2-1-2019

Longitudinal Effects of Rti Implementation On Reading Achievement Outcomes

Sally Grapin
Montclair State University, grapins@montclair.edu

Nancy Waldron
University of Florida

Diana Joyce-Beaulieu
University of Florida

Received: 6 April 2018 | Accepted: 24 November 2018

DOI: 10.1002/pits.22222

RESEARCH ARTICLE

WILEY

Longitudinal effects of RtI implementation on reading achievement outcomes

Sally L. Grapin[1] | Nancy Waldron[2] | Diana Joyce-Beaulieu[2]

[1]Department of Psychology, Montclair State University, Montclair, New Jersey

[2]School of Special Education, School Psychology, and Early Childhood Studies, University of Florida, Gainesville, Florida

Correspondence
Sally L. Grapin, Montclair State University, Department of Psychology, 221 Dickson Hall, 1 Normal Avenue, Montclair, NJ 07043.
Email: grapins@montclair.edu

Abstract

Because several studies have investigated student outcomes in schools implementing Response to Intervention (RtI), relatively little research has investigated the impact of implementation on students' long-term achievement outcomes (i.e., several years after exposure). The purpose of this study was to describe one elementary school's RtI implementation process and to examine students' long-term reading comprehension outcomes following their exposure to various phases of implementation. Four cohorts of students who experienced different implementation phases (i.e., a baseline condition or Phases I, II, or III of implementation) during Grade 2 were subsequently followed across Grades 3, 4, and 5 to examine their outcomes on two reading comprehension measures. Results indicated that students who experienced the early phases of RtI implementation (i.e., Phases I and II) during Grade 2 generally had higher mean comprehension scores in Grades 4 and 5 than students in the baseline condition. Implications for practice and future research are discussed.

KEYWORDS
implementation, reading comprehension, response to intervention

1 | INTRODUCTION

A number of federal and state policies, including the Individuals with Disabilities Education Improvement Act (Individuals with Disabilities Education Improvement Act, 2004) and the Every Student Succeeds Act (Every Student Succeeds Act, 2015), have called for schools to evaluate their accountability systems, adopt data-based decision-making procedures, and implement evidence-based instructional practices. In response to these

demands, many schools have implemented Response to Intervention (RtI) models[1] (Balu et al., 2015). Broadly defined, RtI is a system of service delivery that provides high-quality instruction to students through a multitiered framework of prevention and intervention. The RtI model of service delivery is particularly well-suited for meeting the demands of legislation such as the IDEIA and ESSA because it employs a problem-solving framework to guide schools in making data-driven decisions and improving student outcomes (Fuchs & Fuchs, 2006; National Association of School Psychologists, 2016).

Essential features of RtI models include a focus on outcomes for all students, a tiered model of instructional supports, implementation of evidence-based practices for core instruction and intervention, periodic universal screening to identify students who need more intensive supports, progress monitoring of growth in interventions, and systematic use of assessment data to determine the effectiveness of instruction (Fuchs & Fuchs, 2006; Fuchs & Vaughn, 2012). Although RtI frameworks may comprise any number of tiers, traditional models generally incorporate three tiers of intervention. Tier 1 refers to the core instructional supports provided to all students in a school. This core supports should be evidence-based and sufficient for the majority of students in a school (approximately 80%) to meet grade-level academic expectations. Tier 2 interventions are provided to students who struggle academically despite access to Tier 1 supports. They are intended to be more intensive than core supports and, ideally, to accelerate students' rate of learning such that they are able to succeed as full-time participants in Tier 1 instruction (Fuchs & Fuchs, 2017). Tier 2 interventions typically are delivered in small-group settings. Finally, Tier 3 interventions constitute the most intensive level of academic supports. They typically are highly individualized (or conducted with even smaller groups than seen at Tier 2) and are designed to remediate more severe academic skill deficits. Collectively, these tiered instructional supports provide an integrated framework for the delivery of high-quality reading instruction to all students.

1.1 | Outcomes associated with RtI implementation

Over the past decade, a substantial body of evidence has indicated the effectiveness of multitiered instruction for improving academic outcomes for all students (e.g., O'Connor, Harty, & Fulmer, 2005; Vellutino, Scanlon, Zhang, & Schatschneider, 2008). For example, Vaughn et al. (2009) found that students who received intensive intervention within a tiered instructional model made significant gains in the areas of word reading and comprehension. Moreover, the provision of high-quality instruction and intervention has been shown to lead to a significant increase in the number of students who demonstrate proficiency in basic reading skills at the end of first grade (Denton, Fletcher, Anthony, & Francis, 2006).

A considerable body of research has explored the essential components of RtI (screening practices, evidence-based instruction, etc.); however, there is a continuing need for large-scale, longitudinal research on the impact of implementation (Denton, 2012; Hughes & Dexter, 2011; Mellard, Frey, & Woods, 2012). Several studies have investigated student and systems outcomes associated with full-scale RtI implementation. Collectively, this study has suggested that RtI implementation is associated with greater accuracy and decreased numbers of special education referrals, improvements in student achievement, and reduced assessment and placement costs for districts (Burns, Appleton, & Stehouwer, 2005; Lembke, Garman, Deno, & Stecker, 2010; VanDerHeyden, Witt, & Gilbertson, 2007). Whereas these studies have focused largely on outcomes related to early reading skills and special education placements, further research is needed to examine long-term reading outcomes associated with RtI implementation (i.e., students' reading comprehension performance in the latter elementary grades; Hughes & Dexter, 2011).

[1]Many state education agencies, including the Florida Department of Education (FLDOE; n.d.), have transitioned from describing Response to Intervention (RtI) models to describing multitiered systems of support (MTSS). MTSS integrates RtI concepts of intervention and disability identification with principles of systems-level change and school-wide data analysis (Kansas Department of Education, 2010; Sulkowski & Joyce-Beaulieu, 2014). In this study, RtI is conceptualized as a systems-level framework whose goals and scope are aligned with MTSS.

More recently, the National Center for Education Evaluation and Regional Assistance of the Institute for Education Sciences (IES) commissioned a nationwide evaluation of RtI's impact on student reading achievement (Balu et al., 2015). In this evaluation, Balu et al. identified 146 impact schools across 13 states. Impact schools were identified as those schools that were fully implementing RtI, including all of the following practices, for a minimum of three years: (1) Use of three or more tiers of increasing intensity for reading instruction; (2) use of universal screening practices at least two times per year; (3) use of data-based decision-making for placing students in Tiers 2 or 3; and (4) use of progress monitoring data (beyond universal screening) to determine the effectiveness of Tier 2 and Tier 3 interventions. Because random assignment to control and treatment conditions was not possible, the authors used a regression discontinuity design. Presumably, this design involved designating a universal screening cut score, whereby students above the cut score received the intervention (i.e., Tier 2 or Tier 3 interventions) and students below the cut score did not. By examining outcomes for students who performed either slightly below or above the cut score, the authors intended to investigate the impact of RtI interventions on reading achievement. Findings from Balu et al.'s evaluation indicated that the tiered interventions had a significant negative effect on first-grade students' scores on the Early Childhood Longitudinal Study, Kindergarten Cohort Reading Assessment (ECLS-K) but no significant effect on their scores on the Test of Word Reading Efficiency, 2nd Edition (TOWRE-2). Moreover, the interventions had no significant effect on second-grade students' scores on the TOWRE-2 or third-grade students' scores on their end-of-year state tests of reading achievement.

Although this study has been construed as evidence of RtI's ineffectiveness, several design features suggest that it ultimately did not assess outcomes of RtI intervention per se. For example, only 89 of the 146 selected impact schools reported having at least one student in Tiers 1, 2, and 3 among first graders (Balu et al., 2015). This suggests that the remaining 57 schools (and perhaps others as well) were not implementing RtI in an appropriately rigorous manner (or, at least, per the evaluation team's own criteria; Fuchs & Fuchs, 2017). Moreover, contrary to the study's intended design, many students who performed above the designated screen cut score (i.e., students who did not meet the criterion for receiving tiered interventions) were placed in Tier 2 and 3 interventions anyway. As noted by Arden, Gandhi, Edmonds, and Danielson (2017), 45% of schools reported providing Tier 2 interventions to first-grade students who performed above the screening cut score. This suggests that, rather than comparing outcomes for students who did and did not participate in Tier 2 and 3 interventions, the authors ultimately compared outcomes for students who fell either below or above a designated cut score (Fuchs & Fuchs, 2017).

Gersten, Jayanthi, and Dimino (2017) suggested that more field evaluations of RtI are needed to address questions left unanswered by the IES national evaluation. In particular, these authors contended that smaller field evaluations should include both treatment and control groups, or what they referred to as "intervention and "business-as-usual" conditions (p. 252). Designs that incorporate both types of conditions would allow researchers to better understand and trace the impact of RtI interventions on student achievement outcomes.

Because numerous well-designed studies have documented the positive effects of high-quality Tier 2 and 3 reading interventions (Gersten, Newman-Gonchar, Haymond, & Dimino, 2017), critics have argued that the results of Balu et al.'s (2015) national evaluation speak more to widespread problems with RtI implementation than to the efficacy of the tiered interventions themselves (Arden, et al., 2017; Fuchs & Fuchs, 2017; Gersten et al., 2017). As reiterated by Arden et al. (2017) and others (e.g., Fixsen, Naoom, Blasé, Friedman, & Wallace, 2005), "*how* implementation occurs matters just as much as *what* is being implemented" (p. 271). Ultimately, high-quality implementation can only occur when school systems are prepared to engage in comprehensive systems change. This process involves gradually fostering school readiness and building capacity for full implementation.

1.2 | RtI and the systems-change process

As noted above, fully-integrated RtI models are established through a complex systems change process that requires collaborative problem-solving as well as careful evaluation of instructional practices (Batsche, Curtis, Dorman, Castillo, & Porter, 2007; Sansosti & Noltemeyer, 2008). One critical focus of this systems change process

is building capacity in schools for developing, adapting, and sustaining practices that meet the needs of all learners (Fullan, 2016). This involves garnering a collective commitment to improving student outcomes and pursuing changes that are consistent with a shared vision for reform and that promote program coherence (Hargreaves & Shirley, 2008).

Consistent with the Fullan's (2016) model of educational change, RtI implementation can be conceptualized as occurring over three critical phases. The first phase (often referred to as the *initiation phase*) involves garnering stakeholder support and laying the foundations for substantive change. It requires the establishment of consensus regarding the rationale and design of services, the development of a shared vision for instruction, and the identification of training needs. The second phase, often referred to as the *implementation phase*, constitutes the system's first experiences in implementing the various components of the reform. In the context of RtI implementation, this may involve establishing problem-solving and student support teams, developing and refining secondary and tertiary interventions, and establishing data-based decision-making procedures. Finally, the third phase involves the *continuation* or full implementation of the model. In this phase, various components of the model are refined (e.g., enhancing connections among instructional tiers), and further steps are taken to ensure long-term institutionalization. Fullan (2016) estimated that even moderately complex change initiatives may take anywhere between 2 and 4 years to progress through these three phases, with more complex or large-scale changes requiring between 5 and 10 years. Time to full implementation may be impacted by a number of variables, including the quality of instruction before implementation, stakeholder support, and the extent to which supportive resources are made available at the district and state levels (Sansosti & Noltemeyer, 2008).

Because the components and implementation of various RtI models have been described in extensive detail (e.g., Batsche, et al., 2007; Lembke et al., 2010), little is known how about systems shifts toward RtI implementation impact student achievement over time. Lembke et al. (2010) described the RtI implementation process and its associated student outcomes in a diverse, K-5 Midwestern elementary school. The authors identified eight core steps to implementing effective RtI models, including: (1) The establishment of school-based problem-solving teams; (2) selection of an evidence-based, formative assessment system; (3) examination of core academic programs for supporting all learners; and (4) identification of Tier 2 and 3 interventions and procedures for delivering these interventions. Implementation of tiered supports in reading occurred over three academic years (i.e., 2004–2007). Lembke et al.'s results indicated that, by the school's second year of implementation, greater numbers of students were meeting academic benchmarks with Tier I supports alone, and fewer students demonstrated the need of Tier III supports.

Although Lembke et al.'s case study provided a rich description of the systems-change process associated with RtI implementation, they evaluated student outcomes in a descriptive manner only. In other words, they did not use inferential statistics to determine whether students' performance on academic measures changed significantly from year to year. In addition, like most studies of RtI implementation to date, Lembke et al. evaluated immediate student achievement outcomes associated with implementation (i.e., students' academic achievement during the years in which the changes were implemented). Ultimately, further research is needed to examine student outcomes over multiple phases of RtI implementation, including both formative and advanced stages of change. Moreover, additional research is needed to understand how shifts toward RtI implementation impact students' long-term academic achievement (i.e., their academic performance several years after their exposure to the model).

1.3 | Present study

The purpose of this study was to investigate the impact of the systems change process (in relation to RtI implementation) on later student achievement. More specifically, this study examined students' reading comprehension outcomes in the latter elementary grades following their exposure to RtI implementation in Grade 2. The following research questions summarize the objectives of this study: (1) How did a shift in reading instructional practices in Grade 2, as a result of gradual RtI implementation, impact students' later achievement in

Grades 3–5? and (2) are there significant differences in long-term reading comprehension performance among students who experienced different phases of RtI implementation? We hypothesized that students who experienced later phases of implementation (i.e., phases in which the model was more developed) during Grade 2 would perform significantly better on measures of reading comprehension in Grades 3–5.

2 | METHOD

2.1 | Participants and setting

Participants were 489 students enrolled in a K-5 public, university-affiliated research school in Florida. The student population of this school is selected by lottery and designed to reflect racial and income demographics represented across the state of Florida. The racial and ethnic composition of students in this school was as follows: 47% Caucasian, 27% African American, 17% Hispanic, 2% Asian, and 5% multiracial or other backgrounds. Approximately 22% of students were identified as having a disability (under the Individuals with Disabilities Education Improvement Act, 2004), and 18% qualified for free or reduced-price lunch. Approximately half identified as female (49%) and half as male (51%).

Students in this school commute from a variety of surrounding small and rural cities and towns in North Florida. The school employs a highly trained faculty and staff. Approximately 78% of teachers possess graduate degrees.

2.2 | RtI implementation process

This school initiated RtI implementation in 2004 and achieved full implementation in 2010. The development of its RtI model occurred over three phases of change (i.e., Phases I, II, and III), each of which lasted approximately 2 years. Table 1 provides timeframes and descriptions of the essential systems changes that occurred during each phase.

During the Baseline Phase, core instruction was largely fragmented, with teachers implementing a variety of curricula and instructional strategies in K-2 classrooms. Subsequently, Phase I change initiatives centered on strengthening core instruction, such that all students received high-quality instruction in the five main areas of reading identified by the National Reading Panel (2000). This was achieved in part through participation in the Florida Reading Initiative (FRI), a research-based school-wide reform effort dedicated to ensuring the provision of

TABLE 1 Descriptions of phases

Phase	Years	Description
Baseline	2001–2002 2002–2003	• Business as usual • No formal or systematic pullout intervention
Phase I	2003–2004 2004–2005	• Participation in the Florida Reading Initiative (FRI) • Enhanced core instructional practices to address five main areas of reading identified by the National Reading Panel (2000) • Developed pullout intervention for students with the most intensive needs • Implemented universal screening procedures (using CBM)
Phase II	2005–2006 2006–2007	• Adoption of evidence-based intervention programs • Development of decision-making framework for identifying students for pullout intervention (based on CBM data) • Review of screening data at grade level problem-solving team meetings
Phase III	2007–2008 2008–2009	• Full implementation of RtI (e.g., progress monitoring, decision rules fidelity) • Tier 2 teacher-directed instruction in small groups ($4 \leq n \leq 6$) • Enhanced connections between the core and Tier 2 instruction • Developed Tier 3 instruction with an interventionist (individual and groups of $n < 4$)

Note. CBM: curriculum-based measure.

evidence-based reading instruction to all students (Batsche et al., 2007). Along with participants from approximately 80 other schools statewide, teachers in this school received intensive, ongoing professional development in providing explicit, systematic instruction within a 90-min core reading block. Whereas reading interventionists provided both push-in and pullout classroom supports to struggling students, these supports were broadly targeted and not necessarily matched to specific student needs in a systematic manner. To monitor fidelity, reading interventionists maintained intervention schedules that were monitored by a head reading coach.

The focus of Phase II was on implementing critical components of the multitiered framework. This involved establishing problem-solving teams for reviewing school-wide data, distinct secondary and tertiary tiers of intervention, and data-based decision-making procedures. For Tier 2 and Tier 3 interventions, more intentional, systematic efforts were undertaken to match specific, evidence-based interventions to the particular skill needs of students. These evidence-based interventions included programs such as Road to the Code (Blachman, Ball, Black, & Tangel, 2000) and Great Leaps (Mercer & Campbell, 1998). The school's head reading coach for grades K-2 continued to ensure fidelity by monitoring intervention schedules. In addition, problem-solving team meetings were held quarterly to discuss the effectiveness of core instruction and to identify students in need of further intervention. Participants in team meetings included classroom teachers, interventionists, and school leaders.

Also during Phase II, progress monitoring procedures for the group and individualized interventions were introduced. These progress monitoring tools included measures from the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), such as the Phoneme Segmentation Fluency (PSF), Nonsense Word Fluency (NWF), and Oral Reading Fluency (ORF) probes. Data from universal screening and progress monitoring measures were reviewed at problem-solving team meetings.

The third and final phase (Phase III) marked the full implementation of the model. In this phase, personnel refined and differentiated the intensity of secondary and tertiary interventions, enhanced connections among the three tiers, and further clarified the roles of classroom teachers and interventionists. Specifically, classroom teachers assumed responsibility for implementing Tier 2 interventions, whereas reading interventionists were responsible for implementing pullout Tier 3 interventions. Classroom teachers received extensive, ongoing training from the head reading coach regarding how and when Tier 2 interventions would be implemented. For both classroom teachers and interventionists, options for evidence-based interventions expanded to include programs such as Wilson Fundations (Wilson Reading Systems, 2002).

Throughout Phase III, students' progress in Tier 2 and Tier 3 interventions continued to be reviewed regularly at problem-solving team meetings, which increased in frequency from quarterly to monthly. Screening and progress monitoring measures were administered by trained staff who had previously achieved appropriate inter-rater reliability for administration and scoring. To monitor implementation fidelity during Phase III, periodic observations of Tier 2 and Tier 3 groups were conducted by the head reading coach. Moreover, interventionists maintained detailed intervention logs that were monitored by the head reading coach and reviewed at the monthly problem-solving team meetings.

2.3 | Measures

2.3.1 | Florida Comprehensive Assessment Test (FCAT)

The Florida Comprehensive Assessment Test (FCAT) is a state-developed assessment of achievement in the areas of reading, math, and science. It was administered in a group format to students in Grades 3–5 in the spring of each academic year. Specifically, the Reading section of the FCAT required students to read a series of literary and informational passages that ranged from 100 to 700 words in length. Based on these passages, students answered between 50 and 55 multiple choice questions that assessed skills such as identifying main ideas, plot, and purpose, vocabulary, and inferential reasoning. The FCAT Reading section provides a Developmental Scale Score (DSS) that ranges from 86 to 3,008.

Regarding the technical adequacy of the FCAT, internal consistency reliability coefficients (i.e., Cronbach's alpha values) for the third, fourth, and fifth-grade Reading sections range from 0.85 to 0.89 (Harcourt, 2007). The criterion-related evidence of the validity of this test with several other measures of language, basic reading, and reading comprehension skills has also been established (Schatschneider et al., 2004). For example, correlations of scores on this test and the Stanford Achievement Test (SAT) have been found to range from 0.70 to 0.81 (Crist, 2001).

2.3.2 | Gates MacGinitie Reading Tests, 4th Edition

The Gates MacGinitie Reading Tests (GMRT; MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2000) are group-administered, norm-referenced, broad-based tests of reading achievement that measure performance in areas such as decoding, comprehension, and word knowledge across a range of grade and age levels. In particular, students completed the Comprehension subtest of this measure, in which they were presented with a series of passages and subsequently prompted to answer relevant comprehension questions. Items on the Comprehension subtest assessed students' skills in understanding text, making inferences, and determining the meaning of words using contextual information. The GMRT generates an Extended Scale Score (ESS) for test takers. Internal consistency reliability coefficients exceed 0.90 for the third, fourth, and fifth-grade Comprehension subtests of the GMRT (Johnson, 2005).

2.4 | Study design and analysis

Participants were subdivided into four cohorts, each of which completed Grade 2 during one of four phases: Baseline, Phase I, Phase II or Phase III (as described above). The Baseline phase referred to a two-year period directly preceding Phase I in which the school was not implementing any components of RtI. Phases occurred across consecutive, two-year periods. For example, the Baseline cohort comprised participants who completed Grade 2 during either the 2001–2002 or 2002–2003 school year, and the Phase I cohort comprised participants who completed Grade 2 during either the 2003–2004 or 2004–2005 academic year.

All participants completed the FCAT and GMRT during Grades 3, 4, and 5, respectively. Students completed the FCAT and GMRT at the end of each academic year (i.e., spring). Archival test data were retrieved from school records. Due to missing school records, approximately 22.9% of data were missing from the sample. This rate of missing data is consistent with rates commonly reported in educational and psychological research (Peng, Harwell, Liou, Ehman, & Sawilowsky, 2006; Peugh & Enders, 2004). Accordingly, missing data were addressed using full information maximum likelihood (FIML) estimation.

Analyses included calculations of means and standard deviations as well as Pearson correlations between all measures. PROC MIXED, a SAS procedure that can conduct a variety of analyses, was used to conduct a multivariate analysis of variance (MANOVA) with phase as the independent variable (IV) and GMRT and FCAT scores at each of Grades 3, 4, and 5 as the six dependent variables (DVs). The model used in PROC MIXED allowed for the possibility of unequal variance-covariance matrices across phases. PROC MIXED provides full information maximum likelihood estimates and therefore data for all 489 children were included in the analysis. Hypothesis testing was conducted using the Kenward–Roger procedure, which provides improved estimates of standard errors and more accurate degrees of freedom that are obtained when the procedure is not used. See, for example, Littell, Milliken, Stroup, Wolfinger, and Schabenberger (2006). Following this analysis, 36 pairwise comparisons were conducted to identify potentially significant mean differences between phases at each grade level. Corrections for multiplicity were applied to all contrasts using the Benjamini–Hochberg (BH) procedure, which controls the false discovery rate (FDR; Benjamini & Hochberg, 1995), rather than the alpha level. The What Works Clearinghouse (What Works Clearinghouse, 2017) has recommended using the BH procedure to account for multiplicity when

multiple outcome measures in the same domain are tested. In the present analyses, an FDR of 0.05 was maintained across all 36 contrasts.

To provide evidence regarding the size of significant mean differences, Glass's delta was computed using the maximum likelihood estimates of the means for the contrast and the standard deviation from the Baseline phase for the test and grade. Like Cohen's *d*, Glass's delta represents the difference between two means divided by a standard deviation. For Cohen's *d*, the standard deviation is a pooled standard deviation that reflects average variability across groups, whereas for Glass's delta, the standard deviation is that of a single group (i.e., the control or baseline group). Glass's delta is most appropriate and commonly used as a measure of effect size when a control or baseline group is clearly delineated, as in the present study (e.g., Glass, 1976; Smith & Glass, 1977).

3 | RESULTS

Table 2 displays descriptive statistics (i.e., means and standard deviations) for FCAT and GMRT scores by grade level and phase. For Grade 3, FCAT and GMRT mean scores increased from Baseline to Phase I but decreased somewhat across Phases I, II, and III. For Grade 4, mean scores on both the FCAT and GMRT increased from Baseline through Phase II but declined slightly for Phase III. Finally, Grade 5 mean FCAT scores increased from Baseline through Phase II but declined slightly for Phase III. GMRT scores at this grade level increased from Baseline through Phase I but declined in Phases II and III. Table 3 displays maximum likelihood estimates of means, which were compared in the multivariate analysis described below.

Table 4 presents Pearson correlations between Grades 3, 4, and 5 FCAT and GMRT scores, respectively. As noted in the table, all correlations were statistically significant ($p < 0.001$). Generally, correlation coefficients were highest for measures completed during the same academic year. For example, correlation coefficients for the Grades 3 FCAT and GMRT ($r = 0.71$) and the Grades 5 FCAT and GMRT ($r = 0.75$) were the two highest values.

TABLE 2 Descriptive statistics FCAT and GMRT scores by phase

	FCAT			GMRT		
Phase	***n***	***M***	***SD***	***n***	***M***	***SD***
Grade 3						
Baseline	59	1,515.9	271.1	59	497.0	26.5
Phase I	100	1,525.4	332.0	111	500.0	40.3
Phase II	103	1,507.0	309.0	102	494.5	36.6
Phase III	99	1,432.1	313.2	106	489.7	38.6
Total	361	1,493.0	312.0	378	495.2	37.0
Grade 4						
Baseline	105	1,607.4	238.8	114	506.5	35.8
Phase I	105	1,663.4	307.9	112	521.7	41.7
Phase II	102	1,730.2	280.6	92	527.2	40.2
Phase III	111	1,633.7	319.6	91	521.4	50.0
Total	423	1,657.8	291.5	409	518.6	42.3
Grade 5						
Baseline	100	1,652.2	247.5	102	517.8	36.1
Phase I	89	1,762.2	361.8	97	536.4	45.5
Phase II	106	1,768.9	251.6	82	530.0	29.0
Phase III	58	1,727.2	310.9	56	529.2	32.4
Total	353	1,727.3	294.8	337	528.0	37.6

Note. FCAT: Florida Comprehensive Assessment Test; GMRT: Gates MacGinitie Reading Tests.

TABLE 3 Maximum likelihood estimates of means

Grade/phase	FCAT	GMRT
Grade 3		
Baseline	1,482.9	493.1
Phase I	1,516.1	500.4
Phase II	1,503.2	493.4
Phase III	1,454.0	492.1
Grade 4		
Baseline	1,613.4	506.6
Phase I	1,657.1	520.8
Phase II	1,731.5	525.2
Phase III	1,639.1	521.0
Grade 5		
Baseline	1,663.2	518.7
Phase I	1,752.1	534.1
Phase II	1,764.4	532.1
Phase III	1,679.5	523.5

Note. FCAT: Florida Comprehensive Assessment Test; GMRT: Gates MacGinitie Reading Tests.

A MANOVA was conducted to examine the extent to which phase accounted for variation in GMRT and FCAT scores. Results of this analysis indicated a significant main effect of phase on reading comprehension scores ($F(18, 298) = 2.31$, $p = 0.002$). To identify significant mean differences in FCAT and GMRT scores between phases, follow-up pairwise comparisons were conducted for each grade level. For Grade 3, no significant mean differences in scores for either test were observed between phases. However, for Grade 4, *t*-tests revealed a significant mean difference in FCAT scores between the Phase II and Baseline cohorts ($t(213.2) = 3.36$, $p < 0.001$, $g = 0.49$). They also revealed significant mean differences in Grade 4 GMRT scores between the Phase I and Baseline cohorts ($t(222.5) = 2.76$, $p = 0.006$, $g = 0.39$) and between the Phase II and Baseline cohorts ($t(237.5) = 3.75$, $p < 0.001$, $g = 0.51$). No other pairwise comparisons reached statistical significance for the Grade 4 FCAT or GMRT data.

Regarding Grade 5 test scores, results of pairwise comparisons indicated a significant mean difference in FCAT scores between the Phase II and Baseline cohorts ($t(224.9) = 2.96$, $p = 0.003$, $g = 0.39$). In addition, significant mean differences in Grade 5 GMRT scores were found between the Phase I and Baseline cohorts ($t(208.8) = 2.77$, $p = 0.006$, $g = 0.42$) and between the Phase II and Baseline cohorts ($t(217.5) = 3.00$, $p = 0.003$, $g = 0.37$). No other pairwise comparisons for Grade 5 data yielded significant results.

TABLE 4 Pearson correlations

	Grade 3 FCAT	Grade 3 GMRT	Grade 4 FCAT	Grade 4 GMRT	Grade 5 FCAT	Grade 5 GMRT
Grade 3 FCAT	1					
Grade 3 GMRT	0.71	1				
Grade 4 FCAT	0.66	0.61	1			
Grade 4 GMRT	0.56	0.53	0.57	1		
Grade 5 FCAT	0.66	0.66	0.70	0.60	1	
Grade 5 GMRT	0.61	0.66	0.63	0.58	0.75	1

Note. FCAT: Florida Comprehensive Assessment Test; GMRT: Gates MacGinitie Reading Tests; all correlations were statistically significant at $p < 0.001$.

4 | DISCUSSION

The results of this study revealed significant mean differences in later reading comprehension outcomes for students who experienced different phases of RtI implementation in Grade 2. Overall, patterns in reading comprehension test scores varied somewhat across Grades 3, 4, and 5. For Grade 3, descriptive data indicated a slight decline in mean FCAT and GMRT score across Phases I, II, and III, although mean scores for the Phase I cohort were higher than scores for the Baseline cohort. Notably, however, no significant mean differences in scores were observed between phases for Grade 3, indicating that these differences were not statistically meaningful.

Patterns in comprehension performance were somewhat different for Grades 4 and 5. Descriptive data indicated that Grade 4 FCAT and GMRT mean scores increased steadily across the Baseline, Phase I, and Phase II cohorts, although they declined slightly for the Phase III cohort. Despite the observed increase in scores across phases, only the Phase II cohort performed significantly better on the FCAT than the Baseline cohort. Pairwise comparisons for the GMRT indicated that both the Phase I and Phase II cohorts performed significantly better than the Baseline cohort on this test. As measured by Glass's delta, effect sizes for these contrasts ranged from 0.39 to 0.51. According to Cohen's (1988) criteria, these are best characterized as medium effect sizes.

Similarly, for the Grade 5 analyses, descriptive data indicated that mean FCAT scores steadily increased across the Baseline, Phase I, and Phase II cohorts but declined slightly for the Phase III cohort. Although mean GMRT scores did not exhibit this same pattern, they evidenced a sizeable increase from Baseline to Phase I. Pairwise comparisons revealed that students in the Phase I cohort performed significantly better on the FCAT than those in the Baseline cohort. Grade 5 students in both the Phase I and II cohorts performed significantly better on the GMRT than those in the Baseline cohort. Medium effect sizes were observed for these contrasts.

One of the most interesting and unexpected findings of this study was that significant increases in mean test scores between phases were evident for Grades 4 and 5 but not for Grade 3. This is surprising, given that students completed the Grade 3 measures only shortly after their exposure to the RtI-enhanced instruction (i.e., one year later), whereas they completed the Grades 4 and 5 measures 2–3 years following exposure. One possible explanation for these findings is that, as demands on reading comprehension skills increased across grades, a strong command of basic reading skills became increasingly important when students entered Grades 4 and 5. Notably, this school's Grade 2 reading instruction focused heavily on foundational skills in reading fluency, which has been empirically and theoretically linked to comprehension skills in elementary school-aged children (Fuchs, Fuchs, Hosp, & Jenkins, 2001; LaBerge & Samuels, 1974). Perhaps students who experienced the RtI-enhanced instruction in Grade 2 reaped the greatest benefits of this exposure in the latter elementary grades, which posed the most substantial demands in reading comprehension.

Another interesting finding of this study is that, when statistically significant mean differences in test scores were observed, they were most likely to be observed between the Baseline cohort and either the Phase I or II cohorts, respectively. Although not anticipated, these findings are somewhat unsurprising. Arguably, the most substantial phases of the RtI change process in this school were Phases I and II, given that Phase I involved the revamping of core instruction (to which all students were exposed) and that Phase II involved the introduction of a tiered system of supports. Although important, Phase III generally was focused on refining core features of the model that had already been established during Phases I and II.

It was unexpected, however, that increases in test scores were not sustained for the Phase III cohort in either Grades 4 or 5. One possible explanation for this finding is that the Phase III cohort simply comprised a larger number of struggling students than did the Phase I and II cohorts. Unfortunately, due to the cohort-sequential design of this study, it was not possible to compare the reading achievement of students in the Phase III cohort with that of students in preceding cohorts before RtI implementation.

Given that systems change often is not a linear process, it may not be all that surprising that students' achievement scores evidenced a minor decline in Phase III. In fact, when proceeding with educational reform, Fullan (2006) cautioned personnel against expecting linear improvements in student outcomes. Rather, he advised that

sustainability (which was a major focus of Phase III in this study) is a cyclical process in which innovative momentum may require periodic revival to maintain forward movement. Moreover, it is possible that "the set of strategies that brought initial success are not the ones—not powerful enough—to take us to higher levels" (Fullan, 2006, p.120). Consistent with Fullan's message, the decline in Phase III scores observed in this study may simply be a reflection of the nonlinear nature of the systems change process.

4.1 | Limitations

Several limitations of this study should be acknowledged. First, the FCAT is no longer used in the Florida K-12 schools, as it was replaced by the Florida Standards Assessments (FSA) in 2015. Although these two tests are similar in many respects, it is unclear whether outcomes from the present study could be generalized to other settings that are utilizing this new state test. Whereas it would have been desirable to examine systems-level change using the most up-to-date achievement measures, archival data were necessary to track student outcomes from the beginning of RtI implementation. Moreover, archival data were needed to establish a baseline condition, which Gersten et al. (2017) argued was necessary for producing higher-quality RtI implementation research. Since state tests are revised periodically, and RtI implementation requires at least three years (if not more) to be successful, it is logical that research in this area would need to rely on archival test data, to some extent.

Other limitations of this study concern its design. Although this study incorporated a baseline condition (unlike some prior research), it was not possible to rule out all potential threats to internal validity using the cohort-sequential design. For example, changes in achievement scores over time could have been attributable, in part, to other internal (e.g., staff turnover) or external events (e.g., changes in legislation) that directly or indirectly influenced instructional practices. Due to these design limitations, definitive causal inferences about the impact of RtI implementation on student achievement cannot be made. Future research that employs more rigorous experimental designs is needed to better understand the impact of RtI implementation on students' long-term reading achievement. Finally, because this study relied on archival data from more than a decade ago, some FCAT and GMRT data were missing. Whereas it is not possible to ascertain their true ramifications in this sample, missing data can potentially impact results (e.g., by leading to reduced statistical power; Newman, 2014). However, FCAT and GMRT data were missing primarily due to gaps in school records (which, in and of itself, would not lead to systematic patterns of missing data). Moreover, they were handled using recommended procedures, as described above (Newman, 2014).

4.2 | Implications and directions for future research

The findings of the present study have a number of implications for research and practice. First, the results of this study suggest that RtI implementation in the early elementary grades may impact students' long-term reading achievement, particularly in the area of comprehension. As suggested by Fullan (2006, 2016) and the present data, improvements in student achievement may be evident but nonlinear. School personnel may need to exhibit patience when evaluating student outcomes, whereas also advocating for timely change. They also may need to solicit innovative improvement strategies and reinvigorate fervor for reform, particularly in the latter stages of the change process.

As indicated above, further research is needed to investigate the impact of RtI implementation in the early elementary grades on students' later reading outcomes. In addition to examining reading achievement outcomes, researchers should consider a variety of other indicators, including special education referrals/placements and numbers of students receiving interventions at various tiers (Shapiro & Clemens, 2009). Moreover, these studies should strive to implement more robust experimental designs that allow researchers to clarify the causal relationships between variables. In particular, this study should use random assignment to intervention and control (i.e., "business as usual") conditions (Gersten et al., 2017). This study also should monitor fidelity of implementation.

Because this type of high-quality research is often logistically difficult to conduct in schools, it is essential for making sense of the disparate findings that have emerged from the RtI literature to date. Overall, further research in this area may help researchers and practitioners draw clearer conclusions about the impact of RtI implementation in K-12 schools.

ORCID

Sally L. Grapin http://orcid.org/0000-0001-8119-3681

REFERENCES

Arden, S. V., Gandhi, A. G., Zumeta edmonds, R., & Danielson, L. (2017). Toward more effective tiered systems: Lessons from national implementation efforts. *Exceptional Children*, *83*(3), 269–280.

Balu, R., Zhu, P., Doolittle, F., Schiller, E., Jenkins, J., & Gersten, R. (2015). *Evaluation of response to intervention practices for elementary reading (NCEE 2016-4000)*. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Batsche, G., Curtis, J., Dorman, C., Castillo, J., & Porter, L. J. (2007). The Florida problem solving/response to intervention model: Implementing a statewide initiative. In S. Jimerson, M. Burns, & A. VanDerHeyden. (Eds.), *Handbook of response to intervention: The science and practice of assessment and intervention* (pp. 378–395). New York: Springer.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300.

Blachman, B., Ball, E. W., Black, R., & Tangel, D. M. (2000). *Road to the code: A phonological awareness program for young children*. Baltimore: Brookes.

Burns, M. K., Appleton, J. J., & Stehouwer, J. D. (2005). Meta-analytic review of responsiveness-to-intervention research: Examining field-based and research-implemented models. *Journal of Psychoeducational Assessment*, *23*(4), 381–394.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Crist, C. (2001). *FCAT briefing book*. Tallahassee: Florida Department of Education.

Denton, C. A. (2012). Response to intervention for reading difficulties in the primary grades: Some answers and lingering questions. *Journal of Learning Disabilities*, *45*(3), 232–243.

Denton, C. A., Fletcher, J. M., Anthony, J. L., & Francis, D. J. (2006). An evaluation of intensive intervention for students with persistent reading difficulties. *Journal of Learning Disabilities*, *39*(5), 447–466.

Every Student Succeeds Act. (2015). P. L. 114-95.

Fixsen, D. L., Naoom, S. F., Blasé, K. A., Friedman, R. M., & Wallace, F. (2005). *Implementation research: A synthesis of the literature*. Tampa: University of South Florida. Retrieved from. http://ctndisseminationlibrary.org/PDF/nirnmonograph.pdf

Florida Department of Education (n. d.). *Florida Department of Education's multi-tiered system of support*. Retrieved from http://www.fldoe.org/finance/school-business-services/fl-department-of-edus-multi-tiered-sys.stml

Fuchs, D., & Fuchs, L. S. (2006). Introduction to response to intervention: What, why, and how valid is it? *Reading Research Quarterly*, *41*(1), 93–99.

Fuchs, D., & Fuchs, L. S. (2017). Critique of the national evaluation of Response to Intervention: A case for simpler frameworks. *Exceptional Children*, *83*(3), 255–268.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading*, *5*(3), 239–256. https://doi.org/10.1207/S1532799XSSR0503_3

Fuchs, L. S., & Vaughn, S. (2012). Responsive-to-intervention: A decade later. *Journal of Learning Disabilities*, *45*(3), 195–203. https://doi.org/10.1177/0022219412442150

Fullan, M. (2006). The future of educational change: System thinkers in action. *Journal of Educational Change*, *7*(3), 113–122.

Fullan, M. (2016). *The new meaning of educational change*. New York, NY: Teachers College Press.

Gersten, R., Jayanthi, M., & Dimino, J. (2017). Too much too soon? Unanswered questions from national Response to Intervention evaluation. *Exceptional Children*, *83*(3), 244–254.

Gersten, R., Newman-Gonchar, R. A., Haymond, K. S., & Dimino, J. (2017). *What is the evidence base to support reading interventions for improving student outcomes in grades 1–3?* Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast. (REL 2017–271). http://ies.ed.gov/ncee/edlabs

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher. Educational Researcher*, *5*(10), 3–8.

Harcourt (2007). *Reading and mathematics: Technical report for 2006 FCAT test administrations*. San Antonio, TX: Harcourt Assessment.

Hargreaves, A., & Shirley, D. (2008). The fourth way. *Educational Leadership*, *66*(2), 56–61.
Hughes, C. A., & Dexter, D. D. (2011). Response to intervention: A research-based summary. *Theory into Practice*, *50*(1), 4–11.
Individuals with Disabilities Education Improvement Act. (2004). Pub. L. No. 108-446, 118. Stat. 2647.
Johnson, K. M. (2005). Review of the Gates-MacGinitie Reading Tests, Fourth Edition, Forms S and T. In R. A. Spies & B. S. Plake (Eds.). *The sixteenth mental measurements yearbook* [electronic version]. Retrieved from EBSCOhost Mental Measurements Yearbook database.
Kansas Department of Education. (2010). *Kansas multi-tier system of supports: The integration of MTSS and RtI*. Retrieved from http://www.ksde.org/Portals/0/Title/ESOL/TheIntegrationofMTSSandRtI.pdf
LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*(2), 293–323.
Lembke, E., Garman, C., Deno, S., & Stecker, P. (2010). One elementary school's implementation of response to intervention (RTI). *Reading & Writing Quarterly*, *26*(4), 361–373.
Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). SAS system for mixed models, Cary, NC: SAS Institute Inc.,.
MacGinitie, W., MacGinitie, R., Maria, K., Dreyer, L., & Hughes, K. (2000). *Gates-MacGinitie Reading Tests-4*. Itasca, IL: Riverside.
Mellard, D. F., Frey, B. B., & Woods, K. L. (2012). School-wide student outcomes of response to intervention frameworks. *Learning Disabilities*, *10*(2), 17–32.
Mercer, C. D., & Campbell, K. U. (1998). *Great Leaps reading program, Kindergarten-2nd grade*. Gainesville, FL: Diarmuid.
National Association of School Psychologists. (2016). *Building capacity for student success: Every Student Succeeds Act opportunities*. https://www.nasponline.org/research-and-policy/current-law-and-policy-priorities/policy-priorities/the-every-student-succeeds-act/essa-implementation-resources/essa-and-mtss-for-school-psychologists
National Reading Panel (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*. Washington, DC: U.S. Department of Health and Human Services.
Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods*, *17*(4), 372–411.
O'Connor, R. E., Harty, K. R., & Fulmer, D. (2005). Tiers of intervention in kindergarten through third grade. *Journal of Learning Disabilities*, *38*(6), 532–538.
Peng, C.-Y. J., Harwell, M., Liou, S.-M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. In S. Sawilowsky (Ed.), *Real data analysis* (pp. 31–78). Greenwich, CT: Information Age.
Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of educational research*, *74*(4), 525–556.
Sansosti, F. J., & Noltemeyer, A. (2008). Viewing Response-to-Intervention through an educational change paradigm: What can we learn? *The California School Psychologist*, *13*(1), 55–66.
Schatschneider, C., Buck, J., Torgesen, J., Wagner, R., Hassler, L., Hecht, S., & Powell-Smith, K. (2004). *A multivariate study of factors that contribute to individual differences in performance on the Florida Comprehensive Reading Assessment Test (FCRR Report No. 5)*. Tallahassee, FL: Florida Center for Reading Research.
Shapiro, E. S., & Clemens, N. H. (2009). A conceptual model for evaluating systems effects of RTI. *Assessment for Effective Intervention*, *35*(1), 3–16.
Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760.
Sulkowski, M. L., & Joyce-Beaulieu, D. K. (2014). School-based service delivery for homeless students: Relevant laws and overcoming access barriers. *American Journal of Orthopsychiatry*, *84*(6), 711–719.
VanDerHeyden, A. M., Witt, J. C., & Gilbertson, D. (2007). A multi-year evaluation of the effects of a Response to Intervention (RTI) model on identification of children for special education. *Journal of School Psychology*, *45*(2), 225–256.
Vaughn, S., Wanzek, J., Murray, C. S., Scammacca, N., Linan-Thompson, S., & Woodruff, A. L. (2009). Response to early reading interventions: Examining higher responders and lower responders. *Exceptional Children*, *75*(2), 165–183.
Vellutino, F. R., Scanlon, D. M., Zhang, H., & Schatschneider, C. (2008). Using response to kindergarten and first grade intervention to identify children at-risk for long-term reading difficulties. *Reading and Writing*, *21*(4), 437–480.
What Works Clearinghouse. (2017). *Procedures handbook, version 4.0*. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_handbook_v4.pdf
Wilson Reading Systems (2002). *Foundations*. Oxford, MA: Wilson Language Training.

How to cite this article: Grapin SL, Waldron N, Joyce-Beaulieu D. Longitudinal effects of RtI implementation on reading achievement outcomes. *Psychol Schs*. 2019;56:242–254. https://doi.org/10.1002/pits.22222