Learning Disabilities Research & Practice, *33*(1), 37–50

DOI: 10.1111/ldrp.12161

Examining the Impact and School-Level Predictors of Impact Variability of an 8th Grade Reading Intervention on At-Risk Students' Reading Achievement

Hank Fien, Daniel Anderson, Nancy J. Nelson, Patrick Kennedy, and Scott K. Baker
University of Oregon
Michael Stoolmiller
Michigan State University

The purpose of the present article is to report on a large-scale investigation of six school districts' implementation of an initiative aimed at reducing dropout rates by improving reading achievement in the middle grades. Data for the Middle School Intervention Project (MSIP) were collected in 25 middle schools across the state of Oregon. We examined (a) the degree to which the schools improved reading achievement for struggling readers in 8th grade, and (b) whether we could account for school differences in the treatment effect through measured explicit and intensive intervention factors. At the end of 8th grade there was no evidence of significant or positive effects on the two primary reading outcome measures.

INTRODUCTION

In the last decade, there has been increased attention and funding for local and state education agencies to partner with institutes of higher education to conduct rigorous evaluations of locally developed programs, practices, and policy. At the heart of this movement is demonstrating what works and what does not work in public schools in the United States. For example, the Institute of Education Sciences, the statistical and research arm of the U.S. Department of Education, released for the first time in 2009 a special funding topic area devoted to the *Evaluation of State and Local Education Programs and Policies*. Since that time, IES has funded 22 partnerships between research institutions and state and local education agencies to rigorously evaluate whether programs and policies implemented by states and districts are having a beneficial impact on student education outcomes (https://ies.ed.gov/funding/ncer_rfas). Likewise, the Office of Innovation and Improvement, through the Investing in Innovation (i3) Fund has allocated over $1.5 billion for funding large-scale efficacy and effectiveness trials to determine what works in education models and approaches across the United States (https://www2.ed.gov/programs/innovation/funding.html).

Through these funding mechanisms and others, there is a call in educational research not only determine what programs, practices, and policies work or do not work in U.S. schools, but also to examine factors that make programs and practices more or less effective, or to identify why programs simply do not work. This context is particularly relevant when considering theories of change that examine the role of alterable variables (i.e., malleable factors that can be easily adjusted by school staff) for enhancing or muting intervention effects. In theory, these investigations are meant to help educators to increase or improve implementation of practices that amplify intervention effects, and to discontinue practices that do not deliver on expected outcomes.

The purpose of the current study is to report on a large-scale investigation of six school districts' implementation of a concerted effort to reduce dropout rates by improving reading achievement in the middle grades. Data for the Middle School Intervention Project (MSIP) were collected in 25 middle schools across the state of Oregon in the Pacific Northwest. Specifically, we examined (1) the extent to which the schools improved reading achievement for struggling readers in 8th grade, (2) whether schools varied in how much they improved outcomes for students in the reading intervention treatment groups, and (3) whether we could explain school to school differences in the treatment effect based on factors hypothesized to support stronger reading growth for struggling readers, namely through explicit and intensive intervention factors (e.g., number of explicit teacher-student interactions, intervention duration, frequency, dosage).

Context for MSIP Evaluation

Like many states, Oregon is in the midst of reform to substantially increase high school diploma requirements (Oregon Department of Education: ODE, 2008). Widespread demand for graduates who are better prepared for challenging postsecondary opportunities—particularly advanced education—is fueling graduation reform (Friedman, 2005; Carnevale & Desrochers, 2003). Currently, 90 percent of the fastest-growing, high-wage jobs in the country require postsecondary education (Alliance for Excellent Education, 2008), and among high school graduates, only about

Requests for reprints should be sent to Hank Fien, University of Oregon. Electronic inquiries should be sent to ffien@uoregon.edu.

LEARNING DISABILITIES RESEARCH & PRACTICE

50 percent are adequately prepared for college (ACT, 2006; Greene & Winters, 2005). The call for more prepared high school graduates is striking in light of the recent ACT report that indicated that only 23 percent of high school graduates who took the ACT in 2009 met the college readiness benchmark in reading (ACT Inc., 2009). Of students who actually enroll in college, 42 percent of community college freshmen and 20 percent of freshmen in four-year institutions must take at least one remedial course (e.g., basic math, remedial writing; Alliance for Excellent Education, 2009). In response, State education agencies (SEAs) are instituting reforms to better prepare high school graduates. Many SEAs are increasing the number of credits needed for graduation, as well as the number of academically challenging courses students must take to graduate (Guy, Shin, & Thurlow, 1999; Johnson & Emanuel, 2000).

Currently, approximately 30 percent of students drop out of school and never graduate (Editorial Projects in Education (EPE), 2008; Laird, KewalRamani, & Chapman, 2008). In Oregon, an estimated 75 students drop out of school each day (Diplomas Count, 2008). Dropouts are at substantially higher risk than graduates for life-long difficulties associated with unemployment, poverty, illiteracy, incarceration, and chronic stress (Finn & Owings, 2006; Harlow, 2003; McCaul, 1989). Nationally, dropout rates are substantially elevated for specific student groups, including students from high-poverty and minority backgrounds (Greene & Winters, 2005; Neild & Balfanz, 2006; Rumberger, 1995). In Oregon, 72 percent of White students graduate from high school, while the graduation rate is only 50 percent for Hispanic students, and 61 percent for African American students (Greene & Winters, 2005). The costs associated with dropping out of school are measurable in terms other than diminished post-secondary education opportunities as well. Nationally, 1.23 million dropouts from the graduating class of 2007 will cost the nation nearly $329 billion in lost income, taxes, and productivity over their lifetimes (Alliance for Excellent Education, 2007; EPE, 2008). In Oregon alone, the 13,500 dropouts from the graduating class of 2008 are expected to cost the state $3.5 billion in lost wages, taxes, and productivity over their lifetimes (Alliance for Excellent Education, 2009).

To address these issues locally, ODE has been working with districts across Oregon to implement a multi-tiered delivery system that integrates academic and behavioral supports for students in K-12 settings. The *Effective Behavioral and Instructional Support System* initiative (EBISS) is multi-tiered in that *universal,* Tier 1 or school-wide academic and behavior supports are implemented to benefit all students in a school, and *targeted,* Tier 2 academic and behavior interventions are delivered to students at-risk for school failure (e.g., three-tiered model; Baker & Baker, 2007; Baker et al., 2007; Kame'enui et al., 2002; Nelson et al., 2016; Smith, Fien, Basaraba, & Travers, 2009; Sugai & Horner, 1999). EBISS supports districts through ongoing technical assistance and financial support (i.e., providing district-level funds for implementation coordinators). To date, EBISS has been implemented in approximately 60 school districts and 200 schools across Oregon.

However, districts and schools have varied considerably in their implementation of the EBISS initiative. For example, some districts have focused their efforts in elementary schools within their district, while other districts have elected to focus on middle or high school. The state has devoted millions of dollars to EBISS, yet, prior to the MSIP initiative, there had not been a systematic evaluation of the impact of this widely implemented initiative on student outcomes. In the MSIP evaluation, we partnered with six districts that had made a serious commitment to implementing multi-tiered supports in both academic and behavior areas in the middle grades (6–8) to study the effects of their intervention systems on student reading achievement (Baker, Crone, & Fien, 2010).

Although districts were completely responsible for developing what their middle school interventions would look like – the specific programs, strategies, and practices they would implement – the research team was charged with developing a research design and measurement net to determine, in a causal paradigm, whether district-adopted and -implemented interventions were having their desired effect on student reading outcomes. The goals of the evaluation were to measure impact on student reading achievement and study implementation factors hypothesized to improve reading outcomes (i.e., delivery of evidence-based intervention features) for students experiencing reading difficulty who were at risk for eventual high school dropout.

Adolescent Literacy Intervention Literature Review

Numerous recent publications have focused on adolescent literacy, and on the need to provide explicit and intense interventions for struggling adolescent readers (Biancarosa & Snow, 2006; Kamil et al., 2008; National Association of State Directors of Special Education, 2006; Torgesen et al., 2007). Some of these documents provide clear recommendations for districts, as well as middle and high schools. The recent *IES Practice Guide, Improving Adolescent Literacy: Effective Classroom and Intervention Practices* (Kamil et al., 2008) was produced to "present specific and coherent evidence-based recommendations that educators can use to improve literacy levels among adolescents in upper elementary school, middle school, and high school" (p. 1). The recommendations in the Practice Guide closely mirror recommendations outlined in *Academic Literacy Instruction for Adolescents,* produced by the Center on Instruction (Torgesen et al., 2007), and *Reading Next* produced by the Carnegie Corporation (Biancarosa & Snow, 2006). All three of these prominent documents on adolescent literacy address what content area teachers (e.g., English Language Arts, Social Studies, Science) should do during content area instruction to improve students' comprehension of content area texts, and, most relevant for this project, what needs to be in place to provide effective reading interventions for struggling adolescent readers.

Although research specific to adolescent literacy is not as extensive as research on beginning reading (Boulay, Goodson, Frye, Blocklin, & Price, 2015; Herrera, Truckenmiller, & Foorman, 2016; National Institute of Child Health and Human Development, 2000; Snow, Burns, & Griffin, 1998),

there is a strong and growing consensus that if what we currently know about literacy instruction for adolescents were more broadly applied in practice, there is "little doubt that levels of adolescent literacy would improve" (p. 1, Torgesen et al., 2007). A recent quantitative synthesis of reading programs for adolescents found 33 studies published between 1970 and 2007 involving 39,000 students (Slavin, Cheung, Groff, & Lake, 2008). The important outcome of this synthesis, and one that has direct bearing on how reading interventions were evaluated in this project, is that interventions and approaches that focused on what teachers and students did in the classroom during reading instruction (i.e., the strategies used to teach reading and engage students) had a larger impact than interventions that implemented a particular curriculum or program.

Evidence from high-quality studies (Kamil et al., 2008) also indicates there is strong support for the assertion that explicit instruction is a necessary foundation for reading interventions with struggling adolescent readers (e.g., Duffy et al., 1987; Fuchs et al., 1997; Herrera et al., 2016; Klingner, Vaughn, & Schumm, 1998; Schumaker & Deshler, 1992). Explicit and systematic instruction involves a series of sequenced instructional steps that include: (a) teachers explaining and modeling strategy use, (b) teachers guiding students in using the strategy or strategies (i.e., guided practice), and (c) students demonstrating their ability to use the strategies independently under the supervision of the teacher (Gersten, Fuchs, Williams, & Baker, 2001; Kamil et al., 2008). The power of explicit instruction cuts across multiple content areas as a method for providing effective reading instruction for adolescent readers (and younger readers), as it can be used to teach word-level reading, reading fluency, vocabulary, and comprehension (Biancarosa & Snow, 2006; Kamil et al., 2008; Scammacca et al., 2007; Torgesen, et al., 2007).

One conclusion from the recent spate of research on reading interventions for older students is how difficult it is to improve outcomes for struggling readers in the upper grades Solis et al., 2014). For example, in a recent synthesis of extensive reading interventions that lasted a minimum of 75 sessions delivered to students in Grades 4 through 12 only small, positive effects were indicated on such outcomes as reading comprehension, reading fluency, word reading, and spelling (Wanzek et al., 2013). Interestingly, hypothesized moderators of intervention effects related to intervention group size, hours of intervention, and grade level of intervention were found to be significant.

While a good number of seemingly promising interventions have been tested, many have not demonstrated significant and meaningful improvements under rigorous research conditions. In the last 20 years, more than 7,000 peer-reviewed studies of adolescent literacy interventions have been published, yet only 33 of these studies have met WWC standards with or without reservations (Herrera et al., 2016). Of these 33 rigorous studies, 12 were identified as having a positive or potentially positive effect on vocabulary, reading comprehension, or general literacy skills, and all of these involved explicit instruction or the use of instructional routines to teach reading. The vast majority of the studies demonstrating positive or potentially positive effects also involved ongoing support or coaching for instructors, who were most likely to be typically hired school staff. Importantly, eight of the 12 studies observed small to moderate effects on a high-stakes assessment, such as a state accountability measure.

It is also important to note that the Herrera et al. (2016) review summarized research findings from *all* studies of adolescent literacy interventions that met review criteria, which includes studies planned and conducted by researchers to test the effects of particular interventions under ideal conditions, as opposed to only studies that are implemented under naturalistic conditions (i.e., those in which districts select the programs and practices they will implement, even if the evaluation is conducted through support from an external evaluator, as is the case in the current study). In reviews of studies of literacy interventions where districts and schools select and implement interventions, even fewer studies demonstrate positive or potentially positive effects on student literacy outcomes.

For instance, a recent review of the Striving Readers grant program, in which 16 school districts were paired with an external evaluator to study the effects of district-selected and district-implemented literacy interventions, identified only three interventions that resulted in positive or potentially positive effects on a literacy-related outcome, although effect sizes were small, ranging from 0.0 to 0.21 (Boulay et al., 2015). Consistent with other research findings, the interventions that were identified as being effective incorporated explicit instruction to teach reading skills. In addition, dosage and intensity were higher for the interventions that were effective (e.g., 90 minutes of supplemental intervention per day) compared to those that were not.

PURPOSE OF THE STUDY

The purpose of the current study is to evaluate the impact of locally supported reading interventions for 8th grade students with reading difficulties and increased risk for dropping out of school (Chen & Kaplan, 2003; Finn & Rock, 1997). The Middle School Intervention Project was conducted in six school districts across the state of Oregon. Individual schools used a rank-ordered list of composite z-scores comprised of students' scores on the 7th grade spring state reading assessment and an oral reading fluency measure to select a normative cut point for determining who would receive reading intervention. Each school selected this cut point based primarily on the school's available resources in terms of capacity for delivering interventions in their building (e.g., staffing, schedule, space). Students scoring below the normative cut score were assigned to the school-selected and school-implemented reading intervention (in most cases, in addition to a regular, English Language Arts [ELA] class) or to the comparison condition in which students received typical reading instruction, generally in the form of a regular, ELA class. Our specific research questions in this study are as follows:

(1) What is the impact of a district-implemented adolescent literacy intervention on at-risk students' reading achievement?
(2) Is there variability in the treatment effect between schools?

(3) Does school-to-school variability in the average number of explicit teacher-student interactions across the school year, intervention intensity, duration, and alignment to intervention need co-vary with school-to-school differences in treatment outcomes (i.e., do these factors predict variability in school-level discontinuities)?

METHOD

Participants

Participants in the study were teachers and students from six school districts in Oregon. Three of the districts were medium to large suburban districts located adjacent to two of the three largest metropolitan areas in the state, and three were small to midsize city districts located in the same two metropolitan areas. Districts ranged in size from 5,659 to 39,941 students (NCES, 2015), and were recruited to participate because: (a) we had worked with them on previous research and outreach projects, including prior years of the MSIP project; (b) they had existing interventions in place to improve reading outcomes for struggling readers; and (c) they were interested in evaluating the impact of those interventions on student outcomes.

Twenty-five comprehensive middle schools participated in the study, representing 61 percent of the districts' schools that served eighth-grade students. Schools did not participate if (a) they did not offer a traditional middle school curriculum ($n = 10$), or (b) fewer than 10 percent of eighth grade students were eligible for the intervention ($n = 6$). Participating schools ranged in size from 339 to 1,049 students, and were larger and served a more academically diverse student population than non-participating schools.

All students who attended a participating school for any portion of eighth grade and had a valid score on at least one of the two reading assessments used for assignment to condition were included in the study. These assessments were administered the prior year, when students were in seventh grade. Condition assignment was based on whether students scored above or below a reading cut point. Students below the cut point were assigned to receive the school-planned intervention, and students above the cut point were assigned to the comparison condition. A total of 5,753 students were assigned to any condition. The sample was comparable to both state and national averages on a range of demographic characteristics. For example, 22 percent of students in the study were Hispanic, and 17 percent represented other minority groups, compared to 22 percent and 25 percent Hispanic, and 13 percent and 25 percent other minority, at the state and national levels, respectively. Similarly, 59 percent of the sample was eligible for free or reduced-price lunch, compared to 54 percent and 52 percent eligibility at the state and national levels, respectively.

Description of Intervention

Participating schools had intervention components in place prior to the start of the study. This underscores a central purpose of this analysis: to evaluate the impact of districts' existing intervention practices on student outcomes. As a condition of participation, schools agreed to adhere to two common implementation criteria: (1) provide reading interventions only to students who scored below the cut point, and (2) monitor all intervention students and use data teams to make ongoing decisions about the interventions.

Districts and schools made all intervention decisions independent of the research team, although they received ongoing support from project staff to understand features affecting the intensity of interventions and were encouraged to match intervention intensity to student need. Interventions varied widely among schools with respect to curricula used; frequency, duration, and length of interventions; staff qualifications; and ratio of teachers to students. To document intervention features, we conducted direct observations of reading intervention classes, English Language Arts (ELA) classes, and data team meetings.

Reading Intervention

The purpose of the reading intervention was to provide targeted support to improve reading achievement for struggling readers. To measure the dosage of reading interventions, we documented through intervention tracking sheets completed by school-based research staff (1) the number of days per week reading intervention classes were held, (2) the average duration of intervention sessions, and (3) the frequency with which published programs were used. Across the project, reading interventions met more than four days per week ($M = 4.3$, $SD = .82$), but most of the variability came from one district. In three districts, every reading intervention class met five days per week. In the remaining three districts, reading intervention classes met 4.97 ($SD = .18$), 4.83 ($SD = .48$), and 4.07 days per week ($SD = 1.42$), on average. The average reading intervention session lasted 58 minutes ($SD = 21$), but session length varied substantially by district, ranging from a low of 47 minutes ($SD = 8$), to a high of 70 minutes ($SD = 27$). More than three quarters of reading interventions used a published program (M proportion $= .76$, $SD = .43$). Two districts used a published program in every reading intervention class and three others used a published program in between 73 percent and 92 percent of intervention classes. In the sixth district, only 42 percent of reading interventions used a published program. Across the schools, the most frequently reported intervention programs were: teacher-created materials (30.1 percent of classes), Language! (20.0 percent of classes), Step Up to Writing (9.4 percent), Read 180 (8.9 percent), Corrective Reading (7.5 percent), Reading Advantage (6.1 percent), and Word Generation (4.7 percent).

Data Teams for Decision Making

The purpose of the school data teams was threefold: (1) summarize and report ongoing student progress, (2) use these data to evaluate the effectiveness of the interventions, and (3) modify interventions as necessary to improve outcomes (Crone et al., 2016). Data teams were observed at least

twice per year using a standard observation protocol that documented information about data team composition, students and data sources reviewed, and ratings of data team structures and processes. Overall, meetings were frequently well-attended, with an average group size of 7 members (*SD* = 2), at least one of whom was an administrator more than three quarters (78 percent) of the time. However, teams spent an average of fewer than 2.5 minutes discussing students and even less time was spent discussing intervention students (roughly 1.5 minutes) specifically. Data team discussions frequently focused on high-stakes state test scores or grades, rather than student performance on specific skills and content: 35 percent of student discussions included state test scores, whereas 22 percent of discussions included formative assessment data, 10 percent of discussions included attendance data, and 6 percent of discussions referenced quantitative behavioral data. Data teams made relatively few actionable decisions (i.e., deciding to maintain a student in his or her current intervention, identifying a goal, assigning a person, or establishing a timeline) in meetings: Only 40 percent of student discussions of reading issues and 34 percent of discussions of behavioral concerns resulted in any type of decision made.

MSIP Summer Institutes

In addition to the two primary intervention components, the study design included an ongoing dialogue with project schools about project goals and the purposeful dissemination of project outcomes. A primary focus of those efforts was a series of conferences known as the MSIP Summer Institutes. These institutes convened key stakeholders from each school to review project data related to the intervention components and plan implementation for the coming year and included both formal presentations by project staff and a series of breakout sessions during which the districts and schools reviewed and discussed their own data. Breakout sessions were led by a district leader, with support from project staff, and included a discussion of the relationship between school-specific findings and interventions planned for the next year. Through these discussions, school stakeholders generated an action plan for implementing MSIP in their school that fall, and shared those plans within their districts and with other participating districts. Nearly all participants (94 percent) reported anonymously that they found the breakout sessions valuable, 82 percent said they learned information at the Institute that would inform their professional practice and 86 percent said they planned to share information they learned at the Institute with colleagues.

Measures

We collected two measures of reading proficiency at the end of seventh grade to assign students to condition and again at the end of eighth grade to measure reading achievement: (1) the Oregon Assessment of Knowledge and Skills, Reading/Literature subtest (OAKS); and (2) a measure of oral reading fluency, the easyCBM Passage Reading Fluency measure (Alonzo, Tindal, Ulmer, & Glasgow, 2006). For the exploratory analysis that is part of the current study, students with low scores on the OAKS were considered in need of higher order reading instruction, while students with low scores on the PRF were considered in need of foundational reading instruction. We also collected direct classroom observation data three times across the year (fall, winter, and spring), using the MSIP Classroom Observation Tool, in intervention and comparison settings. These classroom observation data are used in the current study as evidence of the explicitness and intensity of instruction, and to demonstrate the type of instruction (i.e., foundational reading or higher-order reading) that students received.

Oregon Assessment of Knowledge and Skills Reading/Literature (OAKS)

The OAKS (Oregon Department of Education [ODE], 2012) is a criterion-referenced test aligned with grade-level content standards. It is an untimed, multiple-choice, computer adaptive test in which roughly 80 percent of the items address reading comprehension, and 20 percent address vocabulary knowledge (ODE, 2012). Student performance is reported as an Item Response Theory scaled score. Students were given up to two opportunities to take the OAKS during eighth grade (ODE, 2012); their highest score was used for analysis. The adaptive administration maximized the test information function (thereby reducing the standard error of measurement) by matching item difficulty with student ability. The reliability of the test was therefore high across a broad range of student abilities (Oregon Department of Education, 2007). During the initial scale creation, the Rasch theta estimates were transformed to a RIT scale (Rasch Unit), which was centered on 200 with a standard deviation of 10 ($RIT = (\theta \times 10) + 200$, where θ represents students' Rasch-scaled ability estimate). Items were anchored to the initial calibration values across administration years, making scores directly comparable between years within each grade.

easyCBM Passage Reading Fluency (PRF)

easyCBM PRF (Alonzo et al., 2006) is a standardized, individually administered measure of oral reading fluency. Students read aloud from a 250- to 350-word passage for 1 minute. Skipped, misread, and omitted words are counted as errors, and the score used for analysis was the number of words read correctly in 1 minute. The average correlation between a reference PRF passage and 19 other seventh-grade passages was .89 (Alonzo & Tindal, 2008). easyCBM PRF is also moderately predictive of performance on the OAKS. Based on a convenience sample from three districts, the correlation between seventh-grade fall PRF scores and seventh-grade OAKS scores was .68, accounting for 15 percent of OAKS variance (Anderson, Alonzo, & Tindal, 2010).

MSIP Classroom Observation Tool

The MSIP Classroom Observation Tool (MSIP-COT) was designed to document various attributes of literacy

instruction, developed by the project research team using other published observation instruments as models (e.g., Doabler & Nelson, 2009; Grossman et al., 2010; Pianta & Hamre, 2009; Smolkowski & Gunn, 2012). The portion of the MSIP-COT germane to the current study involves a modified version of the Classroom Observations of Student Teacher Interactions (MSIP-COSTI). The COSTI has been field-tested and validated in more than 1,000 classroom observations of reading and math instruction in elementary settings. In our previous work with the COSTI, predictive validity coefficients with reading and mathematics outcomes have ranged from .25 to .55.

The MSIP-COSTI utilizes event recording to document the quantity and types of content-relevant interactions that occurred between teachers and students (i.e., each individual instance of teacher demonstrations, teacher corrective and confirmatory feedback, individual student practice opportunities, group practice opportunities, peer-to-peer practice opportunities, and student initiations that occurred during the observation). The student-teacher interactions provide evidence for the explicitness of instruction; classrooms with more teacher modeling, student practice, and feedback are arguably indicative of more explicit instruction. Using the MSIP-COSTI, student-teacher interactions are coded within blocks of time corresponding to (a) content domains that could occur during literacy instruction (i.e., decoding, reading connected text, reading comprehension, writing, vocabulary, other literacy-relevant, other non-literacy) and (b) grouping structures employed (i.e., independent work, 1:1 instruction, small group teacher-led, small group independent, and large group teacher-led), in each observed classroom.

Trained data collectors who demonstrated inter-rater agreement (.80 or higher) with the expert lead were eligible to conduct observations in live classrooms. These trained data collectors used the MSIP-COT to observe all reading intervention classes and each section of ELA instruction three times per year – in the fall, winter, and spring – to record the frequency of student-teacher instructional interactions during foundational reading (i.e., decoding and reading connected text) and higher-order reading (i.e., vocabulary and comprehension) skill instruction.

Assignment of Students to Condition

Assignment to condition was based on a standardized cut score. Normative information was used to standardize students' OAKS and PRF scores into a z-score distribution. The z-score for OAKS was based on the statewide standard deviation; the z-score for PRF was based on the standard deviation of the study sample, as statewide normative information was not available. Students' z-scores for each measure were averaged to create a composite score for each student. The score for students who took only one of the two assessments was the z-score for that assessment. Prior to intervention, a rank-ordered list of cut scores was distributed to each school. Key staff (e.g., principals, specialists) selected the cut point that their school used for assignment to condition, usually determined by the number of students they could serve in the intervention group. Students with cut scores below this point were assigned to the intervention group, and students with scores above it were assigned to the comparison group. Schools followed project guidelines so that at minimum 20 percent of students received the intervention, helping maintain statistical power for the RD analysis.

The mean cut score across project school was -0.60 on the standardized assignment variable, with a standard deviation of 0.19. At the school level, school provided treatment to, on average, 82 percent of students who scored less than or equal to the cut, which varied with a standard deviation of 13 percent, ranging from 55 percent to 100 percent. Similarly, schools provided treatment to approximately 12 percent of students, on average across schools, who scored above the cut point, which varied with a standard deviation of 14 percent and ranged from 3 percent to 62 percent.

Strategies to Increase Study Compliance

Multiple Cut Points for Condition Assignment

The use of multiple cut points for condition assignment in RD studies is not typical in education research (e.g., Bloom, 2012; Nomi & Allensworth, 2009), but was critical to maintaining school participation. A single cut point for all schools would have resulted in some schools serving a very different proportion of students in the reading intervention than they typically served, placing disproportionate demands on district and school resources, and compromising a fundamental premise of the study, which was to investigate effects as schools and districts typically applied interventions.

To account for different cut points in analyses that pooled students across schools, we centered the cut scores within each school by subtracting from each student's cut score the individual school's cut point so that, for analysis and interpretation purposes, all schools had the same cut point (i.e., zero). Comparable to using school mean centering in a multilevel model (Raudenbush & Bryk, 2002), we included the school cut point value in all models as an additional school-level predictor.

Use of Exemptions in the Assignment Process

Schools could exempt up to 5 percent of their students from placement into the condition indicated by the student's cut score. These exemptions were necessary because some school staff had strong a priori views that certain students should or should not be placed into intervention and their professional judgment occasionally needed to override the cut score (e.g., students with Individual Education Plans who needed to be served in the intervention group even if the cut score selected did not result in that particular student being assigned to the intervention condition). These students are included in the analysis, but the effect size estimates pertain only to the students who complied with their assignment.

Analysis

Misspecification of functional form is perhaps the greatest threat to RD designs (Bloom, 2012). If a linear trend is specified when the data follow a curvilinear trajectory, then a gap at the cut point may be estimated when no gap exists, or vice versa. To guard against this threat, we fit generalized additive models (GAMs), which relax the linearity assumption by replacing one or more model parameters with smoothing functions. Each term in a GAM can be specified with a different smooth, or not smoothed at all (i.e., linear relation). For the smoothed terms, the functional form is determined, in part, by the data themselves, similar to non-parametric methods (i.e., Chib & Greenberg, 2014).

We fit all GAMs using R (R Development Core Team, 2016) with the *gamm4* package (Wood & Scheipl, 2014) and thin-plate spline smooths. The amount of smoothing was estimated via generalized cross-validation procedures (Wood, 2006), with the nonlinearity reported by the effective degrees of freedom (EDF). When the EDF = 1.00 the relation is linear (greater EDF values indicate more nonlinearity; Shadish, Zuur, & Sullivan, 2014). To simplify the model, we replaced all smooths with an EDF of 1.00 with linear parameters, which increased power and reduced the likelihood that the observed results were sample-specific (i.e., the bias-variance tradeoff; see James, Witten, Hastie, & Tibshirani, 2013).

We began our model-building process by first fitting the impact models for OAKS and ORF to evaluate the effect of treatment. We anticipated variability between schools in the treatment effect. We therefore planned post-hoc exploratory analyses to examine whether observational data aggregated to the school level could predict treatment effect variability between schools.

Impact Models

In our study, the data were inherently multilevel, with students (Level 1) nested in schools (Level 2), and each school setting its own cut-point for treatment. The basic impact model was therefore estimated with a multilevel GAM. At the student level, the basic impact model was defined as

$$\begin{aligned} y_{ij} = {} & \beta_{0j} + \beta_{1j}(LEC_{ij}) \\ & + s_1(LEC_{ij} * pre_{ij}) + s_2(AC_{ij} * pre_{ij}) + e_{ij} \end{aligned} \quad (1)$$

where y_{ij} represents the outcome (OAKS or ORF) for student i in school j. The *LEC* and *AC* variables were dummy-coded vectors indicating whether students' scores on the assignment variable, *pre*, were less than or equal to the cut (LEC), or above the cut (AC), for treatment in school j. These dummy variables were both included as interactions with the assignment variable such that separate smooths, $s_p()$, could be estimated for students on either side of the cut point for treatment. At the school level the model was defined as

$$\beta_{0j} = \gamma_{00} + \gamma_{10}(cut_j) + u_{oj} \quad (2)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (3)$$

where cut_j represents the location of the cut score for school j, γ_{00} represents the model intercept (average achievement for students above the cut), and γ_{10} represents the average gap in the regression discontinuity (i.e., treatment effect). The u_{oj} and u_{1j} terms represent school-level deviations from these averages, respectively, and were estimated with an unstructured variance-covariance matrix.

In a sharp RD design, the γ_{10} term represents the average effect of treatment. In our sample, however, approximately 9 percent of students who scored above the cut point received treatment (crossovers), while approximately 18 percent of students who scored below the cut point did not receive treatment (no-shows), making the design fuzzy. The average treatment effect is still estimable in fuzzy RD designs, however, provided there is a significant gap in the probability of receiving treatment at the cut-point (Bloom, 2012). This implies first modeling the probability gap through a model equivalent to Equations (1)–(3), but with treatment receipt as the outcome and a multilevel logistic GAM estimated. The sharp RD estimate, γ_{10}, is then divided by the estimated probability gap to provide the fuzzy RD point estimate, and similar transformations are available to compute fuzzy standard errors (see Bloom, 2012; Schochet, 2008). Interpretation of the effect size, however, is restricted to those who complied with the design.

Post-hoc Exploratory Modeling

The u_{1j} term in Equation (3) represents random deviations between schools in the estimated treatment effect, which was assumed normally distributed with variance τ_{11}. As a post-hoc exploratory analysis, we were interested in examining whether classroom-level observational data aggregated to the school level would account for any of this variability. We fit separate models for instructional variables related to *foundational* and *higher-order* reading skills, for a total of four models (two for each outcome). Each model included three school-level variables for each corresponding reading skill area: (a) average hours of instruction, (b) average number of teacher-student interactions, and (c) proportion of students whose reading need matched the instruction they received. These variables were on dramatically different scales and were thus standardized prior to analysis. The conditional models included each variable predicting school-level variability in the treatment effect, with separate smooths estimated for each side of the cut for each variable. Equation (3) was redefined as

$$\begin{aligned} \beta_{1j} = {} & \gamma_{10} + s_3(LEC_{ij} * \overline{hours}_j) + s_4(AC_{ij} * \overline{hours}_j) \\ & + s_5(LEC_{ij} * \overline{int}_j) + s_6(AC_{ij} * \overline{int}_j) \\ & + s_7(LEC_{ij} * pmatch_j) + s_8(AC_{ij} * pmatch) + u_{1j} \end{aligned} \quad (4)$$

where $\overline{hours}_j$, $\overline{int}_j$, and $pmatch_j$ represent the average hours, average number of teacher-student interactions, and proportion of students whose instruction matched their need in

school *j*. As with the impact models, the model was simplified following the full fit, with smooths with an EDF of 1.00 fit as linear. If the smooth was linear on both sides of the cut, we first estimated a model with separate linear functions estimated on each side of the cut, and compared this to a model with a single linear function. We compared the fit of these two models with chi-square deviance tests as well as changes in Akaike's information criteria (see Burnham and Anderson, 2003), privileging the model that fit the data better or was more parsimonious. After arriving on a final model, the reduction in τ_{11} variance was calculated ($pseudo\ R^2$) and the fit was compared to the impact model.

RESULTS

Impact Models

Figure 1 displays the fitted GAM for the impact models. For both OAKS and ORF, the GAM reduced to a linear fit for students who scored below the cut (EDF = 1.00), while there was some slight nonlinearity for students who scored above the cut (OAKS EDF = 2.14, ORF EDF = 3.18). The model was refit with the treatment slope constrained to be linear. As can be observed in Figure 1, there was essentially no discontinuity in the relation between the assignment variable and either outcome at the cut point. The mean probability of receiving treatment at the cut point was 0.73 for students who scored below the cut, and 0.25 for students who scored above the cut, resulting in a mean probability gap of 0.48, which was significant ($SE = 0.05$, $z = 9.90$, $p < .001$). The raw (sharp) gap was −0.06 for OAKS, which when divided by the probability gap resulted in a fuzzy gap estimate of −0.12 with an associated approximate standard error of 0.72, which was not significant ($z = -0.16$, $p = 0.87$). The raw gap for ORF was −1.16, resulting in a fuzzy estimate of −2.41, which was also not significant ($SE = -2.41$, $z = -0.79$, $p = 0.43$).

Figure 2 provides a visual representation of the variability in the treatment effect between schools for both outcomes. In each panel of the plot, the gray horizontal line represents the average raw gap, while the circles represent point estimates for specific schools. Each point estimate is also displayed with a 95 percent confidence interval. The y-axis represents raw-score deviations from the average effect. For example, the school on the furthest left for OAKS had an average treatment effect slightly less than one point below the average treatment effect of −0.06, while the school to the furthest right had an average treatment effect approximately one point above the average effect. The confidence intervals for essentially all schools cross the zero point, implying that they are not statistically different from the average. We compared these models to a more parsimonious model with the treatment effect variance fixed across schools. Both AIC and a chi-square deviance test indicated the more complex model fit the data better (OAKS $p = 0.048$, $\Delta AIC = -3$; ORF $p = 0.015$, $\Delta AIC = -4$). In both models, approximately 2 percent of the total variance was attributable to treatment effect variance.

Post-hoc Exploratory Analyses

Following the estimation of our impact models, we explored whether observational data would account for between-school treatment effect variability. Note that these analyses used the sharp RD estimates. Table 1 includes a summary of the weighted grand means and standard deviations for the school-level covariates, including (a) proportion of students

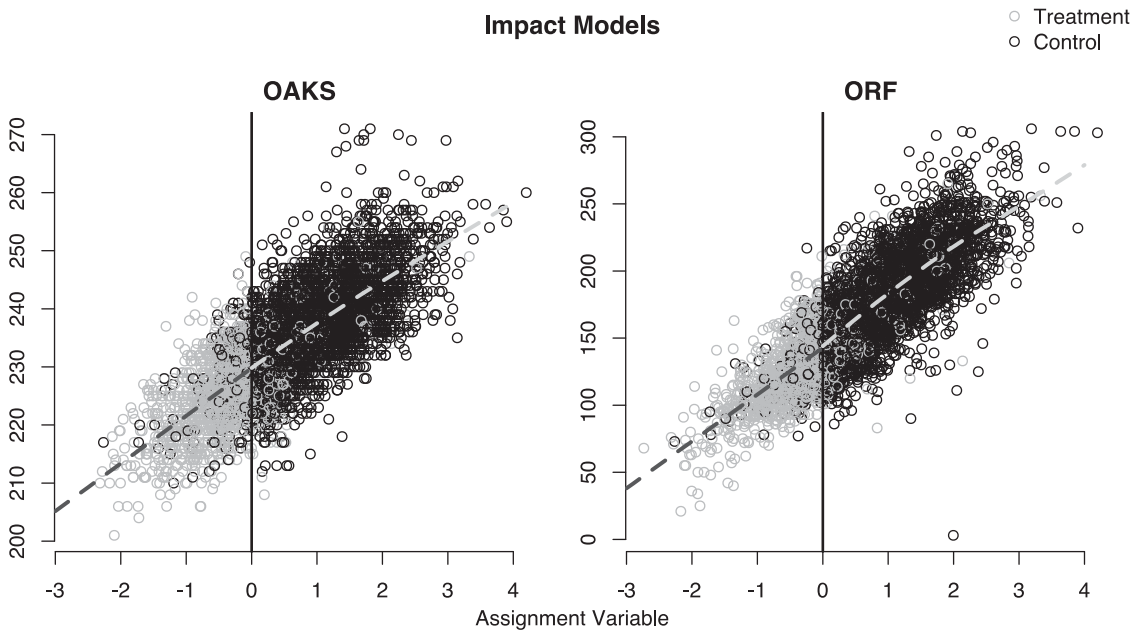

FIGURE 1 Impact models for the fuzzy regression discontinuity design estimated with the multilevel generalized additive model.

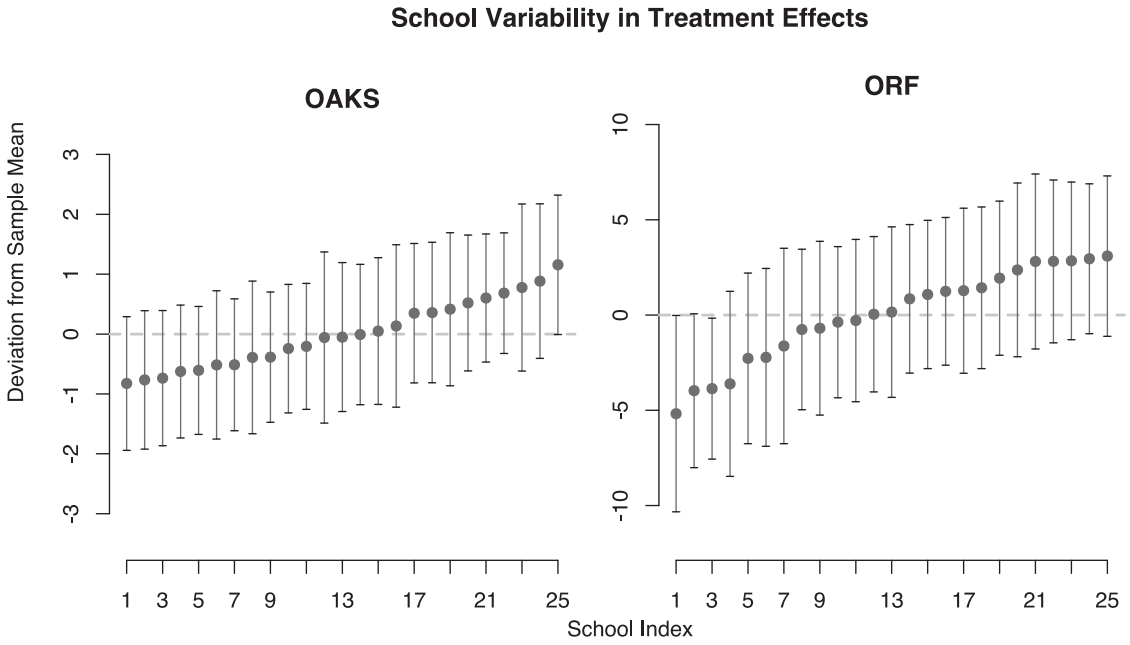

FIGURE 2 Variability in the treatment effect between schools. The dashed gray horizontal line represents the average effect (γ_{10}). Treatment effect estimates for each school are displayed with 95 percent confidence intervals.

TABLE 1
Descriptive Statistics for School-Level Covariates

Variable	*M*	*SD*	*Min*	*Max*
Foundational Instruction variables				
Alignment proportion	0.49	0.32	0.00	1.00
Hours	15.22	9.26	0.00	40.23
Interactions	2393.43	1892.59	0.00	6778.95
Higher-order instructional variables				
Alignment proportion	0.52	0.30	0.00	1.00
Hours	31.63	15.37	7.56	68.49
Interactions	5924.56	3067.56	1203.81	15239.12

Note. Descriptive statistics represent the mean of the school means, and the standard deviation among the school means, respectively (all variables were aggregated to the school-level, and entered as school-level covariates in the models). All variables were scaled to have a mean of 0 and standard deviation of 1.0 prior to be entered in the models.

whose reading need matched the instruction they received, (b) average hours of instruction, and (c) average number of teacher-student interactions.

Foundational Reading Models

For OAKS, several smooths approximated linearity in the initial fit. The model was refit with each constrained to be linear. The proportion of students whose instruction matched their need was significant for the treatment group ($\gamma = 0.83$, $SE = 0.29$, $p < 0.01$). This variable was linear and so interpretation was straightforward, with treatment students scoring 0.83 points higher, on average, for every one standard deviation increase in the proportion of students whose instruction matched their need at the school. Given the scale of the OAKS ($\mu = 234.60$, $SD = 8.73$), this was a small, though significant, effect. The smoothed effect of teacher-student interactions was also approaching significance for the treatment group ($p = 0.08$). Collectively, the measures of features of reading interventions accounted for approximately 55 percent of the between-school variability in the treatment effect. However, the model did not fit the data significantly better than the impact model ($p = 0.09$).

For ORF, all of the measures of reading intervention features with the exception of hours for the treatment group (EDF = 1.72) approximated a linear fit (EDF = 1.00). These were all constrained to be linear and models were compared with separate versus single slopes. Models with separate slopes did not fit the data better than models with a single slope. Only one of the measures was significant: the average number of teacher-student interactions employed during reading interventions. However, this effect was negative, implying that an increase in the average number of interactions at the school led to lower fluency scores, on average ($\gamma = -3.03$, $SE = 1.35$, $p = 0.03$). The proportion of students whose instruction matched their need also approached significance ($\gamma_{15} = 1.84$, $SE = 1.11$, $p = 0.10$). Collectively, the measures of reading intervention features accounted for 15 percent of the variance in ORF scores, and the model fit the data significantly better than the impact model ($p = 0.01$).

Higher-order Models

For OAKS, all higher-order instructional variables reduced to linear fits. The hours of instruction variable fit the data better with separate slopes for treatment and control

($p = 0.04$), while all other variables were adequately described by a single slope. The only significant effect was for hours of instruction for the control group ($\gamma = 0.35$, $SE = 0.14$, $p = 0.01$). Again, however, because of the scale of the OAKS measure, this was a modest effect, with a one standard deviation increase in hours relating to a 0.35 point gain for control students, on average. Collectively, the measures of reading intervention features accounted for 7 percent of the treatment effect variance. The model did not fit the data significantly better than the impact model ($p = 0.11$).

For ORF, there was not enough information in the data to reliably estimate the variance-covariance matrix when smoothing the observational measures with the GAM. This was evidenced by a degenerate variance-covariance matrix with an estimated correlation between intercept variance and treatment effect variance of -1.0 when any of the observational measures were smoothed. The final model therefore included only linear fits for the reading intervention features while maintaining the smooth on the assignment variable for the control group. For both the hours of instruction and teacher-student interactions variables, a single slope adequately described the relation for both treatment and control, while for the instructional match variable separate slopes displayed the better fit to the data ($p < 0.01$). None of the individual predictor variables were significant; however, collectively the variables accounted for approximately 41 percent of the treatment variance, and the model fit the data significantly better than the impact model ($p = 0.01$).

DISCUSSION

The primary objective of this study was to determine the effect of district-adopted and district-implemented reading interventions on 8th grade students' reading achievement. We utilized a RD design to screen 5,753 students into either a school-determined reading intervention, or the comparison condition. Schools were able to choose their own cut scores for determining how many students to serve in intervention, so long as the percentage of students receiving intervention was at least 20 percent of the school population.

A secondary objective was to determine if the treatment effect varied across schools. In other words, we examined whether there was variability in the discontinuity, between the cut scores used to screen students into the intervention compared to students screened out of the intervention. Once we determined whether there was variability between schools in the treatment effect, we examined school-level factors hypothesized to correlate with improved reading outcomes to account for differences in the treatment effect between schools. We expected school-level intensity factors (i.e., the frequency of teacher-student interactions during instruction), dosage (i.e., hours of intervention received), and alignment between need (as measured by the OAKS and the PRF) and the focus of the intervention provided (as measured by the MSIP-COT) to correlate with improved reading outcomes.

Results Summary

Overall, there was not a significant impact of the intervention on either the state outcome assessment (OAKS) or the passage reading fluency measure (PRF). For OAKS, the fuzzy gap estimate of -0.12 was not significant ($z = -0.16$, $p = 0.87$). For PRF, the fuzzy estimate of -2.41 was also not significant ($SE = -2.41$, $z = -0.79$, $p = 0.43$). While the null effects are disappointing, they are not entirely surprising when we examine the level of intensity of the interventions that schools delivered to students with protracted reading difficulties in the middle grades. For example, many of the interventions determined to be effective in reviews of adolescent literacy interventions involved support for teachers to implement the intervention, whereas the current study evaluated districts' existing intervention practices and did not provide teachers with coaching or professional development. Additionally, the interventions in the Striving Readers evaluation (Boulay et al., 2015) that were shown to be effective for improving reading achievement were typically delivered for 90 minutes per day, five days per week, whereas reading interventions schools selected and implemented in the current study were administered for 47 minutes per day, 4.3 days per week.

When considering null effects of any particular intervention, one could consider three sources for the interpretation of null effect: (1) methodological failure, (2) implementation failure, and (3) theory failure (Seftor, 2017). We content that the utilization of a regression discontinuity design with approximately 600 subjects provides a firm basis to reliably conclude that the intervention effects were indeed null. That leaves theory failure or implementation failure as potential explanatory factors. For example, if evidence suggested that 25 milligrams of a statin reduced the likelihood of having a heart attack or other complications related to heart failure, and a study reported that individuals were advised to take 25 mg per day, yet they reported taking, between 5 mg and 15 mg on average, one might conclude that implementation failure contributed to the null, or even negative, effects if the study reported less than desirable outcomes.

Conversely, if implementation failure can be ruled out (e.g., implementation fidelity is high and not variable), then one might consider theory failure as a culprit for null effects. For example, if researchers conduct a theory-driven study that posits that students at risk for math difficulty in the early grades require more time to grapple with math concepts and to discover math principles on their own, as opposed to being explicitly taught those principles, and results suggest both (a) that implementation was high and (b) that the intervention resulted in null or negative effects, theory failure must be considered. Consequently, such questions should be addressed sequentially (i.e., first, we examine implementation, and then we examine intervention impact). If implementation failure cannot be ruled out, it becomes difficult, if not impossible to determine whether theory failure occurred or not.

In the current study, two important findings temper our research question and hypotheses related to school factors that might account for school-to-school differences in the treatment effect. First, while there was significant variability

between schools on the treatment effect, it was such a small portion of the total variance of student reading outcomes (2 percent), that our explanatory analyses for hypothesized factors are tenuous. Second, if one considers whether schools implemented what research suggests should be implemented in terms of intervention intensity, one could consider the null effects of the current study a result of implementation failure. Taken together, these two findings make any meaningful interpretation of the explanatory factors of school-to-school differences in the treatment effect difficult.

Limitations

There are several limitations with the current study. Some limitations are inherent with using RD designs to causally determine whether an intervention is efficacious or not. For example, RD designs depend on researchers fully knowing the functional form of the regression between pre- and post-test. If the functional form is curvilinear, for example, forcing a linear model on the data might indicate a discontinuity at the cut score, not because an intervention was effective, but because the model is mis-specified. Such discontinuities could be an artifact of nonlinearity at or around the chosen cut score. We used a generalized additive model (GAM) with the degree of smoothing estimated from the data to account for any nonlinearity in the data and guard against this potential threat. Further, in the context of a study with null effects, and where schools chose their own cut scores, it is not plausible that nonlinearity issues would affect multiple areas along the cut score distribution. A second limitation related to using RD designs to evaluate intervention efficacy is that most of the power is tightly bound around the chosen cut score. Again, this limitation is somewhat mitigated by the fact that while power may affect the precision of the estimate, it should not necessarily affect the estimate itself, which was very small, and negative in some cases.

Other limitations include the lack of any proximal measures in the post-test measurement net. The OAKS state assessment is a summative test of reading achievement. While one could make a case for the more proximal nature of the PRF measure, the consistent finding of null effects on the fluency construct in reading intervention research (Gersten et al., 2009), and the case for PRF as a general outcome measure situate the measure as a more distal approximation of reading skill in the middle grades. Including measures of word attack or proximal measures of constructs taught in many of the reading interventions that schools delivered might have proven more sensitive to intended skill improvements. Finally, the post hoc nature of the questions related to explaining school-to-school differences in the treatment effect, coupled with the fact that a significant, but very small amount of variance existed between schools (2 percent), make efforts to explain differences in the treatment effect extremely tentative. In addition, our efforts to explore school-to-school variance used the *sharp* RDD estimate, even though the design was inherently fuzzy, with some students scoring above the cut receiving treatment and some students scoring below the cut not receiving treatment. In other words, although the overall impact of the intervention was determined by accounting for the fuzzy design, the post-hoc analyses ignored the fuzziness. As such, the exploratory results should be interpreted with caution.

Implications for Research and Practice

Although there are some technical limitations to using RD designs to evaluate intervention efficacy, the design also has many strengths and much merit for conducting rigorous evaluations, as they have the ability to leverage practices that many schools already utilize. For example, many schools implement multi-tier systems of support and/or RTI models that incorporate screening methods to assign students to tiers of increasingly intensive interventions. If schools are willing to maintain student assignment to the intervention for a fixed period of time without introducing "comparison" students to the intervention, they meet the basic parameters for rigorously evaluating a range of interventions in various content areas across grade levels. By partnering with a university research center or other external evaluator, or otherwise employing methodologists, these districts may be well prepared to conduct a rigorous evaluation of intervention practices.

Another merit of RD designs is that, in contrast to randomized-controlled trials, those students that are most in need of the intervention would have access to the treatment, avoiding the need to randomly assign some students that need an intervention into the treatment group and others into a control group. This may not be such an issue for interventions that have not yet proven efficacious, but may be more difficult for interventions with promise of improving student outcomes. The current study demonstrates that districts can effectively partner with researchers and rigorously evaluate the effects of an 8th-grade reading intervention in a number of schools and districts serving a very large sample of students ($n = 5{,}753$).

Although in the current study there was very little variability between schools in the treatment effect, we believe we have offered a useful technical approach for using school-level hypothesized factors to predict school-to-school differences in the RD effect. While the RD models used may not be completely novel, employing them to describe differences in school-to-school variability in discontinuities is, to our knowledge, a new approach; we have not found evidence of such an approach in the education intervention research literature. Such theory-driven models to account for differences in treatment effects should be incorporated in more researcher-practitioner partnerships to determine what works (or not), and why a given intervention did or did not work.

In our very first discussions with each of the respective superintendents from the districts involved in the current study that occurred a year prior to beginning the Middle School Intervention Project, one of the district superintendents made a forceful case that they had many initiatives occurring in his district, and that he had no data to help him determine whether any given initiative was effective or not. He also lamented that initiatives seemed to come and go, with no continuity across school years or sometimes within a year. Each of the superintendents strongly agreed with this message and committed

to a multi-year focus on improving reading achievement for their middle school students, to reduce high school dropout, a consequence of poor achievement that has been notorious in Oregon. However, by the end of the project, only one of the six superintendents was still employed in project districts. Whereas the founding superintendents participated in the project design and subsequent meetings to discuss project progress, their replacements did not. This point is salient given what we know about the number of years it takes for comprehensive school reform efforts to significantly improve student outcomes (Borman et. al., 2002) Therefore, it seems warranted that districts and schools incorporate systems and structures to weather the frequent and predictable turnover of leaders.

A final implication for practice of this study involves the level of intensity that is necessary for solving such intractable issues as improving adolescent reading achievement. Although the evidence base is converging on the level of intensity that should be brought to bear to improve reading outcomes for older readers, and the schools and districts involved in the current study indicated they were acutely aware of the level of intensity necessary for reading interventions to be efficacious in the middle grades, they chose to implement intervention plans that fell quite short of the these evidence-based recommendations (Boulay et al., 2015; Herrera et al., 2016). As evidenced by the few studies of adolescent reading interventions that have improved reading achievement for students in middle school, the intensity required to turn the dial on student outcome is substantial. In addition, districts seeking to improve adolescent literacy outcomes should consider matching interventions to student need, using interventions that employ explicit instruction, and providing ongoing support for teachers to implement interventions.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E100041 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

REFERENCES

ACT Inc. (2006). *Act high school profile report: The graduating class of 2009 act national profile report*. Washington, DC: ACT Inc.

ACT Inc. (2009). *National overview. Measuring college and career readiness: The class of 2009*. Washington, DC: ACT Inc.

Alliance for Excellent Education. (2007). *High school dropouts in America*. Washington, DC: Alliance for Excellent Education.

Alliance for Excellent Education. (2008). *Dropouts, diplomas, and dollars: U.S. High schools and the nation's economy*. Washington, DC: Alliance for Excellent Education.

Alliance for Excellent Education. (2009). *High school dropouts in America*. Washington, DC: Alliance for Excellent Education.

Alonzo, J., & Tindal, G. (2008). Examining the technical adequacy of fifth-grade reading comprehension measures in a progress monitoring assessment system (Technical Report No. 0807). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Alonzo, J., Tindal, G., Ulmer, K., & Glasgow, A. (2006). *easyCBM online progress monitoring assessment system*. Eugene, OR: University of Oregon.

Anderson, D., Alonzo, J., & Tindal, G. (2010). easyCBM® mathematics criterion related validity evidence: Oregon state test (Technical Report No. 1011). Eugene, OR: Behavioral Research and Training, University of Oregon.

Baker, S. K., & Baker, D. L. (2007). Teaching English language learners to read in Oregon schools: Moving from theory to practice. Paper presented at the COSA Conference, Seaside, OR.

Baker, S. K., Crone, D. A., & Fien, H. (2010). The Middle School Intervention Project (MSIP). (U.S. Department of Education, Institute for Educational Sciences, CFDA Num: 84.305E, 2010-2015, Funding Number: R305E100043, awarded $7,251,436.00).

Baker, S. K., Smith, J. L. M., Fien, H., Otterstedt, J., Katz, R., Baker, D. L., et al. (2007). Three year report on Oregon Reading First: Impact and implementation (May 15). Eugene, OR.: Oregon Reading First Center.

Biancarosa, G., & Snow, C. E. (2006). *Reading next: A vision for action and research in middle and high school literacy*. Washington, DC: Alliance for Excellent Education.

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

Borman, G. D., Hewes, G. M., Overman, L. T., & Brown, S. (2002). Comprehensive school reform and student achievement: A meta-analysis. *Review of Educational Research*, *73*, 125–230.

Boulay, B., Goodson, B. D., Frye, M., Blocklin, M., & Price, C. (2015). *Summary of research generated by striving readers on the effectiveness of interventions for struggling adolescent readers (NCEE No. 2016–4001)*. Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Burnham, K. P., & Anderson, D. R. (2003). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.

Carnevale, A. P., & Desrochers, D. M. (2003). *Standards for what? The economic and demographic roots of standards based reform*. Princeton, NJ: Educational Testing Service.

Chen, Z.-Y., & Kaplan, H. B. (2003). School failure in early adolescence and status attainment in middle adulthood: A longitudinal study. *Sociology of Education*, *76*, 110–127.

Chib, S., & Greenberg, E. (2014). *Nonparametric bayes analysis of the sharp and fuzzy regression discontinuity designs*. St. Louis, MO: Washington University in St Louis.

Crone, D. A., Carlson, S. E., Haack, M. K., Kennedy, P. C., Baker, S. K., & Fien, H. (2016). Data-based decision-making teams in middle school: Observations and implications from the Middle School Intervention Project. *Assessment for Effective Intervention*, *41*, 79–93.

Diplomas count 2008: School to college. Can state p-16 councils ease the transition? (2008). Retrieved from https://www.edweek.org/ew/toc/2008/06/05/index.html

Doabler, C. & Nelson-Walker, N. J. (2009). Ratings of Classroom Management and Instructional Support. Available from the Center on Teaching and Learning at the University of Oregon, Eugene, OR.

Duffy, G. G., Roehler, L. R., Sivan, E., Rackliffe, G., Book, C., Meloth, M. S., et al. (1987). Effects of explaining the reasoning associated with using reading strategies. *Reading Research Quarterly*, *22*, 347–368.

EPE Research Center. (2008). *School to college: Can state p-16 councils ease the transition? A special supplement to education week's: Diplomas count 2008*. Bethesda, MD: Editorial Projects in Education Research Center.

Finn, J., & Owings, J. (2006). *Adult lives of at-risk students: The roles of attainment and engagement in high school No. NCES 2006–328)*. Washington, DC: National Center for Education Statistics.

Finn, J., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology*, *82*, 231–234.

Friedman, T. L. (2005). *The world is flat: A brief history of the 21st century*. New York: Farrar, Strauss, and Giroux.

Fuchs, L. S., Fuchs, D., Karns, K., Hamlett, C. L., Katzaroff, M., & Dutka, S. (1997). Effects of task-focused goals on low-achieving students with and without learning disabilities. *American Educational Research Journal*, *34*, 513–543.

Gersten, R., Complon, D., Santoro, L. E., Dimino, J., Linan-Thompson, S., & Tilly, D. (2008). *Response to intervention (RTI) & muttitier intervention for reading in the primary grades*. Washington, DC: U.S. Department of Education, Institute for Education Sciences.

Gersten, R. M., Fuchs, L. S., Williams, J. P., & Baker, S. K. (2001). Teaching reading comprehension strategies to students with learning disabilities: A review of research. *Review of Educational Research*, *71*, 279–320.

Greene, J. P., & Winters, M. A. (2005). Public high school graduation and college-readiness rates: 1991–2002 (Working paper No. 8). New York: Center for Civic Innovation at the Manhattan Institute.

Grossman, S. R., Shylakhter, I., Karlsson, E. K., Byrne, E. H., Morales, et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science*, *327*(5967), 883–886.

Guy, B., Shin, H., & Thurlow, M. L. (1999). State graduation requirements for students with and without disabilities (Technical Report No. 24). Minneapolis, MN: National Center on Educational Outcomes, University of Minnesota.

Harlow, C. W. (2003). Education and correctional populations (Bureau of Justice Statistics Special Report No. January 2003, NCJ 195670). Washington, DC: U.S. Department of Justice, Office of Justice Programs.

Herrera, S., Truckenmiller, A. J., & Foorman, B. (2016). Summary of 20 years of research on the effectiveness of adolescent literacy programs and practices (REL No. 2016–178). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast.

James, G., Witten, D., Hastie, T. J., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Vol. *103*). New York: Springer. https://doi.org/10.1007/978-1-4614-7138-7

Johnson, D. R., & Emanuel, E. J. (2000). *Issues influencing the future of transition programs and services in the United States*. Minneapolis, MN: Institute on Community Integration, University of Minnesota.

Kame'enui, E. J., Francis, D., Fuchs, L. S., Good, R. H., O'Connor, et al. (2002). *Final report on the analysis of reading assessment instruments for k-3*. Eugene, OR: Reading Assessment Committee, Institute for the Development of Educational Achievement, College of Education, University of Oregon.

Kamil, M. L., Borman, G., Dole, J., Kral, C. C., Salinger, T., & Torgesen, J. (2008). Improving adolescent literacy: Effective classroom and intervention practices: A practice guide. (No. NCEE #2008-4027). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Klingner, J. K., Vaughn, S., & Schumm, J. S. (1998). Collaborative strategic reading during social studies in heterogeneous fourth-grade classrooms. *The Elementary School Journal*, *99*, 3–22. https://doi.org/10.1086/461914

Laird, J., KewalRamani, A., & Chapman, C. (2008). Dropout and completion rates in the United States: 2006 (Report No. NCES 2008–053). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, US Department of Education.

National Association of State Directors of Special Education. (2006). *Response to intervention: Policy considerations and implementation*. Alexandria, VA: National Association of State Directors of Special Education.

National Center for Education Statistics. (2015). The nation's report card: 2015 mathematics & reading assessements. *National achievement level results, 8th grade*. Washington, DC: National Center for Education Statistics.

National Institute of Child Health and Human Development. (2000). *Teaching children to read: An evidence-based assessment of scientific research literature on reading and its implications for reading instruction*. Bethesda, MD: NICHD Clearinghouse.

Neild, R. C., & Balfanz, R. (2006). *Unfulfilled promise: The dimensions and characteristic's of Philadelphia's dropout crisis, 2000-2005*. Philadelphia, PA: Philadelphia Youth Network; University of Pennsylvania; Johns Hopkins University.

Nelson, N. J., Smith, J. L. M., Fien, H., Crone, D. A., Baker, S. K., & Kame'enui, E. J. (2016). Researcher-practitioner partnerships: Lessons learned from the first year of the Middle School Intervention Project (MSIP). *Journal of Special Education Leadership*, *29*, 46–59.

Nomi, T., & Allensworth, E. (2009). "Double-dose" algebra as an alternative strategy to remediation: Effects on students' academic outcomes. *Journal of Research on Educational Effectiveness*, *2*, 111–148. https://doi.org/10.1080/19345740802676739

Oregon Department of Education. (2007). Oregon early childhood foundations - birth to 3. Retrieved from http://www.ode.state.or.us/search/page/?id=1352.

Oregon Department of Education. (2012). Oregon's statewide assessment system. Test administration (Technical Report Volume No. 5). Salem, OR: Oregon Department of Education.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standarized observation can leverage capacity. *Educational Researcher*, *38*, 109–119.

R Development Core Team. (2016). *R: A language and environment for statistical computing (Version 2.2.5 (Very, Very Secure Dishes)) [Computer Software]*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rumberger, R. W. (1995). Dropping out of middle school: A multilevel analysis of students and schools. *American Educational Research Journal*, *32*, 583–625. https://doi.org/10.3102/00028312032003583

Scammacca, N., Roberts, G., Vaughn, S., Edmonds, M., Wexler, J., Reutebuch, C. K., et al. (2007). *Interventions for adolescent struggling readers: A meta-analysis with implications for practice*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Schochet, P. Z. (2008). Technical methods report: Statistical power for regression discontinuity designs in education evaluations No. NCEE 2008–4026). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, US Department of Education.

Schumaker, J. B., & Deshler, D. D. (1992). Validation of learning strategy interventions for students with learning disabilities: Results of a programmatic research effort. In B. Y. L. Wong (Ed.), *Contemporary intervention research in learning disabilities: An international perspective* (pp. 22–46). New York: Springer.

Seftor, N. (2017). Raising the bar. *Evaluation Review*, *41*(3), 212–239.

Shadish, W. R., Zuur, A. F., & Sullivan, K. J. (2014). Using generalized additive (mixed) models to analyze single case designs. *Journal of School Psychology*, *52*, 149–178. https://doi.org/10.1016/j.jsp.2013.11.004

Slavin, R. E., Cheung, A., Groff, C., & Lake, C. (2008). Effective reading programs for middle and high schools: A best-evidence synthesis. *Reading Research Quarterly*, *43*, 290–322. https://doi.org/10.1598/RRQ.43.3.4

Smith, J. L. M., Fien, H., Basaraba, D., & Travers, P. (2009). Planning, evaluating, and improving tiers of support in beginning reading. *Teaching Exceptional Children*, *41*(5), 16–22.

Smolkowski, K., & Gunn, B. (2012). Reliability and validity of the classroom observations of student-teacher interactions (costi) for kindergarten reading instruction. *Early Childhood Research Quarterly*, *27*, 316–328.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academny Press, National Research Council.

Solis, M., Miciak, J., Vaughn, S., & Fletcher, J. M. (2014). Why intensive interventions matter: Longitudinal studies of adolescents with reading disabilities and poor reading comprehension. *Learning Disability Quarterly*, *37*, 218–229.

Sugai, G., & Horner, R. H. (1999). Discipline and behavioral support: Preferred processes and practices. *Effective School Practices*, *17*(4), 10–22.

Torgesen, J. K., Houston, D. D., Rissman, L. M., Decker, S. M., Roberts, G., Vaughn, S., et al. (2007). *Academic literacy instruction for adolescents: A guidance document form the center on instruction*. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Wanzek, J., Vaughn, S., Scammacca, N. K., Metz, K., Murray, C. S., Roberts, G., et al. (2013). Extensive reading interventions for students with reading difficulties after grade 3. *Review of Educational Research*, *83*, 163–195.

Wood, S. (2006). *Generalized additive models: An introduction with r*. Boca Raton, FL: CRC Press.

Wood, S., & Scheipl, F. (2014). Gamm4: Generalized additive mixed models using mgcv and lme4 (Version 0.0-4) [Computer software]: The Comprehensive R Archive Network.

About the Authors

Hank Fien, Ph.D., is the Director of the Center on Teaching and Learning and the National Center for Improving Literacy at the University of Oregon and also serves as an Associate Professor in the College of Education. His research interests include reading and mathematics development in young children, instructional design, and empirically validated interventions aimed at preventing or ameliorating student academic problems.

Daniel Anderson is a Research Associate for Behavioral Research and Teaching at the University of Oregon. His research interests lie primarily with mining large-scale data and applications of data science to educational research - i.e., the intersection of computer science and statistics.

Nancy J. Nelson, Ph.D., is a Research Assistant Professor and CTL's Director of Clinic and Outreach. Dr. Nelson is a PI or Co-PI on eleven externally funded projects to develop, implement, or evaluate math and reading interventions, including the National Center on Improving Literacy. Dr. Nelson is a licensed school psychologist and special education teacher with expertise in the implementation of academic interventions and the use of data-based decision making to support students in multi-tiered systems of support.

Patrick C. Kennedy, Ph.D., is a Research Associate at the Center on Teaching and Learning at the University of Oregon. His research interests include analytic methods (e.g., hierarchical linear modeling, structural equation modeling, and Item Response Theory), observation-based measurement, formative assessment, and data-based decision-making.

Scott K. Baker is a Research Professor at the Center on Research and Evaluation (CORE) at Southern Methodist University (SMU), and a Senior Research Associate at the Center on Teaching and Learning (CTL), University of Oregon. Dr. Baker is interested in the impact of interventions on child outcomes, mechanisms that underlie effective interventions, and how intervention impact varies by factors intrinsic and extrinsic to the child.

Mike Stoolmiller is an assistant professor at Michigan State University, College of Human Medicine. His research interests include reading and mathematics proficiency in childhood, health risking behaviors in early adolescence and advanced statistical methods for social science research designs.