. A. (2016). *PowerUpR* (1.0.4).  
[powerupr.shinyapps.io/index/](http://powerupr.shinyapps.io/index/)
- Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel’s meta-analysis. *Review of Educational Research*, 71(3), 393–447.
- Firpo, S., Fortin, N. M., & Lemieux, T. (2009). Unconditional Quantile Regressions. *Econometrica*, 77(3), 953–973. <https://doi.org/10.3982/ECTA6822>
- Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology*, 107(3), 884.

- Foorman, B. R., & Moats, L. C. (2004). Conditions for sustaining research-based practices in early reading instruction. *Remedial and Special Education, 25*(1), 51–60.
- Foorman, B. R., & Schatschneider, C. (2003). Measurement of teaching practices during reading/language arts instruction and its relationship to student achievement. *Reading in the Classroom: Systems for the Observation of Teaching and Learning*, 1–30.
- Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children, 85*(2), 126–128. <https://doi.org/10.1177/0014402918811924>
- Fuchs, D., Kearns, D. M., Fuchs, L. S., Elleman, A. M., Gilbert, J. K., Patton, S., Peng, P., & Compton, D. L. (2019). Using moderator analysis to identify the first-grade children who benefit more and less from a reading comprehension program: A step toward Aptitude-by-Treatment Interaction. *Exceptional Children, 85*(2), 229–247. <https://doi.org/10.1177/0014402918802801>
- García, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research, 84*(1), 74–111.
- Geraci, M., & Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing, 24*(3), 461–479. <https://doi.org/10.1007/s11222-013-9381-9>
- Gottfredson, N. C., Cole, V. T., Giordano, M. L., Bauer, D. J., Hussong, A. M., & Ennett, S. T. (2019). Simplifying the implementation of modern scale scoring methods with an automated R package: Automated moderated nonlinear factor analysis (aMNLFA). *Addictive Behaviors, 94*, 65–73. <https://doi.org/10.1016/j.addbeh.2018.10.031>

- Gough, P. B., & Tunmer, W. E. (1986). Decoding, Reading, and Reading Disability. *Remedial and Special Education, 7*(1), 6–10. <https://doi.org/10.1177/074193258600700104>
- Hagan-Burke, S., Coyne, M. D., Kwok, O., Simmons, D. C., Kim, M., Simmons, L. E., Skidmore, S. T., Hernandez, C. L., & McSparran Ruby, M. (2013). The effects and interactions of student, teacher, and setting variables on reading outcomes for kindergarteners receiving supplemental reading intervention. *Journal of Learning Disabilities, 46*(3), 260–277. <https://doi.org/10.1177/0022219411420571>
- Hart, S. A., Al Otaiba, S., Connor, C. M., & Norris, C. U. (2021). *Project KIDS Item level Data* [Data set]. <https://doi.org/10.33009/ldbbase.1620837890.bcf8>
- Hart, S. A., Schatschneider, C., Reynolds, T. R., Calvo, F. E., Brown, B. J., Arsenault, B., Hall, M. R. K., van Dijk, W., Edwards, A., Shero, J. A., Smart, R., & Phillips, J. S. (2020). *LDbase*. <http://doi.org/10.33009/ldbbase>
- Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127–160.
- Hornburg, C. B., Rieber, M. L., & McNeil, N. M. (2017). An integrative data analysis of gender differences in children’s understanding of mathematical equivalence. *Journal of Experimental Child Psychology, 163*, 140–150. <https://doi.org/10.1016/j.jecp.2017.06.002>
- Jansen, M., Lüdtke, O., & Robitzsch, A. (2020). Disentangling different sources of stability and change in students’ academic self-concepts: An integrative data analysis using the STARTS model. *Journal of Educational Psychology, advance online*.



- Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology, 101*(4), 765.
- Killewald, A., & Bearak, J. (2014). Is the Motherhood Penalty Larger for Low-Wage Women? A Comment on Quantile Regression. *American Sociological Review, 79*(2), 350–357.  
<https://doi.org/10.1177/0003122414524574>
- Koenker, R. (2017). *Quantile regression 40 years on*. The IFS.  
<https://doi.org/10.1920/wp.cem.2017.3617>
- Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society, 46*, 33–50.
- Lam, E. A., & McMaster, K. L. (2014). Predictors of responsiveness to early literacy intervention: A 10-year update. *Learning Disability Quarterly, 37*(3), 134–147.  
<https://doi.org/10.1177/0731948714529772>
- Leijten, P., Raaijmakers, M., Wijngaards, L., Matthys, W., Menting, A., Hemink-van Putten, M., & Orobio de Castro, B. (2018). Understanding who benefits from parenting interventions for children's conduct problems: An Integrative Data Analysis. *Prevention Science, 19*(4), 579–588. <https://doi.org/10.1007/s11121-018-0864-y>
- Lonigan, C. J., Burgess, S. R., & Schatschneider, C. (2018). Examining the simple view of reading with elementary school children: Still simple after all these years. *Remedial and Special Education, 39*(5), 260–273.
- Lovett, M. W., Frijters, J. C., Wolf, M., Steinbach, K. A., Sevcik, R. A., & Morris, R. D. (2017). Early intervention for children at risk for reading disabilities: The impact of grade at

- intervention and individual differences on intervention outcomes. *Journal of Educational Psychology*, *109*(7), 889. <https://doi.org/10.1037/edu0000181>
- MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). *Gates-MacGinitie reading tests* (4th ed.). Riverside Publishing.
- McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). *Technical Manual. Woodcock-Johnson III normative update*. Riverside Publishing.
- Nation, K. (2019). Children's reading difficulties, language, and reflections on the simple view of reading. *Australian Journal of Learning Difficulties*, *24*(1), 47–73.  
<https://doi.org/10.1080/19404158.2019.1609272>
- National Center for Education Statistics. (2019). *The condition of education 2018*. (NCES 2018-144). Government Printing Office.
- National Institute of Child Health and Human Development (NICHD). (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH 00-4754). Washington DC: U. S. Government Printing Office.
- Oakhill, J. V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, *16*(2), 91–121.
- Ouellette, G., & Beers, A. (2010). A not-so-simple view of reading: How oral vocabulary and visual-word recognition complicate the story. *Reading and Writing*, *23*(2), 189–208.
- Pellegrini, A. (2001). Some theoretical and methodological considerations in studying literacy in social context. *Handbook of Early Literacy Research*, *1*, 54–65.

- Petscher, Y., & Logan, J. A. R. (2014). Quantile Regression in the Study of Developmental Sciences. *Child Development, 85*(3), 861. <https://doi.org/10.1111/cdev.12190>
- Porter, S. R. (2015). Quantile Regression: Analyzing Changes in Distributions Instead of Means. In M. B. Paulsen (Ed.), *Higher Education: Handbook of Theory and Research* (Vol. 30, pp. 335–381). Springer International Publishing. [https://doi.org/10.1007/978-3-319-12835-1\\_8](https://doi.org/10.1007/978-3-319-12835-1_8)
- R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rhoads, C. H. (2012). Problems with Tests of the Missingness Mechanism in Quantitative Policy Studies. *Statistics, Politics, and Policy, 3*(1). <https://doi.org/10.1515/2151-7509.1012>
- Roberts, G. J., Vaughn, S., Roberts, G., & Miciak, J. (2019). Problem behaviors and response to reading intervention for upper elementary students with reading difficulties. *Remedial and Special Education, 0741932519865263*. <https://doi.org/10.1177/0741932519865263>
- Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.
- Sabatini, J. (2015). Understanding the basic reading skills of US adults: Reading components in the PIAAC literacy survey. *ETS Center for Research on Human Capital and Education*.
- Simmons, D. C., Taylor, A. B., Oslund, E. L., Simmons, L. E., Coyne, M. D., Little, M. E., Rawlinson, D. M., Hagan-Burke, S., Kwok, O., & Kim, M. (2014). Predictors of at-risk kindergarteners' later reading difficulty: Examining learner-by-intervention interactions. *Reading and Writing, 27*(3), 451–479. <https://doi.org/10.1007/s11145-013-9452-5>
- Snow, C. E. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Rand Corporation.

- Snow, C. E., Porche, M. V., Tabors, P. O., & Harris, S. R. (2007). *Is literacy enough? Pathways to academic success for adolescents*. Brookes.
- Solari, E. J., Denton, C. A., Petscher, Y., & Haring, C. (2018). Examining the effects and feasibility of a teacher-implemented Tier 1 and Tier 2 intervention in word reading, fluency, and comprehension. *Journal of Research on Educational Effectiveness, 11*(2), 163–191. <https://doi.org/10.1080/19345747.2017.1375582>
- Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of the individual differences in the acquisition of reading. *Reading Research Quarterly, 21*, 360–407.
- Vadasy, P. F., Sanders, E. A., & Abbott, R. D. (2008). Effects of supplemental early reading intervention at 2-year follow up: Reading skill growth patterns and predictors. *Scientific Studies of Reading, 12*(1), 51–89. <https://doi.org/10.1080/10888430701746906>
- van Dijk, W., Norris, C. U., Otaiba, S. A., Schatschneider, C., & Hart, S. A. (2022). Exploring individual differences in response to reading intervention: Data from Project KIDS (Kids and Individual Differences in Schools). *Journal of Open Psychology Data, 10*, 2. <https://doi.org/10.5334/jopd.58>
- Waldmann, E. (2018). Quantile regression: A short story on how and why. *Statistical Modelling, 18*(3–4), 203–218. <https://doi.org/10.1177/1471082X18759142>
- Wanzek, J., Petscher, Y., Al Otaiba, S., Kent, S. C., Schatschneider, C., Haynes, M., Rivas, B. K., & Jones, F. G. (2016). Examining the average and local effects of a standardized treatment for fourth graders with reading difficulties. *Journal of Research on Educational Effectiveness, 9*(sup1), 45–66. <https://doi.org/10.1080/19345747.2015.1116032>

Wenz, S. E. (2019). What Quantile Regression Does and Doesn't Do: A Commentary on

Petscher and Logan (2014). *Child Development*, 90(4), 1442–1452.

<https://doi.org/10.1111/cdev.13141>

Wolff, U. (2016). Effects of a randomized reading intervention study aimed at 9-year-olds: A 5-

year follow-up. *Dyslexia*, 22(2), 85–100. <https://doi.org/10.1002/dys.1529>

Woodcock, R. W., McGrew, K. S., Schrank, F. A., & Mather, N. (2007). *Woodcock-Johnson III*

*Normative Update*. Riverside Publishing.

Table 1

*Participant Characteristics*

Variable	Total #	Treatment	Control	Group differences
Total	3,197	1,760	1,436	
Project 1	641	362	279	
Project 2	514	261	253	
Project 3	331	331	0	
Project 4	804	410	394	
Project 5	395	245	150	
Project 6	512	279	232	
Gender				$\chi^2 = 4.19, df = 1, p = 0.04^*$
Male	1,491	845	645	
Female	1,616	976	640	
Missing	90	67	23	
Ethnicity				$\chi^2 = 0.30, df = 1, p = 0.58$
Hispanic	142	87	55	
Non-Hispanic	2,763	1,618	1,144	
Missing	292	173	119	
Race				$\chi^2 = 18.20, df = 7, p = 0.01^*$
American Indian/Alaskan	7	3	4	
Asian	69	44	25	
Black	1,300	804	496	
Hawaiian/Pacific islander	25	11	14	

Variable	Total #	Treatment	Control	Group differences
White	1,369	772	596	
Multi-racial	91	57	34	
Other	51	18	26	
Missing	285	173	112	
LEP status				$\chi^2 = 0.58, df = 1, p = 0.45$
Not LEP	2,571	1,534	1,037	
LEP	50	33	17	
Missing	576	321	254	
Eligibility for FRL				$\chi^2 = 0.84, df = 3, p = 0.83$
No	1,165	688	477	
Yes	1,199	698	501	
Missing	833	502	330	
ESE services				$\chi^2 < 0.001, df = 1, p = 1.00$
No	249	149	100	
Yes	12	7	5	
Missing	2,935	1,732	1,203	

Table 2

*Missing Data for Key Variables*

Variable	# missing (%)
Treatment indicator	1 (< 1)
Word reading pre	120 (4)
Word reading post	507 (16)
Vocabulary pre	105 (3)
Vocabulary post	469 (15)



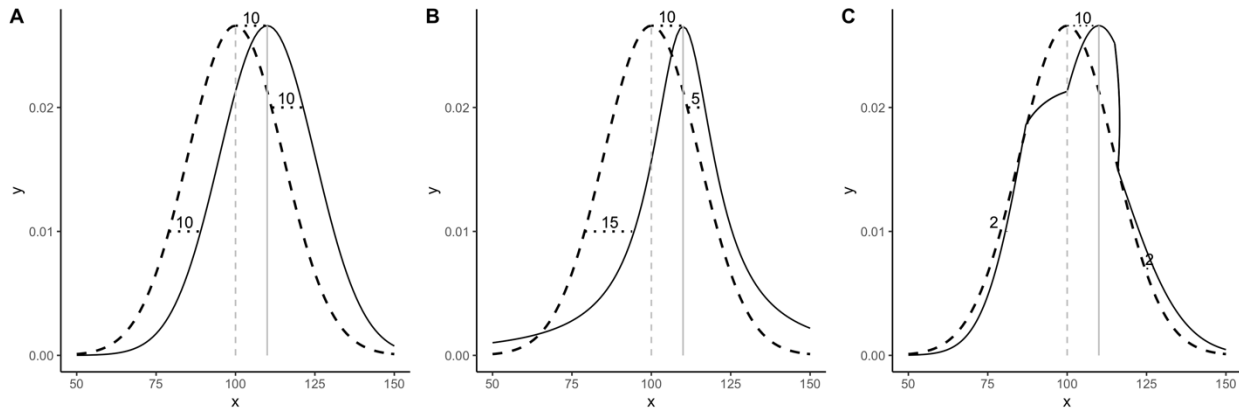
Table 3

*Minimum Expected Sample Sizes*

RQ	%missing outcome	%missing predictor	Minimum expected N from 3,197 total
1a/b	15/16	-	2,717/2,685
2ai/bi	15/16	4	2,609/2,578
2aii/bii	15/16	3	2,636/2,605

Figure 1

Possible scenarios of intervention effects.



*Note.* Three possible scenarios of intervention effects. In each scenario the x-axis represents scores on a hypothetical reading assessment. The y-axis represents the density of the distribution. The dashed line represents the distribution of skills of students in a control group; the solid line represents the distribution of scores of students in an intervention group. In each scenario, the mean of both distributions are represented by grey dashed (for control) and solid (for intervention) lines. A. In this scenario the difference in scores between control and intervention groups are equal across the distribution, namely students in the intervention groups scored 10 points higher than those in the control group. B. In this scenario, the intervention leads to more gains for students with lower scores (i.e., 15 points) than students with higher scores (i.e., 5 points). The difference at the mean is still 10 points. C. In this scenario the intervention only leads to increased scores for students at the average range (i.e., 10 points); whereas the difference for students either with lower or higher scores is almost 0.

## Appendix

### Specifics on Multiple Imputation Techniques

Using multiple imputation will enable us to avoid the casewise deletion imposed by the *lqmm* package (Geraci & Bottai, 2014) that we will use to estimate the models and use the complete data set. As stated above, students with missing variables on both the pre- and postintervention scores will be excluded from the sample before the MI procedures, since they will not have the scores required to impute scores. MI will take all available data into account (i.e., scores on other pre-and postintervention variables not of interest in the main analyses), as well as the nested structure of the data). We will generate 25 imputations based on 25 iterations using the *mice* package (Buuren & Groothuis-Oudshoorn, 2011) to ensure random convergence of the MCMC. To evaluate the imputed scores, we will visually inspect stripplots for overlapping distributions of the original and imputed data sets. We will then replace the missing values in the original data set with the mean value of the 25 imputed data sets. We will provide pre-imputation descriptives of the variables of interest and compare them to post-imputation descriptives to show comparability of the two datasets.