The Incremental Effects of Advanced Phonemic Awareness Training for Struggling Second- and Third Grade Students Receiving Repeated Reading

By

Alexander D. Latham IV

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(School Psychology)

at the

UNIVERSITY OF WISCONSIN – MADISON

2025

Date of final oral examination: 05/06/2024

The dissertation is approved by the following members of the Final Oral Committee:

David A. Klingbeil (Committee Chair), Assistant Professor, Educational Psychology

Craig A. Albers, Associate Professor, Educational Psychology

James E. Pustejovsky, Associate Professor, Educational Psychology

Martha W. Alibali, Professor, Psychology

Dedication

"Dedicated to the homies in the pen; hit me!"

- Kendrick Lamar, 2014

This dissertation is also dedicated to the late Drs. Scott Lilienfeld and Daniel Kahneman. Drs. Lilienfeld and Kahneman both radically transformed me through their writing and advocacy on the critical importance of data, evidence, and sound decision-making in Psychology and in everyday life. I hope this document makes strides towards the same ends.

Acknowledgements

A key takeaway from School Psychology classes is that ‘two things can be true at once’. It is true that grad school was both some of the most fun, fulfilling, and exciting years of my life and some of the most challenging, emotional, and draining years of my life. Yet I still find it difficult to accept that this chapter is almost over, and for the first time I will no longer be a student. I attribute much of this resistance to all the wonderful people who have made my graduate school and life such an amazing experience.

I want to acknowledge Andrew and Katie for your company, support, friendship, and care especially living together in 2020-2021. That was the year I truly made Madison home, thanks in large part to you. I hope to acknowledge Matt, Reed, and John for being exactly the friends I needed. I have been anything but a consistent communicator, and nevertheless you always bring me in and brighten my weekend or evening like we were still living at Breese. To Pastry Club, if I made Madison my real home in 2020-2021, I found my real home community the summer of 2021. I could not have needed such a supportive and hilarious group of friends more, when all my college friends were leaving, and you continue to be such an important part of my life.

I want to acknowledge my cohort, for always being willing to lean on one another, advocate for each other, and commiserate together. More specifically, I must acknowledge Maddy S. for being an incredible lab partner and friend since the very beginning of grad school. It takes a very special person to make WebPlotDigitizer fun. I also must acknowledge the cohort above me, for being invaluable friends by welcoming me into graduate school and including me in so many wonderful memory-making things like frisbee and all the NASP events. I want to acknowledge too the faculty of the School Psychology area and ITP; looking back over my materials, I am so grateful that you somehow read through my verbosity, puppy-dog excitement,

naivete, and borderline hubris to still see potential in me. I will disagree firmly that I deserved all the support you gave but hope sincerely that I've made the most of it since. To my committee members as well – Craig, Martha, and James – you have made my project not only so much better, defensible, and comprehensive than I ever imagined it could be, but also pushed me to grow in areas in which I was sure I had plateaued.

Intentionally placed between my school related and family related acknowledgments, my graduate school would have been very different without my advisor Dave. You accepted me without ever having met me, encouraged me to follow circuitous and far-out research interests, and always knew when to meet me where I was at or when to expect and encourage more from me. I am so lucky to have been your advisee, and even more lucky I didn't have to go to Texas to do it (which somehow might have been worth it).

I also must acknowledge my Mom, Dad, and sister, Eliza. I have cherished and needed the comfort of my mom's perspective, care for family history, and tireless focus on creating special moments, the buoys of my dad's cheerful check-ins, jokes, and tik-toks, and the awe of my sister's growth, timely advice, and opinions. I am leaving graduate school with so much more awareness and appreciation of how you've impacted and shaped me.

Lastly, and anything but least, I am so thankful and lucky that I get to acknowledge my partner, Amanda. Even with the support of the village above, I don't know if I would've been able to do any of this without you. Not only because of your daily support and care, constant ability to make me smile and laugh, inexplicable curiosity for what my niche research means, thoughtful advice and guiding through the hardest things I've faced, and omnipresent talent to make any mundane day (like a sunny Friday afternoon in April) into a warm memory, but more because of who you've inspired me to be. The pretentious, indecisive, delicate, workaholic who

you agreed to date long-distance during a pandemic was in no way deserving of such monumental grace. The sheer awe of that grace, though, encouraged me to grow to become someone who was.

Abstract

Students who struggle to read text fluently (i.e., quickly and accurately) often receive supplemental reading instruction tailored to their specific skill needs. Some researchers contend that advanced phonemic awareness (Advanced PA) – the ability to perceive and manipulate phonemic sounds without referencing written text – is an important skill target for improving text reading. Advocates for Advanced PA training argue that verbal phoneme manipulation skills can facilitate readers' ability to consolidate words to memory which in turn facilitates automatic word recognition. Others conclude that evidence does not exist to support Advanced PA instruction among readers who can access text, for whom print-based instruction will be more valuable. With data from a statewide supplemental reading program, I used an inverse propensity weighting approach to compare reading fluency gains among struggling second- and third-grade readers who received traditional instruction with or without Advanced PA. Outcome regression models included fixed- or random-effects for school site. Results showed that students who received Fluency instruction with Advanced PA experienced significantly lower reading fluency gains (between -8.52 and -15.80 Words Correct Per Minute) than if they had received only Fluency instruction. Survival analyses extend the consequences of these diminished gains; students who received Advanced PA met intervention exit criteria less frequently and graduated intervention after longer periods of time than their peers in Fluency intervention alone.

Table of Contents

List of Tables

List of Figures

Chapter I. Introduction

Between 2019 and 2022, state legislative bodies across the country enacted 81 pieces of legislation aimed at standardizing and improving reading instruction (Neuman et al., 2023). For decades, multidisciplinary research evidence has shown that students often best learn to read from a phonics-based approach to early reading instruction (Castles et al., 2018; Ehri et al., 2001). Phonics-based instruction refers to a type of instruction which teaches students to read words based on the links between their component letters (i.e., graphemes) and corresponding sounds (i.e., phonemes). This contrasts with a whole-word approach to reading instruction in which students learn to read independently and are taught to use context clues, visual cues and guesses to read words (Fountas & Pinnell, 2012; Goodman, 1967). Whole word reading instruction has dominated the landscape of American schooling despite an abundance of research documenting its inefficacy relative to phonics-based instruction (Baumann et al., 1998; Hess, 2023; Rayner et al., 2002).

In light of new legislation for the inclusion of systematic phonics instruction in early literacy, policy makers and district administrators face the new problem of choosing among the many reading curricula marketed as being phonics-based or research-backed (Girard, 2023; Schwartz, 2019). It is important to state clearly that the term research-backed is not synonymous with the term phonics-based. Phonics skills are crucial instructional targets that have been notably underemphasized or absent in popular curricula. Evidence is clear that students need phonics, and equally clear they also need instruction in and exposure to vocabulary, high-quality text, comprehension strategies and more to become expert readers (Castles et al., 2018; Seidenberg, 2017). Some who misunderstand scientific approaches to reading instruction mischaracterize the science of reading as only promoting phonics instruction at the expense of all

other literacy skills (Shanahan, 2020). Scientific approaches to reading instruction more accurately promote the important role of phonics as one of several research-backed components of reading development. These misconceptions add to the confusion educators face when comparing curricula labelled as phonics- or research- based.

Debates within the science of reading community further add to this confusion (e.g., Peng & Goodrich, 2020, Shanahan, 2020). Some debates specifically question choices about effective instructional practices and instructional targets across multi-tiered systems of support (Clemens et al., 2021). Multi-tiered systems of support (MTSS) offer a way to effectively serve students across a range of skill levels to prevent the emergence of intractable skill deficits. In MTSS, all students receive core instruction (Tier 1), a smaller population of students with moderate skill delays or deficits receive supplemental intervention (Tier 2), and even fewer students with more significant deficits receive intensive intervention (Tier 3; Jimerson et al., 2016). The focus of this study pertains to a debate regarding instructional targets for supplemental (Tier 2) reading intervention.

This area of debate surrounds the utility of skills known as advanced phonemic awareness skills (Advanced PA; Clemens et al., 2021). Phonemic awareness (PA) is the ability to perceive the letter sounds that compose spoken words. Researchers have separated PA into two categories of tasks. Advanced PA skills describe the ability to manipulate the sounds of a word by deleting (e.g., say cat without saying /c/), substituting (e.g., say cat but instead of /c/ say /b/), or reversing them (e.g., say /k/i/n/k/i/p/ backwards). These skills extend upon basic PA skills which involve phoneme identification, blending phonemes to create words, and segmenting words into phonemes. Both Basic and Advanced PA tasks are verbal and disconnected from text. Some of the most widely used reading curricula in the country emphasize Advanced PA skills

(e.g., Kilpatrick, 2015; Literacy Resources, 2023), whereas others stop after Basic PA (University of Florida Literacy Institute, 2023). Scholars who promote the value of Advanced PA skills argue that they are crucial for increasing the speed at which readers can read connected text (Kilpatrick & O'Brien, 2019). They contend that Advanced PA skills are particularly valuable for struggling readers who benefit from this additional working knowledge of phonemic awareness (Kilpatrick et al., 2022). This emphasis for struggling readers makes this debate a relevant empirical question for tiered interventions which schools use to support student who need additional support beyond standard classroom instruction.

Recently, critics of Advanced PA skill training reviewed the evidence base supporting their utility. Clemens and colleagues (2021) argued that Advanced PA skills do not benefit readers more than do Basic PA skills, which readers often learn before they are developmentally able to access text. Their review further summarized that evidence cited in support of Advanced PA instruction was either non-experimental or did not directly examine Advanced PA skills. Without evidence for the efficacy of Advanced PA, Clemens and colleagues (2021) concluded that precious instructional time would be better spend focusing on text-based phonics instruction. Other research teams also concluded that phonemic awareness instruction complements early learning of letter sounds and basic words that "unlocks" the text-based instruction from which students grow the most (Ehri, 2014; Perfetti, 2007; Student Achievement Partners, 2020).

Statement of The Problem

Both Clemens and colleagues (2021) and Kilpatrick et al., (2022) agree that, in line with the recommendations for building an evidence base called for in the Every Student Succeeds Act (ESSA; P.L. 114-94), experimental trials should test the role of Advanced PA in supplemental reading intervention. There has only been one experimental investigation of the effects of

isolated Advanced PA instruction (Coyne et al., 2021, as cited in Clemens et al., 2022). This trial examined universal (i.e., Tier 1) Advanced PA skill instruction among first grade students. Results indicated that students performed better on phonemic awareness measures when receiving this treatment, but no distal reading outcomes differed by group. No experimental studies have investigated the efficacy of Advanced PA instruction for older, struggling readers. The present study aims to fill this gap by conducting a rigorous quasi-experimental evaluation of Advanced PA instruction provided to second- and third-grade students who are receiving supplemental reading intervention due to concerns about their reading. In doing so, I evaluated the following research questions:

1. Does Advanced PA instruction cause differences in reading fluency gains among second grade Tier 2 students when compared to text-based fluency intervention?
2. Do second- and third-grade students receiving Tier 2 reading fluency intervention who also receive Advanced PA instruction take different amounts of time to meet intervention exit performance criteria than similar peers who receive only reading fluency focused intervention?
3. Are there subpopulations of students who respond differently to fluency intervention with or without Advanced PA instruction?

Chapter II: Literature Review

Background

Thirty-seven percent of fourth grade students do not meet national standards for proficiency in reading, as of 2022 (U.S. Department of Education, 2023). This percentage has grown since 2015. In Wisconsin, the state in which the author resides, less than half of students in Grades 3 through 8 reach grade level proficiency standards in reading (Klingbeil et al., 2022). Educators must thoughtfully design instruction informed by both reading science and learning science for students with diverse ranges of reading skills. To accomplish the difficult task of simultaneously delivering individualized yet rigorous, systematic instruction, many school districts have adopted Multitiered Systems of Support (MTSS) as a model of academic service delivery (Jimerson et al., 2016; VanDerHeyden & Burns, 2010).

Educators using MTSS differentiate instruction by partitioning instructional practices into three tiers. Tier 1 represents whole-school or whole-class instruction and is typically characterized as the curriculum and instructional practices used by teachers (Stoiber & Gettinger, 2016). In MTSS, all learners participate in Tier 1 instruction, making this the most important and widely influential component of instruction. Tier 1 instruction in reading is intended to grow the reading skills of all readers and in turn must be differentiated enough for the weakest readers to avoid levels of challenge too difficult to learn from (Betts, 1946; Shapiro & Clemens, 2023). More explicit and individualized differentiation, however, comes at the second tier of MTSS.

Tier 2 is designed to offer supplemental instruction and practice for students who are ‘at-risk’ of academic skill problems (Stoiber & Gettinger, 2016; Wanzek et al., 2016). Tier 2 interventions aim to help students who demonstrate some academic skill deficits reach grade

level expectations. Many Tier 2 interventions use the same effective instructional practices used in Tier 1, allowing students extra practice opportunities and additional exposure to direct instruction. In well-designed MTSS, students in Tier 2 receive targeted instruction for 30 minutes three-to-five times per week in addition to Tier 1 instruction and eventually develop sufficient skills to be successful in Tier 1. If Tier 2 interventions are unsuccessful, then students receive even greater levels of support at Tier 3. Tier 3 supports are more robust and resource intensive than Tier 2. These supports are reserved for the smallest subset of students whose skill deficits are more intensive and need more than the moderate level of support provided at Tier 2 (Fuchs et al., 2012).

Sound MTSS structures aim to serve not only as instructional and intervention delivery models, but also as a prevention focused approach to education (Fuchs et al., 2012; Stoiber & Gettinger, 2016). MTSS aims to prevent learning disabilities and other more intensive challenges before they occur by encouraging schools provide additional supports for struggling students before significant skill deficits (1 or more grade levels) become intractable. Of course, learning disabilities are not entirely due to insufficiently intense instruction and some students who participate in MTSS will still be referred for special education. However, the intention is that these students will have already received months or years of increasingly intensive and personalized supports by the time of referral, attenuating at least in part their skill deficits. Evidence also favors using students' lack of response to intervention as a decision-making tool for learning disability identification (Kovaleski et al., 2022; Maki & Adams, 2020). Evidence-based and efficacious MTSS interventions are critically important for ensuring that students can (a) successfully exit additional supports as quickly, and (b) avoid unnecessary referrals to special education (Jimerson et al., 2016; VanDerHeyden & Burns, 2010). Despite the crucial role of

evidence-based instruction across all three MTSS levels, the most popular instructional strategies most have long lagged behind those recommended by the research community.

Recent media attention has spotlighted this deep gap between scientific findings and existing reading instruction practices (e.g., Hanford, 2022, Closson, 2023). Although research consensus formed around the importance of phonics-based approaches to reading instruction some twenty years ago (National Reading Panel, 2000; Rayner et al., 2002), whole-language approaches that minimize instruction of students' letter-sound skills have remained widespread (Castles et al., 2018). Experimentally, phonics-based approaches to reading instruction have outperformed whole-language approaches for increasing young students' reading performance with near universality across settings and contexts (Ehri, 2001; Machin et al., 2016; Rose, 2006; Torgerson et al., 2006; Wyse & Styles, 2007). State legislative bodies have in turn enacted 81 pieces of legislation aimed at standardizing and improving reading instruction between 2019 and 2022 (Neuman et al., 2023). Many of these legislative acts mandate that school districts select from a pool of approved phonics-based programs or curricula.

Phonics-based instruction is not monolithic; however, and valid debates remain lively (e.g., Pearson et al., 2020; Peng & Goodrich, 2020; Shanahan, 2020). Consequently, curriculum vendors differ dramatically in the scope and sequence of their reading programs (Petscher et al., 2020). The heterogeneity of phonics-based approaches creates a new problem for policymakers seeking to revitalize reading instruction: how can educators choose necessary and effective reading programs within the broad category of 'phonics-based' instruction? Many modern approaches to reading instruction fall under the umbrella of 'phonics', though they may lead to wholly different instructional decisions (Petscher et al., 2020). Rather than summarize the considerable number of theoretical models on which these instructional programs are based, I

summarize only a cross-section of models most relevant to the present study. Given that the present study examined intervention effects on students' oral reading of connected text, I will summarize influential models that emphasize word reading processes rather than later developed skills like comprehension. For example, the popular and robustly supported Lexical Quality Hypothesis describes word reading processes but focuses more on predictions of – and solutions for – *understanding* written text (Perfetti & Hart, 2002).

Key Terms in Early Literacy

Phonics-based instruction, in its broadest sense, can describe any method in which instructors *explicitly* teach students the smallest units of sound in the English language called phonemes and their corresponding graphemes (i.e., the smallest units of printed text; Ehri et al., 2001). More specifically, phonics is defined by the National Reading Panel as the skills that involve learning the alphabetic system, and how letter sound-correspondences can be applied to read and spell words (National Reading Panel, 2000). Any English word can be broken down into a sequence of pronounced phonemes (this is what phonetic spelling represents) or graphemes (this is what spelling represents). Most of the 44 phonemes found in English are sounds made by single letters (e.g., the s in sail, the e in bet) or by blending two letters (e.g., the ow in cow, the ph in phoneme; Shapiro & Clemens, 2023).

Phonemic awareness (PA), in contrast, is the specific ability to hear or perceive these small subcomponents of words without referencing their graphemes. PA skills are critically important for early readers (National Reading Panel, 2000). One example of a PA skill is blending the verbal phonemes /c/, /a/, and /t/ to say the word "cat" (Kilpatrick, 2022). For an instructional example, kindergarten students may listen to a teacher verbally present list of letters and respond by calling out each letter's sound. Alternatively, students may call out each letter's

sound when a teacher writes several letters on a board. The former example is phonemic awareness skill practice, while the latter is graphophonemic practice due to the inclusion of the printed letters (i.e., graphemes) that represent the sounds.

Typically, the earliest readers begin their learning careers with Basic PA instruction before transitioning to graphophonemic instruction, at which point they practice identifying phonemes based on their graphemes: the earliest approximation of reading (Ehri, 2020; Kilpatrick, 2022). Phonics instruction often includes little PA training beyond early years (e.g., Kindergarten), opting instead for graphophonemic instruction once readers begin to form strong letter sound correspondences (National Reading Panel, 2000). Although multiple graphemes can refer to the same phoneme (e.g., the s in sail is the same sound as the c in cinder), and the same grapheme can refer to multiple phonemes (e.g., the o in jog sounds different than the o in cone), understanding the general rules that relate graphemes to phonemes is sufficient to read a large percentage of English words (Byrne & Fielding-Barnsely, 1989; Castles, 2009; Ehri, 2020). When students apply this knowledge of grapheme-phoneme correspondence to read a word by sounding out its phonemic units they use the skill of *decoding* – the act of using graphophonemic linkages to read written words (Gough & Tunmer, 1986; Perry et al., 2007). Decoding is among the most critical skills for early readers (e.g., Lonigan et al., 2018; Florit & Cain, 2011; Hoover & Gough, 1990) and I describe this skill and its importance in greater depth when summarizing models of reading development below.

One skill complementary to decoding is that of orthographic mapping (OM). OM is the process by which written words are consolidated to memory to be read instantly from sight without relying on decoding (Ehri, 2005; 2014; 2020). Occasionally, the act of using OM to read is referred to as sight word reading. OM is an important conduit towards reading comprehension.

When students read words instantly, they expend neither time nor effort decoding words by their component sounds and can spend this effort and time comprehending the content of what they read (Ehri, 2005; 2020; LaBerge & Samuels, 1974; Shapiro & Clemens, 2023). In order to read automatically using OM, however, students must progress through a hierarchy of reading skills from letter-sound correspondence to decoding. Students who read words by sight before developing sufficient alphabetic or decoding skills are likely to struggle to read novel words independently (Castles & Nation, 2010; Torgesen, 2002). Many theories and models of reading seek to describe and explain necessary and sufficient skills for successful independent reading.

Simple View of Reading

The Simple-View of Reading (SVR) is among the most longstanding theories of reading development, as it was among the first to challenge the presumption that reading was a simple "psycholinguistic guessing game" (Goodman, 1967, pp. 126). The SVR asserts that reading comprehension is the product of word reading skills and linguistic comprehension (Hoover & Gough, 1990). Word reading most often refers to decoding, which describes students' ability to use their knowledge of letter-sounds (i.e., phonemes) to sound-out words. Updates to the SVR insist, however, that word reading is a broad category that can include skills like orthographic mapping (Duke & Cartwright, 2021; Tunmer & Hoover, 2019). Linguistic comprehension encompasses students' ability to apply their understanding of language to extract meaning from words or phrases. Importantly, linguistic comprehension is an oral and audible skill, not a text-based skill. Text represents only a symbolic representation or encoding of language; therefore, language or linguistic comprehension is dissociated from reading (Gough & Tunmer, 1986; Shapiro & Clemens, 2023). Students can hold excellent language comprehension skills such as listening comprehension without being able to comprehend what they read. When combined with

word reading abilities, however, excellent language comprehension skills will produce the act of reading and comprehension according to the SVR (Hoover & Gough, 1990).

The SVR's simplicity provides excellent communicability. The SVR provides a digestible, friendly, helpful framework for training professionals and stakeholders adjacent to reading science such as teachers, special education staff, and caregivers (Duke & Cartwright, 2021; Kirby & Savage, 2008). The United Kingdom's national reading curriculum is rooted in the SVR, and teachers are trained extensively on the model (Department for Education, 2023). The SVR also highlights the important idea that students do not struggle with reading comprehension. This premise leads to the instructionally critical conclusion that assessing and intervening upon students' precursor skills will more effectively boost their distal reading comprehension abilities (Shapiro & Clemens, 2023; Lonigan et al., 2018). Two students with equally poor reading comprehension skills may benefit from dramatically different supports if one student has high linguistic comprehension and low decoding skills while the other is a strong decoder with minimal linguistic comprehension abilities.

What the SVR lacks in complexity it makes up for in robustness; decoding skills and linguistic comprehension are inarguable precursors to and highly predictive of reading ability (Kendeou et al., 2009b; Lonigan et al., 2018). Linguistic comprehension and decoding skills have been shown to account for 85% or more of variance observed in reading comprehension among a large sample of late elementary students, for example (Lonigan et al., 2018). Other estimates, however, have been more modest (e.g., 72-85%, Hoover & Gough, 1990; < 62%, Tilstra et al., 2009). The robustness of the SVR is further strengthened by its flexibility across ages. The model's original authors describe that the relative contributions of decoding and language comprehension may change over time (Hoover & Tunmer, 2018; 2022). For instance,

younger readers may have little decoding skills to rely on, while more advanced readers with years of decoding practice are unlikely to be limited by these skills. Meta-analytic evidence supports this theoretical position. Decoding more strongly predicts reading comprehension than does language comprehension among early readers, but the opposite is true for advanced readers; reading comprehension outcomes for students with more years of reading instruction are better predicted by their linguistic comprehension abilities (Florit & Cain, 2011; Lonigan et al., 2018). In fact, Lonigan and Burgess (2017) found that decoding skills and reading comprehension skills produce only a single reading factor until third grade, at which point the two skills become distinct and linguistic comprehension begins to limit reading skills. These finding highlights not only the importance of early decoding skills, but also the multiplicative relationship of the SVR. Starting from zero, decoding skills will drive the most impactful changes in early reading ability (i.e., the gains from a multiplicative factor of 1 vs 0, are more consequential than those from 4 vs 3).

The independence of decoding and language comprehension as predictors also changes with age or reading experience. Kendeou et al. (2009a) found that oral language was highly correlated with decoding skills ($r = 0.53$) in a sample of pre-Kindergarten students. However, the two variables were minimally related only one year later, in the same sample of then-Kindergarten students ($r = 0.17$). In a separate sample, the relationship dropped further by second grade ($r = 0.11$). Despite the decreasing collinearity of the predictors, oral language skills and word identification both remained highly predictive of reading comprehension in Grade 2. Contrarily, factor analysis of elementary students' reading performance shows that more variance in reading comprehension outcomes is explained by shared variance between decoding

and linguistic comprehension factors than by each factor's unique variance (Lonigan et al., 2018).

The Less Simple Views of Reading

The robustness, influence, and utility of the SVR cannot be understated. Even so, its authors and critics alike agree that the view is incomplete, and that more ingredients contribute to reading outcomes than just decoding and linguistic comprehension. Rather than start from scratch, updated and expanded models have been proposed by the authors (Tunmer & Hoover, 2019) and independent research teams (e.g., Duke & Cartwright, 2021). The cognitive foundations framework expands upon the original SVR by further subdividing the two skills contributing to reading comprehension. As shown in Figure 1, the cognitive foundations framework authors specified a hierarchy of cognitive skills necessary to produce the acts of word reading and linguistic comprehension. This update acknowledged the complexity and multifaceted nature of each reading subskill underscored by intervening years of research (e.g., Aaron et al., 2008; Kim, 2017; Seidenberg, 2017). The cognitive foundations framework, like its inspiring SVR, outlines increasingly specific and hierarchical assessment and instructional targets. Although the authors are clear to note that the hierarchy of skills is not linear, instructors can further decompose students' difficulties in word reading to perhaps a strength in letter knowledge but a relative weakness in phonemic awareness. To this end, the cognitive foundations framework seldom departs from the SVR except to guard against critiques related to oversimplification of the reading process.

The Active View of Reading differs substantially in the scope of its updates to the SVR. The Active View of Reading offers predictions and mechanisms as opposed to simply accounts and frameworks (Duke & Cartwright, 2021). Beyond offering more predictability and empirical

testability, the Active View proposes additional factors that contribute to reading skill development. Namely, the Active View distinguishes between the unique and shared variance of word reading and linguistic comprehension by outlining skills which are unique to each category and shared between the two categories. For instance, phonics knowledge is a skill strictly related to word reading, verbal reasoning is solely descriptive of linguistic comprehension, but vocabulary knowledge is related to both word reading and linguistic comprehension. The Active View also posits that cognitive processes such as executive function, self-regulation, or motivation drive activation of reading specific processes. The model, reproduced in Figure 2, begins with cognitive processes related to active self-regulation and motivation, which drive students' engagement with reading activities and stimuli as well as their subsequent activation of reading specific cognitive processes. From there, the model proceeds through each of the three reading subskill categories (word reading, linguistic comprehension, or skills shared by both) towards the outcome of incorporated reading. The original authors of the SVR quickly criticized the Active View for its lack of empirical evidence as well as its de-emphasis of key reading skills such as listening comprehension (Hoover & Tunmer, 2022). In the years since the Active View was first proposed, a component meta-analysis has preliminarily supported the correlative links among pathways within the model (Burns et al., 2023). In addition, two experimental studies have successfully demonstrated causal links between cognitive flexibility and reading comprehension (Hund et al., 2023) and that semantic verbal fluency mediates the relationship between achievement goals and reading comprehension (Cho et al., 2023).

The Direct and Indirect Effects model of Reading (DIER; Kim, 2020a) is another newer theoretical model that, like the Active View, includes executive function and general cognitive processes as upstream determinants of reading (Kim, 2020b). The DIER differs from the Active

View however in several key ways. First, the hierarchical skill level immediately after cognitive and executive function skills mirrors the skills emphasized by the SVR: word reading and listening comprehension (Hoover & Gough, 1990). Kim (2020b) expanded upon the SVR by specifying subskills within these two categories. Under the umbrella of word reading, DIER highlights the classical cognitive psychological triangle of orthography, phonology, and morphology (Plaut et al., 1996; Seidenberg & McClelland, 1989) as underlying the skill of word reading. Orthography and phonology will be familiar to the reader; phonology refers to component sounds of words and orthography refers to component letters that represent the sounds (i.e., graphemes and phonemes; Ehri, 2020). Morphology refers to small, but meaningful units within words. Morphemes can take the form of suffixes, prefixes, or roots of words and serve as an intermediary unit between graphemes or phonemes and entire words.

Under listening comprehension, DIER places foundational oral language skills beneath higher order cognitive and regulatory skills such as reasoning and perspective taking. These higher order skills are absent from the SVR, Active View, and cognitive foundations framework (Duke & Cartwright, 2021; Hoover & Gough, 1990; Tunmer & Hoover, 2019). Uniquely, text reading fluency, the act of reading efficiently, expressively, and accurately, connects students' word reading and linguistic comprehension skills to their reading comprehension.

Text reading fluency is a robustly predictive measure of many reading and other academic outcomes widely used in schools as a benchmark measure for assessing academic proficiency (Kilgus et al., 2014; Kim et al., 2014; 2021). The research on text reading fluency is so broad, its absence as either a predictor or outcome from other models is notable. Text reading fluency falls under the category of word reading in the original Simple View of Reading, however, later studies found that the skill was unique enough to merit its own distinction (Wolf

& Bowers, 1999). Research has noted that the more fluently students can read, the more cognitive resources they can devote to cognitive skills crucial to meaning making and comprehension such as reasoning or connecting to background knowledge (LaBerge & Samuels, 1974; Young & Rasinski, 2009). In the highest level above the intermediary text reading fluency, DIER places the outcome of reading comprehension alongside other factors influencing reading outcomes directly including background knowledge and social emotional characteristics like motivation, self-beliefs, and attitudes towards learning (Kim 2020a; 2020b).

What sets DIER apart from other models is not only the addition of additional and specific skills and precursors, but the path of relationships throughout the model. As opposed to proceeding linearly from subskill to outcome, DIER proposes that all skills and outcomes are dynamically and hierarchically related. Increasing mastery of lower-level skills (e.g., orthographic knowledge) improves performance on higher-level skills (e.g., text reading fluency) which then reciprocally influences other factors such as increasing self-efficacy and motivation to read, absorbing novel background content, and practicing vocabulary. Tunmer and Hoover (2019) discussed similar Matthew effects when updating the SVR to the cognitive foundations framework but describe them strictly as a cyclical process in which practice opportunities strengthen skills and beget further practice opportunities. However, the reciprocal and dynamic interrelations among skills proposed in the DIER are not strictly cyclical but instead resemble a spreading activation model of performance. As such, DIER can acknowledge correlations between word reading and listening comprehension without undermining their independence, for example. These mutually upregulating or downregulating systems also highlight the heterogeneity among early readers that models like the SVR failed to capture. It can be true that word reading and listening comprehension skills highly predict reading comprehension, and that

students with strong skills will seldom perform up to their predicted level in the absence of sufficient attentional control, for example (Burns et al., 2023; Kim, 2020b).

Evidence supports the unique interrelations put forth by DIER. Students assessed on a battery of reading and cognitive skills in second and fourth grade showed several pathways through which reading subskills were related to reading outcomes. Using structural equation modeling, Kim (2020b) found that word reading and listening comprehension directly predicted reading comprehension, while other cognitive skills such as attention and working memory affected reading mostly indirectly, through either listening comprehension, word reading or both. Interestingly, when phonological, orthographic, and morphological skills were not assessed, the remaining cognitive and linguistic precursor skills explained only small proportion of word reading variance. Word reading, clearly, is much more an independent function of its subskills including phonemic awareness. Further, the relative contributions of word reading and linguistic comprehension changed over time, as had been observed in prior studies of the SVR (e.g., Florit & Cain, 2011; Lonigan et al., 2018). Word reading skills were less related to reading outcomes among fourth grade students than among second grade students. This finding is instructionally relevant, such that advanced readers who demonstrate reading difficulties may be better served by interventions targeting linguistic comprehension or its subcomponents.

Models of Reading Skill Development

The SVR, Active View, DIER, and cognitive foundations framework models all offer accounts of reading that are ingredient-based as opposed to developmental in nature. In essence, the outcome of reading is determined by the simultaneous relative values of all precursors. These models are useful in explaining and assessing observable reading performance. Other accounts of reading performance are developmental and longitudinal in nature. One influential example is

Ehri's Four Phases model (Ehri, 2005). Ehri's model builds on prior developmental models which outline many of the same skills influential to models described above but describe their origins and influences as developmental rather than componential in nature. Development models preceding the Four Phases model frequently delineated developmental stages based on distinct actions, behaviors, or skills common to each stage. For instance, Marsh and colleagues (1981) proposed that students use different forms of guessing when beginning to read: linguistic or discrimination guessing. A sharp transition divides the next stage in which students systematically apply decoding skills to read words.

Ehri's (2005) Four Phases model blurs these discontinuous transitions by positing phases that exist along a spectrum. That is, phases are determined by the degree to which readers demonstrate understanding of the alphabetic system, rather than some categorical groups of skills employed. The earliest phase is the pre-alphabetic phase. The degree to which readers understand the alphabetic system is *minimal* in this phase. Written letters represent little more than visual symbols at this phase and in turn students rely predominantly on visual cues to "read" words. Visual cues may be salient first letters, environmental associations (e.g., the McDonalds' M), or even irrelevant paired stimuli (Ehri, 2005; 2020). Students at this phase who were taught to read a simple word that was displayed alongside a thumbprint, vocalized the word they were taught when they saw the thumbprint alone or next to a different word (Gough et al., 1992). This holistic approach to reading words means that because students do not conceptualize letters as components, they lack the ability to form letter-sound relationships in this phase.

The first transition from the pre-alphabetic to partial-alphabetic phases, then, is marked by a difference of degrees; students flow into the partial-alphabetic phase as they acquire *some* degree of understanding of the alphabetic system. Some of the earliest skills acquired on

students' path to understanding the alphabetic system are letter naming and letter sound identification. Without yet developing a full degree of understanding, students still rely partially on visual cues. In this case, a common cue is the first letter of a word. Errors are still common, but errors that misidentify the first letter of a word are less so. This trend led to the development of a common measure of early reading known as first sound fluency (Cummings et al., 2011) in which readers identify the first sound of each word in a list. Despite this inordinate emphasis on initial sounds, students that know more letter sounds at this stage can start to identify rudimentary combinations of letter sounds; letters do mean more to students than mere visual symbols at this point. In a classic example, students at this stage more readily accept and learn to read strings of letters that phonetically represent, simplistically, a target pronunciation. When the pronounced word 'giraffe' was taught to be spelled as JRF or wBc, students at this phase made fewer mistakes and demonstrated faster learning for the JRF spelling. Crucially, students with *pre-alphabetic* knowledge more easily learned the wBc spelling, due to its high degree of visually salient uniqueness (Ehri & Wilce, 1985; Roberts, 2003).

It is easy to mistake these partial-alphabetic skills of mapping letters to their corresponding sounds for decoding abilities. However, partial-alphabetic readers stumble over the similarities among written words. Students who attend eagerly to beginning and final letters are likely to succeed at reading a word such as 'deem', but also say 'deem' when reading words like 'dream', 'drum', or 'dim' (Ehri, 2020). In the partial alphabetic stage, readers probabilistically guess that the grapheme 'd' takes on its most common phoneme 'dee'. The nuance required to understand that combining phonemes modifies their pronunciations does not develop until the full-alphabetic phase. At this phase, readers are able to accurately *decode* components of words above and beyond their initial and final letters. Full-alphabetic readers

better recall the middle letters used to spell words than do partial-alphabetic readers (Ehri & Wilce, 1987). Furthermore, they can distinguish between increasingly similar words with granularity. Although a partial-alphabetic reader may confuse deem and dream, a full-alphabetic reader will succeed in blending the ‘d-‘ and ‘r’ phonemes to decode the correct word. Two main challenges face students throughout this phase of largely successful reading. First, decoding is not a rapid enough process to facilitate fluent sentence or passage reading. Decoding an individual word takes significantly more time and effort than retrieving it from memory (Samuels, 1976; Stanovich, 1981). Therefore, readers at some point must begin to store words in their memory to be identified much more immediately by sight (Ehri, 2005). This need outlines why higher order cognitive skills like rapid automatic naming and working memory are considered by many other models and scholars as important precursors to reading (Denckla & Rudel, 1976; Landerl et al., 2019). The second main challenge, relatedly, is that many English words do not adhere to the phonemic decoding principles that guide so much early reading growth (Frost et al., 1987). This is another problem that can only be solved by incorporating these words into memory, a key focus for much vocabulary and explicit sight word instruction (Ehri, 1997).

A bridge exists, though, between decoding and more fluent reading. Many English words need not be decomposed into its strictly phonemic components but rather into recurring and meaningful blocks of phonemes called morphemes (Carlisle & Stone, 2005; Ehri, 2005). Although still decoding effortfully, students can decode more quickly when dividing words into commonly read syllables like -ike, -ing, -ully, -ents, and -ly (Ehri, 2020). All of these morphemes can be found in the prior sentence, to give a sense of their ubiquity. It is this progression into decoding-by-chunks that defines the final consolidated-alphabetic phase.

Readers have already mastered phonemic awareness skills at this phase and therefore gain little additional new knowledge about letters or phonics henceforth. Rather, the biggest gains that readers achieve throughout this stage come from increasing the speed at which they can decode and read words. Decoding by morphemes is one way to increase this speed (Bowey & Hansen, 1994; Ehri, 2020). However necessary, decoding is not sufficient to facilitate passage reading, even in the original SVR (Hoover & Gough, 1990).

Ehri's Four Phases model is a theory of sight word reading (Ehri, 2005). Learning to read sight words through the developmental progression posited above is a crucial step towards increasing the automaticity and immediacy with which readers can *recognize* these words. Word *recognition* is dangerous during early reading, when readers are relying on visual cues unrelated to graphophonemic linkages. Students trained to recognize words before they have sufficient phonemic awareness are likely to demonstrate low abilities to decode novel words correctly (recall the deem/dream example; Metsala & Ehri, 2013; Skinner & Daly, 2010). However, word recognition that follows only after the development of high levels of phonemic awareness is a crucial skill that allows readers to rely predominantly on recognition for effortless, rapid reading and pause only when necessary to decode novel or challenging words (LaBerge & Samuels, 1974; Ehri, 2020). This consolidation of familiar words to memory is called orthographic mapping (Ehri, 2014). When readers view a word that they have read before and instantly read it from memory without (a) using decoding skills, (b) confusing it with other similar words, or (c) expending effort, they have previously orthographically mapped that word into their memory (Metsala & Ehri, 2013). Orthographic mapping is the advanced form and the goal of graphophonemic linking; mapping occurs only when students have sufficient knowledge of the components of words (graphemes and phonemes) to store the spellings effortless as

pronunciations. Despite serving as a goal, orthographic mapping occurs throughout reading development. In early literacy, readers map simple letters to basic phonemes, and construct an ever-increasing repertoire of instantly available phonemes, syllables, and pronunciations activated when viewing graphemes, morphemes, and words.

Determining the Instructional Focus

The science of reading instruction intersects importantly with learning science. Throughout periods of learning, students' growth can be accelerated by targeted instruction that emphasizes certain subskills (e.g., letter sound correspondence or morphological decoding) or that emphasizes certain instructional formats (e.g., immediate error correction or delayed error correction; Daly et al., 1996). These instructional modifications show temporal effects across a range of skills; specific subskills and instructional formats are more effective at different points in the learning process (e.g., Maki et al., 2021; Szadokierski et al., 2017).

The Instructional Hierarchy (IH; Haring & Eaton, 1978) provides a useful heuristic for describing temporal effects observed in the efficacy of specific instructional emphases. This heuristic is useful for its linkage of assessment to intervention. Information about a students' level of performance can inform decisions about likely effective instructional strategies. The IH, reproduced visually in Figure 3 below, outlines four broad stages through which students proceed when learning a new skill. The earliest of these stages is the acquisition stage. Students in this stage demonstrate a goal skill slowly, and inaccurately. For a relevant, running example, learners in the acquisition stage of passage reading will read text with pauses and effortfully decoded words, identifying more than 10% of words incorrectly (Parker & Burns, 2014; Shapiro & Clemens, 2023; VanDerHeyden & Burns, 2010). Students benefit the most from expert

modeling (e.g., listening to advanced readers), explicit instruction (e.g., teacher directed learning), and immediate error correction throughout this stage of the IH.

When students begin to demonstrate a skill with greater accuracy, but still slowly, they have entered the proficiency or fluency stage. The level of accuracy necessary to be confidently into the fluency stage varies by the type of skill. Eighty percent accuracy is sufficient for students to be identified as in the proficiency stage for discrete, problem-solving skills like math-computation (VanDerHeyden & Burns, 2010). However, for more continuous skills like passage reading, consensus recommendations fall between 90% and 93% (Parker & Burns, 2014). Within the fluency stage, students benefit from independent practice and extended practice with new material. Importantly, since the goal of this stage is to increase the speed at which students can demonstrate skills accurately, students benefit more from delayed error correction that does not interrupt their practice. The skills that students practice in this stage are more related to speed and fluency themselves than to the skill, and therefore feedback should focus less on accuracy but on speed of responses. Continuing the example of passage reading, students in the fluency stage of passage reading should be encouraged to read independently and, when observed, should be praised only at the end of a passage or book. Instructors can deliver corrective feedback at that point as well, encouraging students to re-read words that were pronounced incorrectly.

One example of a popular reading intervention for students in the fluency stage is the Helping Early Literacy with Practice Strategies (HELPS; Begeny et al., 2010) program. In this program, students experiencing delays in reading fluency read an appropriately matched passage aloud. After this, an instructor reads the passage aloud before the student reads the story in its entirety for the second time. After this second read, the instructor picks three to five sentences in

which the student made an error or read slowly and encourages the student to read each sentence aloud three to five times providing any error correction at the end of each sentence. After this practice, students read the full passage for a third and final time and are praised for their speed.

After students can comfortably demonstrate a target skill quickly and accurately, they enter the generalization stage at which point they can continue to demonstrate their skills quickly and accurately on new and increasingly difficult stimuli (Daly et al, 1996). Students at this stage still benefit from independent practice, but also from discrimination training in which they learn more granular rules, and applied practice in which they complete problems in real world contexts (e.g., math word problems). Regarding passage reading, generalization stage readers can reinforce their skills by reading passages of gradually increasing length and complexity, learning new vocabulary along the way (Haring & Eaton, 1978; Skinner & Daly, 2010).

The final stage is the Adaptation stage. In the adaptation stage, students learn how and when to break the guiding rules that shaped their early instruction. Students at this stage also can devote the precious cognitive resources previously depleted by effortful skill demonstration towards learning from and problem solving with their new skill. Readers at this adaptation stage can increasingly learn from what they read and answer complex questions about inferences from the text. Careful readers will note similarities between this stage of the IH and the consolidated alphabetic phase of Ehri's Four Phases model of sight word reading (Daly et al., 1996; Ehri, 2005). Readers in the adaptation stage, like those in the consolidated-alphabetic phase, may learn new rules for breaking apart words and decrease their reliance on laborious and deliberate tools like decoding in favor of automatic sight word recognition.

A crucial extension of the IH is the recognition that the act of reading comprises many discrete skills each with its own path towards skill development. Early readers, then, may exhibit

a profile of skills simultaneously distributed throughout the stages of the IH. For instance, students may possess letter sound skills at the generalization stage but decoding skills in the acquisition stage (Daly et al., 2005). In fact, due to the sequential nature of reading subskills, it is likely that when students enter the fluency stage for a given subskill, they begin to progress through the acquisition phase for the subsequent skill. The DIER makes similar predictions about the sequential and hierarchical nature of reading skills. As readers develop mastery of early skills like letter-sound correspondence, they dramatically increase their access of more complex stimuli to begin to learn higher order skills like orthographic mapping from (Kim, 2020a; 2020b).

The models of reading and instructional sequencing described above make the same prediction that reading instruction should be appropriately sequenced to focus on relevant challenges that limit the ceiling of blossoming readers' abilities (Burns, 2007; Ehri, 2020; Lonigan et al., 2018; Shapiro & Clemens, 2023; VanDerHeyden & Burns, 2010). Instruction that is too challenging (Frustrational), perhaps targeting skills that are too advanced, would interfere with students' ability to learn from their experiences as the gaps between their already mastered skills and the frustrational skills would be too large to sustain and maintain. Instruction that focuses on skills already mastered by students is unlikely to promote reading skill gains because these skills that students can demonstrate with ease are not the skills limiting their reading performance (Burns, 2007; Daly et al., 2005). Careful attention to instructional sequencing is important as schools have limited resources to provide time, staff, and space for delivering tiered interventions in MTSS (Freeman et al., 2015). Instruction that is inappropriately sequenced, much like interventions that are not evidence based, inappropriately consumes precious school resources without benefitting students. Students also experience consequences. Tier 2 or 3

instruction that is poorly sequenced, mis-targeted, or non-evidence-based will be unlikely to prevent undue referrals to special education or the formation of large skill gaps (Hudson & McKenzie, 2016; VanDerHeyden et al., 2014).

Competing Instructional Targets

As previously mentioned, disagreements still exist within the body of phonics research (e.g., Pearson et al., 2020; Peng & Goodrich, 2020; Shanahan, 2020). These disagreements interfere with practitioners' ability to be confident that they are providing evidence-based and appropriately sequenced instruction. One disagreement that particularly confounds instructional sequencing is the role of advanced phonemic awareness (Advanced PA) skills. Recall that Advanced PA skills refer to the class of verbal and audible skills involved in manipulating, changing, subtracting, or reversing phonemes within words. Kilpatrick (2015; 2020b) argues that Advanced PA skills are likely to benefit struggling readers across elementary grades. Though strong evidence exists for the use of basic PA skills (phoneme blending or segmenting) among Kindergarten and first grade students (National Reading Panel, 2000), generalization to older students is largely theoretical. Kilpatrick draws on Ehri's (2005; 2020) model and its emphasis on letter sound awareness as a precursor to the consolidation of sight words to memory to infer that continuous improvement in phoneme analysis will yield continuous improvement in automatic reading (Kilpatrick et al., 2022; Kilpatrick & O'Brien, 2019). Specifically, Kilpatrick theorizes that advanced phonemic awareness skills like substitution or elision substantially increase readers' phonemic and morphemic proficiency, enabling them to represent larger sound or word units more easily in memory (Kilpatrick, 2020b). A mechanistic diagram outlining this proposition can be found in Figure 4. Research indicating that Basic PA skills accelerate gains in reading interventions for struggling readers also offer some support for this hypothesis

(Kilpatrick & O'Brien, 2019; Torgesen et al., 2001; Truch, 1994). In line with this theory, two widespread Advanced PA intervention programs have been popularized throughout the country (Kilpatrick, 2015; Literacy Resources, 2023).

Beyond representing words more easily, Kilpatrick claims that orthographic mapping *requires* advanced phonemic awareness skills (Kilpatrick, 2015). Kilpatrick frequently cites this concept of orthographic mapping (Ehri, 2005) as justification for the necessity of Advanced PA skill instruction. However, where orthographic mapping refers to the connection between phonemes and their graphemes, Advanced PA instruction emphasizes only phonemes with no text. One pioneer of orthographic mapping research highlights this disconnect directly when speaking about Advanced PA skills. Ehri clarifies that orthographic mapping revolves around *graphophonemic* connections that are formed and strengthened as students bond word spellings to pronunciations in memory (Parker, 2022). In describing her Four Phases model, Ehri goes on to note that her theory of alphabetic development refers more closely to the alphabetic connections (i.e., visual cues) students use. Indeed, the key difference between pre-alphabetic and partial-alphabetic readers was their choice of visual, not auditory, cues (Ehri, 2005; Gough et al., 1992). Ehri's theory also progresses from letter sounds to consolidated syllabic sounds. This contrasts with Kilpatrick's instructional guidance which encourages readers to begin with syllabic units before progressing to smaller phonemes (Kilpatrick, 2020a).

Is Advanced PA Training Necessary?

Advanced PA instruction was further thrust into debate by a recent critical review (Clemens et al., 2021). Clemens and colleagues (2021) narratively reviewed the evidence supporting Kilpatrick's (2015; 2019) theory and the popularity of Advanced PA instruction more broadly (Literacy Resources, 2023). Importantly, the authors distinguish between arguments for

Advanced PA throughout ones' reading development and the indisputable value of phonemic awareness for early reading and transitioning from letter knowledge to decoding ability. Several important challenges to the utility of Advanced PA instruction beyond as early literacy precursors are raised.

Dissociations between Advanced PA skills and skilled reading. Kilpatrick's central argument is predicated on the premise that Advanced PA skills will increase distal reading outcome measures (Kilpatrick & O'Brien, 2019). To accept this premise, there must be a relationship between these Advanced PA skills and reading performance. Clemens and colleagues (2021) provide evidence of the absence of this relationship across several contexts. In one study of students who scored high on measures of decoding, many scored low on measures of Advanced PA (Byrne & Fielding-Barnsley, 1993) while another study showed 31% of successful readers spanning multiple ages failed to surpass 50% on a measure of Advanced PA (Scarborough et al., 1998). This study summarized by Clemens et al., also demonstrated the inverse of Kilpatrick's claim; once students reach a certain level of reading performance, their Advanced PA skills stagnate (Scarborough et al., 1998). This inverse relationship suggests at least one viably alternative pathway to successful consolidated reading other than through Advanced PA skills. Conciliatorily, basic phonemic awareness skills (e.g., blending, segmenting) robustly demonstrate large effect sizes for reading outcomes across meta-analyses, however, the Advanced PA skills demonstrate effect sizes less than half as large.

Dissociations between poor Advanced PA skills and reading difficulties. The argument that Advanced PA skills improve skilled reading does not necessarily entail that poor Advanced PA skills will adversely affect reading. However, the proposition that Advanced PA skills represent a distinct and performance enhancing component of reading performance does

include the assumption that the absence of Advanced PA skills should be associated with attenuated reading outcomes. If this were not the case, Advanced PA skills could not be distinct from other, potentially unobserved skills that would better describe differences in reading outcomes. This is not the case, according to the research reviewed by Clemens and colleagues (2021). Though students with poorer reading performance also demonstrated poorer phonemic awareness skills, there were no differences in the type (e.g., basic vs. advanced) of phonemic awareness subskill (Melby-Lervag et al., 2012; Snowling & Melby-Lervag, 2016). This link was also not causal. Therefore, Advanced PA skills cannot be said to affect reading above and beyond basic PA skills, and evidence of the direction of this observed relationship is not clear.

Limited evidence that Advanced PA skills generalize to distal outcomes. A primary critique offered by Clemens and colleagues (2021) is that little evidence exists to directly support the use of Advanced PA. Further experimental evidence must be collected prior to implementing interventions. Evidence that has been cited in support of Advanced PA, according to Clemens et al., is in fact inappropriately cited. The research in Kilpatrick & O'Brien's (2019) review, for instance, is largely nonexperimental or designed without the intention of testing the causal effects of Advanced PA. Experimental research cited in support of Advanced PA has included Advanced PA training as one component of a multicomponent intervention package (Clemens et al., 2021). The only experimental trial to evaluate isolated Advanced PA instruction without text components found that although intervention students made significantly higher gains on proximal measures of phonemic awareness skills, no inter-group differences emerged for reading outcomes (Coyne et al., 2021). This crucial finding calls the direct relationship between Advanced PA and advanced reading skills into question.

Evidence favors the use of text once readers can access it. Clemens and colleagues (2021) are careful to clarify that phonemic awareness practice, as outlined in both empirical studies and most models of reading development, is crucial during the early stages of reading development (Castles et al., 2009; Ehri, 2005; 2020; Shapiro & Clemens, 2023). However, once early readers begin to display decoding and reading abilities, their reading growth rate increases dramatically (Ehri, 2005), and this increase is strengthened from repeated opportunities to practice with reading. This is consistent with Ehri's earlier emphasis on the role of graphophonemic, not just phonemic, connections (Parker, 2022). Phonemic awareness instruction can still occur in conjunction with graphemic instruction, but Clemens and colleagues conclude that Advanced PA without text is unnecessary.

Clemens and colleagues (2021) summarize two key takeaways from their review of the Advanced PA literature. First, without clear evidence of the efficacy of Advanced PA instruction, students' time would be better spent engaging in instructional practices with firm evidence bases tailored to their unique skill deficits. Resources necessary for providing reading interventions at school – like time and staff – are precious and should not be expended without evidence that the expenditure will produce benefits. Second, the researchers diplomatically remind readers that the 'absence of evidence' for Advanced PA does not mean that there is 'evidence of absence [of efficacy]' for Advanced PA. All evidence-based interventions were at one point strictly theory-based ideas. Therefore, further research should directly examine the effect of Advanced PA instruction so that practitioners can make decisions with the best quality of evidence.

A group of researchers in favor of the use of Advanced PA intervention produced a response to the Clemens et al, 2021) critique (Kilpatrick et al., 2022). This response argues that

Clemens and colleagues misrepresent citations and misunderstand central concepts in crafting their critique of PA. Among other procedural critiques, Kilpatrick and colleagues focus on the definition and role of orthographic mapping. This response paper claims that Clemens and colleagues (2021) conflated the definitions of orthographic mapping and decoding. Although the response paper draws a sharp distinction between the two concepts by arguing that PA skills undergird orthographic mapping but are unrelated to decoding, prior research does not draw such a distinction. In fact, models of reading describe the processes as complementary and sequential; as decoding fluency improves, orthographic mapping is facilitated (Kim, 2020a; 2020b; Metsala & Ehri, 2013). Kilpatrick and colleagues (2022) also contend that the Clemens et al. (2021) paper is misguided in its focus on typically developing readers. According to the response paper, advanced PA skill instruction is most appropriate for struggling readers less likely to develop orthographic mapping skills naturally.

The response paper goes on to contend that this instruction is still valuable for struggling readers after first grade, and may help good readers too, "though not always necessary" (pp. 3). To make this claim, the authors describe four methods of making causal inferences that do not involve conducting controlled experiments. Though they do not provide a citation for these alternative pathways to causal inference, they implore the readers to search 'John Stuart Mill's causal inference'. Millsian causal inference does provide the logical bases for rigorous experimental evaluations of interventions. A popular textbook on statistical inference cites Mill's theories (Shadish et al., 2002). However, Kilpatrick and colleagues use Mill's original logic to argue that evidence need not have statistical analysis nor control groups to constitute sufficient causal evidence. This generous definition is inconsistent with modern understandings of causal

inference and policies defining what constitutes evidence in education (Imbens & Rubin, 2015; Shadish et al., 2002; What Works Clearinghouse, 2022).

For instance, they compare two groups of two studies of reading intervention efficacy. One group of studies included phonemic awareness training while the other did not. According to the authors, "The only systematic difference between the two studies is the intensive phonemic awareness [sic]" (pp. 29). Kilpatrick et al., (2022) conclude that because the two studies including PA training found substantially larger reading gains than the two studies lacking PA instruction, Advanced PA instruction must be related to larger reading gains. However, this assertion ignores tremendous differences among the studies. The studies including Advanced PA training were published 10-20 years prior to the studies without Advanced PA. The population of students, as well as the rigor of statistical analysis, has changed dramatically over that time. Further, all four studies examine entirely different populations. One study includes only early elementary students with severe reading disabilities (Torgesen et al., 2001), another included middle school readers across skill levels (Vaughn & Fletcher, 2012), and another yet included participants ranging from school age to adulthood (Truch, 1994). This is not to mention differences in (a) intervention components, (b) comparison conditions, (c) outcome measures, and (d) settings (Amir-Behghadami & Janati, 2020).

Despite the questionable reasoning used to dispute Clemens and colleagues' (2021) review, the two groups agree on one issue. Kilpatrick and colleagues concede that little empirical evidence exists, currently, to support the use of PA instruction for good readers after first grade. The authors praise Clemens and colleagues' for noting that the use of Advanced PA instruction in practice for older students is currently ahead of the research base. While Kilpatrick et al., (2022) argue that comparing certain components of prior studies is sufficient for concluding

causal inference, Clemens and colleagues (2021) call for further empirical research to be conducted. The present study aims to fill this gap in the literature by conducting an empirical investigation of the relative value of Advanced PA instruction among struggling readers above first grade.

Purpose of The Present Study

The purpose of this study is to investigate the incremental value of Advanced PA training for struggling readers in second and third grades. Second and third grade students who are "at-risk" (i.e., Tier 2) for reading concerns likely have sufficient early literacy skills to read text but struggle to read quickly or accurately (Stoiber & Gettinger, 2016; Wanzek et al., 2016). Importantly, students warranting Tier 2 intervention should not be confused with students who demonstrate more intensive needs greater than one year behind expectations (i.e., Tier 3) or those who demonstrate characteristics of specific learning disabilities. Second and third grade students with more intensive needs may struggle to access text, and phonemic awareness training would benefit them as much as it would an early kindergarten reader who could not access text (Shapiro & Clemens, 2023). In contrast, we can examine the incremental value of text-based or sound-based instruction among students demonstrating *some* risk in second and third grade who can access text. Among this group of students, Clemens and colleagues (2021) predict that text-based instruction would be sufficient with no additional benefit of Advanced PA instruction. Kilpatrick and colleagues' (2022) argument suggests the opposite prediction.

In this study, I use data from the Reading Corps program. Beginning in Minnesota, Reading Corps has now expanded to 13 states and D.C. Reading Corps partners with schools to provide targeted reading interventions for second and third grade students demonstrating the

need for additional reading support. In the following section, I describe the Reading Corps program and the general intervention procedures.

Overview of Reading Corps

Reading Corps is an Americorps service branch that provides high-quality literacy intervention in schools for students in Kindergarten through third grade. Reading Corps tutors are Americorps members trained specifically to deliver literacy interventions and dramatically expand school service capacity by providing valuable instruction for large numbers of at-risk students without depleting existing school resources (Markovitz et al., 2022). Tutors are supported by a site-based literacy coach and a Reading Corps coaching specialist. The site-based literacy coach provides daily, individualized support and supervision for tutors and liaises with school staff to coordinate and complement Tier 2 intervention with Tier 1 classroom instruction. The literacy coach and the coaching specialist offer specific support around data-driven decisions, helping tutors select the most indicated interventions for students based on their screening or progress monitoring data.

Students are identified for participation in Reading Corps tutoring through a universal screening procedure. This procedure screens all students from Kindergarten through Grade 3 using grade-level appropriate screening tools. Because the present study focuses on Grades 2 and 3, only the screening procedure for these grades is described. All students in these two grades complete a curriculum-based measure of reading (CBM-R; Christ et al., 2018) probe in the fall. Students who score as “at-risk” on this measure are offered no-cost, supplemental reading tutoring to be delivered during school by Reading Corps tutors. The “at-risk” category represented a range of scores in which students demonstrated reading skills that, though behind grade-level expectations, are not warranting significant, intensive support. Adopting MTSS

language, these students can be considered those warranted “Tier 2” levels of support based on their oral reading fluency (Jimerson et al., 2016). Reading Corps tutors were trained to deliver 10 different interventions each directed at one or more component reading skills: phonological awareness, phonics, fluency, vocabulary, and comprehension (Markovitz et al., 2022).

Intervention Procedure

Students in Reading Corps receive 20 minutes of 1:1, or in some cases 1:2, targeted reading intervention each school day (i.e., 100 mins/week). Tutors held a variety of personal and professional backgrounds (Hammerschmidt-Snidarich et al., 2021) but were uniformly trained in both intervention assignment and delivery by Reading Corps. Reading Corps training materials provide detailed guidelines for assigning intervention based on student screening scores informed by both the Instructional Hierarchy (IH) summarized in chapter two and published empirical studies (Burns et al., 2008; Parker & Burns, 2014). Tutors, with the supervision of literacy coaches and coaching specialists, selected evidence-based text reading fluency interventions from a list of several options on which all tutors were trained. Reading Corps’ text fluency interventions include Newscaster Reading, Duet Reading, Repeated Reading with Comprehension Strategies, and Stop/Go (Markovitz et al., 2022). For an illustrative example, when participating in the Newscaster intervention, students are instructed to read a passage aloud as if they were a news anchor reading the evening news. Tutors provide immediate error correction if students add, delete, substitute, or mispronounce a word. After students read through the passage, the tutor reads the passage smoothly three times while ensuring the student follows along with their finger. For the next three readings, the student and tutor read the passage in sync with the student attempting to match the tutors’ tone and pace. The student reads independently for the final reading (Van Norman et al., 2020).

When providing Advanced PA intervention, tutors also were able to choose one of three programs (Markovitz et al., 2022). Due to the lack of text, Advanced PA interventions are relatively similar in structure. For example, students receiving one intervention, Sound Awareness, listen to their tutor say a word and instruction about how to modify that word (e.g., say ranch but without the –ch). The students verbally respond with the correct answer, and the tutor moves onto the next word and prompt (e.g., say band without the –d).

Tutors collect baseline, weekly progress monitoring data, and end-of-school-year data. Some students who reach a pre-determined performance benchmark are eligible to exit supplemental reading instruction. Tutors are also provided explicit guidance for assigning interventions to students, based on the Instructional Hierarchy (IH; Van Norman et al., 2018). This guidance includes specifications that students in second and third grade should not receive Advanced PA instruction unless three conditions are met: very low fluency (<30 words read correct per minute), very low accuracy (< 50%), and little to no progress with a phonics-based intervention; Parker, 2023). It is important to note that this guidance, as alluded to in Clemens et al., 2021 (pp. 28), is based on "an absence of evidence" for the efficacy of Advanced PA, not evidence of the inefficacy of Advanced PA.

Despite this guidance, many second and third grade students were assigned to Advanced PA intervention regardless of whether they met the three required conditions. With this less systematic form of intervention assignment, modern statistical adjustments can provide the opportunity to compare reading outcome data and reading growth among students who did and did not receive Advanced PA instruction. This comparison would be the first empirical, causal investigation of the relative value of Advanced PA instruction for students beyond early literacy stages (i.e., in second and third grade).

Chapter III. Method

Participants

Data in the original analytic sample were obtained from a larger database maintained by ServeMinnesota, which is the research and evaluation arm of Americorps programs in Minnesota. The sample includes data from 6,089 students in second and 6,487 students in third grade (N = 12,576) who participated in Reading Corps during the 2018-2019 school year across 771 schools. Students in the sample were 47.1% Female. According to administrative records, 3,003 students identified as Black/African American, 1,318 as Hispanic/Latin(x), 173 as American Indian/Alaska Native, 314 as Asian, 330 as multiracial, 59 as Native Hawaiian or Pacific Islander, and 6,480 as White. One-hundred forty-five students received special education services and 1,426 (11.33% percent) were classified as English Learners (ELs). Special education participation and EL status were reported by tutors, while remaining demographic information was provided by school records. Table 1 shows participant background characteristics by intervention received.

Procedures

Intervention procedures are described above in Chapter 2 (see Overview of the Reading Corps Program). Because the present study contributes only analytic research, only analytic procedures are described in the methods section.

Measures

CBM-R. Students completed Curriculum Based Measures of Reading (CBM-R) probes at baseline as part of the universal screening procedure and again at the end of the 2018-2019 school year. CBM-R probes task students with reading a grade-level connected text passage for

one minute while “trying to read each word” (Christ et al., 2018). Tutors follow along while the students read aloud, marking each error (e.g., substitution, deletion) and the last word read by the student at the end of one minute. Each probe administration is scored using three metrics: words correct per minute (i.e., WCPM), words incorrect per minute (i.e., errors per minute), and percent words correct per minute (i.e., accuracy). WCPM represents the scale on which the construct of Oral Reading Fluency (ORF) is measured. CBM-R probes are also used by the tutors to measure weekly growth for students who are in the Reading Corps program, but those data were not used for analyses. Baseline ORF, errors, and accuracy were used as covariates in propensity weight calculation, Spring ORF will serve as the primary outcome variable.

The validity of CBM-R probes is best evaluated for universal screening procedures by diagnostic accuracy metrics (Kilgus et al., 2014; Klingbeil et al., 2017). Modern probes, including those used with students in the present study and have sufficient technical adequacy to predict end of year performance outcomes with a sensitivity – the rate of true positive results – above 0.75 and a specificity – the rate of true negative results –above 0.71 (Christ et al, 2018). To further support the validity of CBM-R, the vendor of probes used in the present study reports concurrent validity estimates above 0.90 with CBMs developed by other vendors and above 0.66 with computer adaptive tests (Christ et al., 2018). CBM-R is also broadly predictive of important future reading outcomes like reading comprehension, word reading, and vocabulary (Reschly et al., 2009; Wayman et al., 2007). The vendor also reports alternate form reliability coefficients between 0.62 and 0.95, with reliability increasing as student grade-level does. For second grade students, correlations range from 0.81 to 0.97 depending on the alternate measure. This range is 0.78 to 0.96 for Grade 3 students.

Exit Status. Reading Corps discontinued intervention services due to a student's move, schedule change, or parent request at any point in their tutoring. Students were eligible for performance-based intervention exit upon meeting a pre-determined performance criterion. According to Reading Corps, students successfully qualified for graduation from services when they demonstrated three or more consecutive progress monitoring data points above their goal with two or more points above the upcoming triennial screening benchmark standard for their grade (Markovitz et al., 2022). Students who exited due to meeting criteria were coded as 1 (experiencing successful exit event) while all others were coded as 0 (no successful exit event). I used Reading Corps tutor records to censor students who moved and those who stayed in the school but discontinued services for any other reason.

Covariates. I used several demographic and descriptive variables as covariates in analytic procedures. These includes students' baseline ORF, errors, and accuracy as well as race (as reported by census categories), sex, entitlement program eligibility (English Learning services, Special Education), home language, and school. Tutors were not included as covariates because most schools had only one tutor in the final sample, meaning they were nearly perfectly nested within schools.

Methods

Design. Reading Corps training materials, as well as literacy coaches and coaching specialists instructed tutors on which students should, and should not, receive Advanced PA intervention. Advanced PA instruction was distributed according to non-random procedures, creating non-equivalent "Treatment" (Fluency with Advanced PA intervention) and "Control" (Fluency intervention only) groups. Hereafter, the terms FL, or 'Fluency' will refer to the control condition of students who only received Fluency instruction, and FLPA will refer to the

treatment condition of students who received Advanced PA instruction in addition to Fluency practice.

Sites did not distribute interventions with strict adherence to Reading Corps' systematic guidance. As a result, the groups of students do not differ systematically, but rather probabilistically. Table 1, below, outlines observed covariates stratified by intervention type. Students who received Advanced PA instruction in addition to traditional text-based instruction (FLPA) began the year with lower average reading fluency scores than students assigned only to tradition text-based instruction (FL). Table 2 shows the significance of this difference, as well as mild imbalances of other covariate differences.

Because students were not selected for FLPA instruction randomly, they differ on meaningful confounding variables and causal inference cannot be inferred by simple comparisons of group outcomes. Estimates of the treatment effect delta would be biased by the uncontrolled influence of imbalanced covariates (Austin, 2011). Formalized in statistical notation:

$$E[Y(1) \mid D = 1] \neq E[Y(1)]$$

Treatment indicator D_i is not valid instrument for outcome Y_i when both values are affected by covariates X_i (Angrist & Pischke, 2014).

However, propensity score weights which account for the differential probability of being assigned to treatment D_i based on background covariate values X_i can be used to create groups sufficiently comparable for inferring causality (Austin, 2011; Hernán & Robins, 2023). Due to students' non-random assignment to treatment or control groups, students with certain background characteristics were more likely, *but not guaranteed*, to be assigned to FLPA instruction. Table 1 shows the high degree of overlap of groups with respect to background

characteristics. Based on these continuous probabilities of intervention assignment, experimentally comparable groups can be created using propensity score weights that systematically account for the influence of background characteristics on treatment assignment (Hernán & Robins, 2023).

Hernán and Robins (2023) described the mathematical logic underlying the validity of propensity weights. Individuals each have a treatment indicator value D_i and covariate values X_i, but in the context of a non-randomized study D_i is dependent on X_i. Propensity scores account for this dependency by assigning inverse probability of treatment weights according to the probability of receiving treatment $D = d$ given $X = x$ as

$$W^a = 1/Pr[D = d \mid X = x].$$

Higher degrees of dependence between D_i and X_i are weighted with more extreme values. If weights are calculated accurately, conditioning on weights W will balance values of prior covariates leading to the conditional independence of potential outcomes on D_i given W. This inference is known as the assumption of strong ignorability (Rosenbaum & Rubin, 1983) which is formalized as:

$$[Y(0), Y(1)] \perp D_i \mid p(X) = p(x)].$$

Crucially, strong ignorability includes the assumption that all potential confounders are observed (i.e., measured), otherwise conditioning on neither propensity scores nor a vector of covariates will produce conditional independence due to the influence of unobserved confounds. A corollary of the strong ignorability assumption stipulates that treatment assignment is not deterministic for any value of X_i ($0 < Pr(D = 1|X) < 1$). If this assumption is violated, no weight could account for the complete dependence of treatment assignment on X_i; knowing $X = x$ would tell you the value of D_i with certainty. If this assumption is not violated, continuous, non-

deterministic probabilities are then computed using probability density functions, as opposed to the observed empirical probability. In turn, inverse probability weights are most accurately notated as $W^A = 1 / f[D|X]$, where f represents the specific method used to estimate weights, which can vary (Hernán & Robins, 2023).

Propensity scores for the present study were estimated using entropy balancing of treatment assignment probability (Hainmueller, 2012; Tübbicke, 2021). Like traditional generalized linear models of propensity score estimation, entropy balancing primarily relies on logit regression to estimate weights. However, whereas traditional estimators use Maximum Likelihood optimization methods, entropy balancing procedures optimize entropy using a specialized loss function. Specifically, entropy balancing minimizes the Kullback (1959) divergence: $h(w_i) = w_i log(w_i / q_i)$ where the function h decreases towards zero as weights w_i approach base weights q_i (Hainmueller, 2012). In this respect, weights are calculated for both their ability to maximize exact balance and minimize extreme weight values. Both entropy balancing and traditional logit propensity scores estimated using the following logit model:

$$logit(\pi i) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_p X_{pi}$$

And differ only with respect to the parameters to optimize.

I used the R Package "WeightIt' (Greifer, 2023) to estimate and compare traditional and entropy balancing propensity score weights based on a vector of ten covariates: race, gender, EL status, special education status, home language, baseline words read correct per minute, baseline errors per minute, school, and tutor. A total of 12,576 students were eligible for propensity score weighting by having documented receipt of fluency intervention alone (FL; n = 10,820) or fluency intervention combined with Advanced PA (FLPA; n = 2,081). Characteristics of this sample are described above in the *Participants* section. This sample was reduced to 1,705

students in Grade 2 and 1,531 in Grade 3 after removing schools and tutors with treatment probabilities of 0 or near 1 (i.e., where zero or almost all students were in the same condition). These blocks with near definitive probabilities of treatment assignment could not produce balanced propensity scores. Characteristics of this trimmed sample are described below in Chapter IV. A PRISMA diagram outlining the flow of participants from original dataset to final sample is provided in Figure 5.

After balancing with entropy balancing weights, the effective sample size in Grade 2 was 397 FLPA and 223 FL students ($\hat{n}$ = 620). In Grade 3, the effective sample size was 181 FLPA and 161.29 FL students ($\hat{n}$ = 342). Table 2 shows covariate balances in standard mean differences by treatment group when (a) unweighted, (b) weighted by generalized linear regression, and (c) weighted by entropy balancing. A love plot visualizing these covariate (im)balances under all three scenarios is provided in Figure 6.

Analytic Approach

Missing data was accounted for using multiple imputation. With balanced experimental groups created using propensity weights, I developed two models of students' end-of-year reading fluency for implementation with the 'estimatr' and 'WeMix' packages in 'R' (Bailey et al., 2023; Blair et al., 2022). I estimated the Average Treatment effect on the Treated (ATT), which represents the impact of receiving FLPA instruction for students who actually received it, as opposed to for all students (Wang et al., 2017). I estimated the ATT using two regression estimators. The first regression model was developed using a fixed effect for school site and a treatment by covariate interaction term, formalized as:

$$Y_{ij} = \beta_0 + \beta_{1j} + \beta_2 D_{ij} + \beta_{k\ldots} X_{ij} + \beta_{z\ldots} D_{ij} X_{ij} + \varepsilon_{ij}$$

Spring ORF scores for student *i* in school *j* are predicted by the overall average β_0, the school fixed effect β_1, effect of treatment received β_2 (where D = 1 for FLPA, and D = 0 for FL), β_k represents the fixed effects for the covariate vector *X*, and β_z represents the effects for a vector of treatment by covariate interactions. The second, multilevel model included random intercepts at the school level, where student *i* is nested within school *j*:

$$Y_{ij} = \beta_0+\beta_1 D_{ij}+\beta_{2\ldots}X_{ij}+\beta_{3\ldots}D_{ij}X_{ij}+\boldsymbol{\mu_{0j}}+\varepsilon_{ij}$$

In this multilevel model, a term is added to allow for the school specific intercept $\boldsymbol{\mu_{0j}}$. As such, estimates will account for the heterogeneity in school level average ORF performance. Calculated propensity weights were included in both models. For both models, standard errors and confidence intervals were clustered at the school level.

To assess the relative value of Advanced PA instruction more comprehensively, I also conducted a survival analysis. Survival analysis aimed to evaluate Research Question 2. As stated above, students who met the Reading Corps program exit criterion were eligible to graduate from Reading Corps tutoring (and return to receiving Tier 1 instruction only). Students graduated from intensive supports when their skills were remediated to fall within the range of grade level expectations (Fuchs & Fuchs, 2017), making graduation from these supports an important outcome of interest for Tier 2 interventions (Jimerson et al., 2016). Using survival analysis terminology, exit from intensive services can be thought of as the event indicator of interest, with intervention duration in weeks or months serving as the time-to-event metric (Klein et al., 2016). The proportional hazards model (i.e., Cox Model) estimates the rate at which individuals experience the target event, given covariates X:

$$\lambda(t|X) = \lambda_0(t)\ exp\{\beta^{\iota}X\}$$

Where λ represents either the baseline (λ_0) or estimated hazard, *t* is time, and β represents model coefficients over terms β_{ij}. This model is among the most popular methods of survival analysis (Houwelingen & Stijnen, 2014). In the context of the present study, I used the proportional hazards model to estimate the likelihood of intervention exit (λ(t)) among participants with characteristics *X* for each treatment D_i. Student level weights were included to account for non-random treatment assignment. Much like when interpreting an odds ratio, the proportional hazards ratio suggests an increased event (intervention exit) probability when *exp(β)* is greater than 1, and a decreased event probability (intervention continuation) when less than 1. Effective treatments will increase the hazard ratio over time *t* indicating that intervention exit is more likely (Klein et al., 2016). Ineffective treatments, reciprocally, will decrease the hazard ratio indicating a diminished likelihood of experiencing intervention graduation over time *t.* Survival analyses were completed in the 'R' package 'survival' (Therneau, 2022).

Research Question 3 sought to investigate whether certain subgroups (e.g., EL students, students receiving special education, students across racial groups) responded differentially to Fluency intervention with or without Advanced PA. I intended to examine the treatment by covariate interaction coefficients to determine whether differences emerged in the treatment effect estimate for students with specific values of covariates such as race or entitlement program eligibility. However, these analyses were unable to be completed. Sparse cell counts in students' race and special education classification prohibited the estimation of valid treatment effects for these subgroups alone. Moreover, the inclusion of students' racial category and special education participation as covariates prohibited accurate estimation of regression coefficients entirely. In both fixed and random effect models, including students' race or special education participation produced overfitting for these covariates and implausible coefficients (i.e., in the order of

millions). As a result, Research Question 3 was not evaluated in the present study. Students' race and special education status were also removed as covariates from both fixed and random effect models in investigations of Research Questions 1 and 2.

Chapter IV: Results

Students, Tutors, and Schools

The final sample – trimmed to adhere to the assumption of strong ignorability such that no schools contained only students in one treatment condition – contained 3,236 students from 442 tutors across 281 schools. A total of 1,705 of these students were in second grade, with the remaining 1,531 students in third grade. On average, one tutor served each school, although schools ranged in their number of tutors from 1-9 (*IQR*: 1-2). Tutors' caseloads were fixed at 15-16 students by Reading Corps. Tutors worked with an average of 6 students in the analytic sample (*IQR*: 5-9), indicating that tutors provided 2nd and 3rd grade level fluency interventions (with or without Advanced PA) to approximately 1/3rd of their caseload on average. Schools contained an average of 9 students in this sample (*IQR*: 6-15), with the range expanding from 4 to 49.

School Level Descriptive Statistics

School characteristics (e.g., level two descriptive statistics) are outlined in Table 3. School characteristics described here represent grouped descriptive statistics clustered at the school level. They do not include information about Reading Corps participants ineligible for propensity score weighting (i.e., Kindergarten and Grade 1 students, students receiving decoding intervention) nor broader public information about school setting (e.g., percent of students receiving FRPL, whole school demographics). School level descriptive statistics are provided solely to briefly describe the variance in observed measures between schools.

The cluster weighted baseline oral reading fluency (ORF) for schools in the final sample was 43.40 words correct per min (WCPM; *SD* = 10.26; *IQR*: 37.59 – 51.00) in second grade and 72.37 (*SD* = 13.77; *IQR*: 63.70 – 83.47) among third grade participants. Schools ranged in their

average baseline ORF from 10.26 – 66.50 in second grade and 28.38 – 92.97 WCPM in third grade. The average school level accuracy rate at baseline was 87.04% (*SD* = 8.5%; *IQR*: 84.4% - 92.3%; *Range* = 44.04 – 100%) for second grade participants and 92.97% (*SD* = 4.71%, *IQR*: 91.02 – 96.12; *Range* = 73.21 – 98.68%) for third grade participants.

In second grade, the average school contained 50.40% male participants (*IQR*: 37.2 – 66.7%). The average school contained 49.17% (*IQR*: 40.0 – 57.1%) male participants in third grade. Approximately 86.36% (*SD* = 25.37%, *IQR*: 83.00-100%; *Range* = 0 – 100%) of second grade participants in each school spoke English as their home language. For third grade participants, the cluster weighted average percentage of participants speaking English at home was 81.75% (*SD* = 28.15%, *IQR*: 71.43 – 100%; *Range* = 0 – 100%). On average, 10.08% (*SD* = 19.00%; *IQR*: 0.0 – 13.4%) of second grade participants were identified by the program as English Learners (EL) within each school, with some schools containing zero second grade EL participants and some containing only second grade EL participants. The average proportion of EL participants rose to 12.80% (*SD* = 19.79%; *IQR*: 0.0 – 17.8%) in third grade. Some schools included zero third grade EL participants, but the highest percentage of third grade EL participants was 86.96%. Second grade participants received special education at a school level average rate of 3.61% (*SD* = 11.63%, *IQR*: 0.0 – 0.0%; *Range* = 0 – 66.67%) and third grade students were enrolled in special education a cluster weighted average rate of 2.82% (*SD* = 11.84%, *IQR*: 0.0 – 0.0%; *Range* = 0 – 100%).

Schools varied in their racial and ethnic composition as well. On average, schools were composed of 50.21% participants who identified as White (*IQR*: 6.67 – 88.89%), 27.28% who identified as Black or African American (*IQR*: 0.00 – 50.00%), 10.68% who identified as Latin(x) (*IQR*: 0.00 – 14.29%), 2.75% as Asian (*IQR*: 0.00 – 0.00), 2.18% as Multiracial (*IQR*:

0.00 – 0.00), 1.41% as Native American or Alaska Native (*IQR*: 0.00 – 0.00), and 0.34% as Native Hawaiian or Pacific Islander.

Unweighted Descriptive Statistics of Included Students

All descriptive statistics are reported using listwise deletion for missing data. Descriptive statistics calculated from pooled multiple imputation datasets are available in Table 4. Raw descriptive statistics are depicted for the updated, trimmed sample in Table 5. Descriptive statistics for the original untrimmed sample can be found in Chapter 3.

Assessing whether baseline characteristics mirror those of included students is necessary to determine the degree to which our trimmed sample accurately represents the larger Reading Corps population of schools. Second grade students who received FL read more words correct per minute (46.19, *SD* = 13.92) than did students who completed FLPA (34.00, *SD* = 16.64). FL students in second grade also had higher reading accuracy (89%, *SD* = 10%) and fewer errors (5.00, *SD* = 4.18) at baseline than FLPA (80%, *SD* = 17% and 6.77, *SD* = 5.47). In second grade, similar proportions of Female students (48.8% vs 47.4%), English Learners (9% vs 10%), students in Special Education programs (3% vs 5%), Indigenous students (1.3% vs 1.3%), Latin(x) students (8.1% vs 9.8%), Multiracial students (2.3% vs 1.8%), and White students (56% vs 52.6%) participated in FL and FLPA groups, respectively. Slightly higher proportions of Asian students (2.9% vs 1.5%) and lower proportions of Black students (24.6% vs 29.7%) participated in the FL group relative to the FLPA group in second grade.

Similar trends emerged for third grade participants. FL students read more words per minute at baseline (74.89, *SD* = 19.05 vs. 56.08, *SD* = 26.31) with higher accuracy (94%, *SD* = 6% vs. 86%, *SD* = 13%) and fewer errors (4.16, *SD* = 3.70 vs. 6.72, *SD* = 4.25) than FLPA students in third grade. Similar proportions of Female students (49% vs 47.5%), students in

Special Education programs (3% vs 2%), Asian students (3.5% vs 4.4%), and Indigenous students (2.0% vs 2.8%), participated in FL and PA groups, respectively. The FL had slightly smaller proportions of English Learner students (13% vs 17%), Black students (35.3% vs 40.3%), Latin(x) students (11.6% vs 14.4%), yet slightly higher proportions of White students (39.7% vs 33.1%) and Multiracial students (2.4% vs 1.1%).

At year's end, second grade students who received FL intervention (Control; 100.13, *SD* = 21.83) still read more words per minute than FLPA students (74.97, *SD* = 27.38) when unweighted. This trend held for Grade 3 students, where unweighted FL students read 116.57 words per minute (*SD* = 24.82) and FLPA students read 87.01 words per minute (*SD* = 33.32).

Characteristics of Excluded Students and Schools

Baseline characteristics of excluded students and schools are depicted in Table 6. Excluded students who received fluency instruction (FL; n = 5351) read an average of 64.79 (*SD* = 22.70) words correct per minute (WCPM) at baseline. Second grade students in FL demonstrated an average of 46.94 WCPM (*SD* = 14.76) compared to third-grade FL students' average of 77.74 WCPM (*SD* = 18.23). The average excluded student who received FLPA read 32.13 WCPM (*SD* = 19.6) at baseline, with second grade FLPA students' average of 28.18 (SD = 16.30) falling below third grade FLPA students' average of 41.97 (*SD* = 23.33). Second grade FL students made an average of 5.02 errors per minute (*SD* = 4.91) producing an average accuracy of 89.00% (*SD* = 11.00%). Students receiving FLPA in second grade made slightly more errors per minute (7.39, *SD* = 4.36) which resulted in a much lower average accuracy value (74%, *SD* = 19%) due to their reduced WCPM. Third-grade students in the FL group made 3.73 errors per minute (*SD* = 4.48) on average, with an average accuracy of 95% (*SD* = 6.00%). In contrast, FLPA students in third grade made 6.55 errors at baseline (*SD* = 4.58) and read with an

average accuracy of 82% (*SD* = 15%). These baseline characteristics mirror closely those of the Included students prior to weighting, indicating that although the sample was trimmed to adhere to strong ignorability, it does not appear as though schools who did or did not contain students in both treatment groups differed significantly based on observed characteristics.

As with the clustered statistics reported for included schools above, school descriptive statistics provided here represent only participant averages weighted by school clusters. A total of 723 schools were excluded, containing a total of 1,221 tutors who served 9,569 2nd and 3rd grade students. On average, excluded schools served 11 excluded participants (*IQR* 7-17), excluded tutors served 7 students (*IQR* 4-11), and schools contained 2 tutors (*IQR* 2-4). The average school contained second-grade excluded participants who scored 42.77 WCPM (*SD* = 13.11) at baseline with 85.85% accuracy (*SD* = 10.47%). School level averages for third grade excluded participants were 75.43 WCPM (*SD* = 15.21) with 93.99% accuracy (*SD* = 5.30%). Demographically, the average school enrolled second grade excluded participants who received EL supports at a rate of 12.71% (*SD* = 24.13%), received Special Education at a rate of 3.66% (SD = 13.78%), and spoke English at home at a rate of 81.50% (*SD* = 32.29%). On average, 50.48% of excluded participants within each school identified as male. In third grade, the average school was comprised of 11.88% excluded participants who received EL supports (*SD* = 23.72%), 3.41% who received special education (*SD* = 13.19%), and 82.47% who spoke English at home (*SD* = 31.64%). An average of 48.32% of third grade excluded participants within each school identified as male. Among schools containing second grade excluded participants, on average 49.69% of excluded participants identified as White, 23.50% as Black or African American, 10.28% as Latin(x), 2.86% as Multiracial, 2.54% as Asian, 1.34% as Native American or Alaskan Native, and 0.43% as Native Hawaiian or Pacific Islander. In third grade,

the average school contained excluded participants who were identified as 55.73% White, 18.16% Black or African American, 10.35% Latin(x), 2.67% Multiracial, 2.57% Asian, 1.48% Native American or Alaskan Native, and 0.70% Native Hawaiian or Pacific Islander.

Weighted Descriptive Statistics of Included Students

Descriptive statistics were calculated again after accounting for propensity score weights. Weighted descriptive statistics are provided in full in Table 7. After weighting, Grade 2 FL students and FLPA students showed similar baseline levels of words read correct per minute (33.90, *SD* = 13.92 vs. 34.00, *SD* = 16.64), errors (6.76, *SD* = 4.18 vs. 6.77, *SD* = 5.47), and accuracy (80.22%, *SD* = 10.38% vs. 80.30%, *SD* = 16.79). Grade 3 students also demonstrated baseline equivalence on key variables after reweighting. FL and FLPA students in the third grade had similar levels of words read correct per minute (56.07, *SD* = 19.05 vs. 56.08, *SD* = 26.31), errors (6.72, *SD* = 3.70 vs. 6.72, *SD* = 4.25), and accuracy (85.76%, *SD* = 5.90% vs. 85.77%, *SD* = 12.79%).

At the end of the school year, differences emerged in the weighted average levels of words read correct per minute. In Grade 2, FL students finished the year reading a weighted average 93.75 (*SD* = 21.83) words correct per minute compared to FLPA students' weighted average of 74.97 (*SD* = 27.38) words. FL students in third grade ended the year with a weighted average of 106.56 words correct per minute (*SD* = 24.82), while FLPA students demonstrated a weighted average WCPM of 87.01 (*SD* = 33.32). The distribution of weighted descriptive statistics is depicted in Figure 7.

Fixed Effects Model

To answer research question 1, the first analytic model investigated differences in Spring oral reading fluency by intervention group (Fluency vs. Fluency with Advanced PA), baseline

and demographic covariates, the interaction of each covariate with intervention group status, and a fixed effect for the school in which students received tutoring. Student weights were incorporated into the linear regression model and covariates were centered within cluster around the mean of the treatment condition. Centered means are depicted in Figure 6b and presented by condition in Table 7b. The full output for fixed effect models is presented in Table 8. Site specific treatment effect estimates are presented visually in Figure 11. During model specification, covariates indicating students' race and special education status were removed for their interference with parameter estimation. The number of students in some racial categories and the number receiving special education services were so small that attempting to estimate effects for these subgroups inhibited the model's ability to credibly estimate all other effects.

When using a listwise deletion approach to missing data, the fixed effect model for second grade students produced a treatment estimate of -15.8 (*SE* = 5.89; *95% CI:* -27.4 - -4.87). This indicates that students who received Advanced PA in addition to fluency intervention (FLPA) can be expected to score 16 words correct per minute lower in spring than if they received fluency intervention alone (FL) when accounting for propensity score weights. Using the more robust multiple imputation approach to missing data, fixed effects models pooled across 20 imputed datasets similarly produced an estimate of -11.8 (*SE* = 2.09; *95% CI:* -15.9 - -7.69). This result provides a more credible estimate while confirming that patterns of missing data are not responsible for the estimate obtained by the listwise deletion model.

For third grade students, a fixed effect model with listwise deletion yielded an estimate of -8.52 (*SE* = 1.79; *95% CI*: -12.00 – -5.01). In essence, third grade students who received FLPA scored an average of 9 fewer words correct per minute when controlling for baseline covariates and disproportionate probabilities of treatment assignment. Pooled multiple imputation estimates

also yielded an estimate of -13.8 (*SE* = 3.72; *95% CI*: -21.10 – -6.52) indicating that this estimate is credible and not due to missing data patterns.

Random Effects Model

To further answer research question 1, as well as part of the third research question, the second analytic model investigated differences in Spring oral reading fluency by intervention group (FL vs. FLPA), baseline and demographic covariates, the interaction of these covariates with intervention group status, a random effect term for school (random intercept), and a random effect term for the interaction between school and intervention group (random slope). Students' race and special education status were not included as covariates in random effects models due to numeric insufficiencies. However, prior literature using the same dataset attests that effects of Reading Corps are robust across racial categories (Markowitz et al., 2022). Student weights were incorporated to the model at level one. Further, for this model, binary and continuous covariates were centered within each cluster around the treatment condition mean. The full output for this random effect model is presented in Table 9 and a caterpillar plot presenting site specific random effects is provided in Figure 8. Further, example site specific treatment effect graphs are provided in Figure 9 for 15 randomly selected schools.

The variance term for the school level in second grade was 351.81 (*SE* = 17.76). The random effects model produced an overall treatment estimate of -13.25 (*SE* = 2.02) for second grade students, indicating that students who received FLA intervention scored an average of 13.25 fewer words correct per minute in spring than they would have if they received only Fluency intervention. Results from 20 multiply imputed datasets produced a nearly identical pooled average estimate of -12.93 (*SE* = 1.82). In Grade 3, the variance term for the school level was 724.23 (*SE* = 26.91). The random effects model across 20 multiply imputed datasets

produced a treatment effect estimate of -8.81 (*SE* = 2.60). This effect, though slightly attenuated, has the same direction as the second-grade effect. When accounting for missing data, baseline and demographic covariates, and unbalanced treatment assignment, students who receive FLPA intervention experienced reduced growth in oral reading fluency from Fall to Spring.

Survival Analysis

Research Question 2 sought to explore whether FLPA or FL intervention led students to successfully "graduate" intervention (i.e., return to only universal, Tier 1 supports) more quickly. Survival curves were modeled using the Cox proportional hazards model (Klein et al., 2016). The number of weeks that students participated in intervention served as the time variable, while intervention exit served as the 'event' variable. Sufficient data was available to censor students who moved schools or discontinued intervention for other reasons. Baseline covariates, a categorical school identifier, demographic variables, and student weights were incorporated into this model. Much like the random effects model, survival curves could not be generated when including students' race as a covariate due to numeric insufficiencies. Race was subsequently excluded as a covariate from Survival analyses. Survival analysis data are presented quantitatively in Table 10 and visualized using survival curves in Figure 10.

The Cox proportional hazards model produced a hazards ratio of 0.21 (*SE* = 0.06) for Grade 2 students who received FLPA intervention relative to those who received only FL intervention. This ratio indicates that when controlling for baseline covariates and biased treatment assignment, students who received FLPA were approximately 1/5th as likely as FL students to successfully graduate intervention within an academic year. This estimate of 0.21, pooled across 20 imputed datasets, closely matches the estimate obtained when using listwise deletion (0.22, *SE* = 0.59).

For Grade 3 students across twenty imputed datasets, those who received FLPA intervention held a hazard ratio of 0.19 (*SE* = 0.06); third grade students were approximately 1/6th as likely to meet exit criteria when receiving FLPA instruction relative to when receiving FL instruction. Interestingly, a Cox model run on the original dataset using listwise deletion produced a discrepant estimate (0.35, *SE* = 0.38). Though the two estimates yield similar takeaways, the magnitude of the effect is significantly stronger when accounting for missing data.

It is important to ensure that the operational definitions of treatment receipt (i.e., 30+ sessions of Advanced PA) did not artificially affect graduation rates between the two groups. To rule out this possibility, I re-ran the survival analysis when including students who received as few as 11 sessions of Advanced PA in the treatment group. Eleven sessions is the minimum possible number of days of intervention a student could complete before meeting graduation requirements; any student who stopped before 11 sessions did not stop due to successful graduation. The hazard coefficient for students who received between 11 and 30 sessions of Advanced PA training was estimated to be 0.24 (SE = 0.42) in Grade 2 and 0.47 (SE = 0.33) in Grade 3. These estimates show the same direction and significance level as those obtained using the original treatment assignment criteria. This indicates that it is unlikely the 30-session minimum treatment assignment mechanism masked detectable amounts of graduation among students who received Advanced PA intervention. If students graduated prior to six-weeks of intervention, they graduated at rates similar to those observed between groups after six weeks of intervention.

Sensitivity Analyses

I conducted sensitivity analyses to examine the degree to which findings were robust to alternative but similarly acceptable analytic decisions. In the first group of sensitivity analyses, I varied the operational definition of treatment assignment by decreasing or increasing the number of Advanced PA sessions required to fall into the treatment category. Estimates of marginal and random effects from sensitivity analyses are presented in Table 11. Results did not meaningfully differ from those reported above when the treatment assignment mechanism was manipulated. Decreasing the number of sessions required to be categorized as a participant in Advanced PA intervention from 30 to 20 sessions slightly changed 2nd grade fixed effect estimate from -11.80 to -11.40, and third grade fixed effect estimate from -13.80 to -12.00. This decrease in the number of required sessions changed random effects from -12.93 to -12.87 in second grade and from -8.81 to -7.68 in third grade. To the same extent, increasing the number of required sessions from 30 to 40 sessions did not significantly change effect estimates. Second grade estimates grew to -14.10 for fixed effect models but shrank to -11.99 for random effect models. Third grade effects changed to -9.36 and -11.87 for fixed- and random-effect treatment estimates, respectively.

In the next group of sensitivity analyses, I varied model specification to evaluate the fit (a) without treatment by covariate interactions, (b) with quadratic terms, and (c) with random slopes in addition to random intercepts. Results were robust to changes in model specifications. Removing treatment by covariate interactions changed marginal effect estimates from -11.80 to -12.80 in second grade and -13.80 to -8.89 in third grade. Random effect estimates changed from -12.93 to -12.82 in second grade and -8.81 to -8.90 in third grade when removing treatment by covariate interactions. A quadratic regression specification changed second grade treatment effect estimates to -11.00 in the fixed effect model and -12.68 in the random effect model. Third

grade treatment effect estimates changed to -3.49 in fixed effect models and to -16.73 in random effect models when pivoting from linear to quadratic regression. Finally, adding random slopes (i.e., allowing treatment effect to vary by school) increased the strength of random effect estimates to -12.91 in second grade and to -20.99 in third grade. Quadratic specification of fixed-effect regression for third grade students was the only circumstance under which sensitivity analyses or main analyses did not statistically favor FL over FLPA ($p = .680$). Therefore, regression-based treatment effect estimates were robust to researcher analytic decisions and not sensitive to model construction nor treatment assignment specifications.

Finally, I varied the methods used for multiple imputation. Keeping all other analytic choices constant, I created separate datasets for treatment and control students within each grade. I then multiply imputed independent datasets for each condition before stitching the now imputed datasets back together. The Fixed Effect estimate for second grade students (-12.3, SE = 2.08) changed minimally from the original estimate (-11.80, SE = 2.09). The estimate for third graders (-21.30, SE = 5.77), however, was moderately more extreme than the original estimate (-13.8, SE = 3.72). Multiply imputing datasets from the original, whole sample did not exacerbate condition difference. If anything, using a single multiple imputation step provided a more conservative estimate of condition differences.

Chapter V: Discussion

I conducted the first empirical evaluation the incremental benefit of Advanced PA training for struggling (i.e., Tier 2) readers above the initial Kindergarten and First grades. More specifically, I compared oral reading fluency outcomes for second- (n = 1,705) and third grade (n = 1,531) students who were receiving supplemental reading instruction from Reading Corps in reading fluency alone (FL) or in combination with Advanced PA (FLPA). Students demonstrated non-equivalent reading fluency, errors, and accuracy scores at baseline. Demographic factors like students' racial background also varied between the two intervention groups. However, when entropy balancing weights calculated from their differential probabilities of treatment assignment were applied, the groups' weighted means at baseline converged. Students' schools were included in the calculation of propensity weights to ensure students were comparable within, as well as across, sites. In my first research question, I evaluated whether Advanced PA caused differences in reading fluency gains over the course of an academic year.

To investigate ORF differences between FL and FLPA interventions, I first fit a regression model accounting for propensity weights, baseline covariates, and fixed effect for school (k = 291). Covariates were centered within school clusters, and verification of this centering's efficacy is presented in Table 12. Results from this model showed that students who received FLPA instruction had significantly lower Oral Reading Fluency (ORF) scores during spring benchmarking than their counterparts who received FL instruction. More precisely, students instructed in Advanced PA read an average of 12 words correct per minute (WCPM) fewer in second grade, and 13 WCPM fewer in third grade than if they had been instructed in Fluency, when controlling for baseline covariates (i.e., baseline ORF, baseline accuracy, baseline errors, EL status, Home Language), and treatment by covariate interactions.

Next, I fit another model with random school (i.e., level 2) intercepts. This model allowed schools to vary in their average reading fluency level. Accounting for this heterogeneity did not meaningfully alter findings. Students in the FLPA condition read 14 fewer WCPM in second grade and 13 fewer WCPM in third grade spring benchmarking after controlling for baseline covariates and treatment by covariate interactions. Overall, the findings from the evaluation of RQ1 – whether Advanced PA caused differences in connected text reading outcomes – demonstrated clear differences in students' spring ORF between students who received FLPA or FL. Students grew significantly less in their reading fluency when a portion of their intervention time was devoted to Advanced PA. Across both random- and fixed-effect models, results were robust to sensitivity analyses indicating that the magnitude and direction of findings cannot be attributed to model misspecification.

Though reading fluency is a critical skill for promoting reading development and comprehension (cf. Kim, 2020a; Shapiro & Clemens, 2023), the goal of Multitiered Systems of Support (MTSS) in reading is not solely to increase students' fluency. The broader goal of MTSS interventions is to allow students to return to Tier 1 instructions with sufficient skills to be successful (Fuchs & Fuchs, 2017). With critically high levels of non-proficiency in reading (Klingbeil et al., 2022), it is important to provide interventions that help students graduate more quickly so other students can be served by oversubscribed school systems. To better understand how Advanced PA instruction affects more distal outcomes like graduating from intervention by meeting Reading Corps' performance-based exit criteria, I also conducted a survival analysis. RQ2 examined whether students receiving FLPA met this criterion and exited intervention at a rate different than that of students in FL intervention. I used a Cox Proportional Hazards model (Klein et al., 2014) to describe students' time-to-exit event and their relative likelihood of exiting

based on the intervention condition, baseline performance, and demographic covariates. Students who received FLPA instruction met exit criteria less frequently and required more time to successfully exit the intervention in comparison to students receiving FL alone. Students in the FL alone condition began to graduate intervention with regularity after approximately 8 weeks (2 months) of intervention, while graduation among students in FLPA intervention was rare until approximately 25 weeks (6 months). Second grade students receiving FL instruction alone were approximately 5 times more likely to exit intervention than students receiving FLPA. Third grade students were nearly 6 times more likely to exit intervention if they received FL as opposed to FLPA.

Support of Original Hypotheses

The present study aimed to evaluate the Advanced PA interventions used by some Reading Corps tutors. Therefore, I did not create original hypotheses to test, as a means to compare the two interventions with as minimal researcher bias as possible. Instead, I investigated whether differences between intervention groups emerged. Given, however, that FL intervention alone has a significant amount of research support (Hudson et al., 2020; Lee & Yoon, 2017; Shapiro & Clemens, 2023) an appropriate null hypothesis for an evaluative study is that no differences would emerge between FL and FLPA. Despite a lack of original hypotheses, findings do contribute important information to contentious hypotheses and claims in the reading instruction literature.

Context of Existing Results & Reading Theory

Competing hypotheses about the role of Advanced PA informed the present study. Researchers in favor of Advanced PA training have hypothesized that bolstering students' Advanced PA skills will strengthen their ability to orthographically map new words and

therefore improve their text reading fluency (Kilpatrick & O'Brien, 2019). Proponents of Advanced PA training contextualize their hypothesis as a natural extension of Ehri's (2005) Phase Model and Share's (1995) Self-Teaching hypothesis, both of which emphasize the importance of orthographic mapping for text reading. Researchers opposed to Advanced PA training hypothesized that Advanced PA skills are unnecessary because they do not add incremental benefit to students' connected text reading over direct instruction in basic reading skills involving graphemes (Clemens et al., 2021). These detractors cite empirical evidence showing the importance of graphemes for phonemic awareness instruction (Ehri et al., 2001) and dissociating Advanced PA skills from skilled text reading (e.g., Foulin, 2005; Scarborough et al., 1998).

This study provides preliminary evidence against the utility of Advanced PA instruction and offer some support for hypotheses offered by those skeptical of Advanced PA. Text reading fluency, measured using ORF in this study, is an important higher-level reading skill that captures several important lower-level skills (e.g., Sight Word Recognition, Word Reading Efficiency, Decoding Ability). Thus, text reading fluency provides an estimation of how efficiently students can use those skills to read text (Kim, 2020a, b; Shapiro & Clemens, 2023). One of the most important tasks necessary for fluent reading is automatic word recognition (Ehri, 2020; LaBerge & Samuels, 1974). Advocates of Advanced PA skills argue that training in these skills is critical for promoting students' ability to store words for automatic retrieval when they encounter them in text by sight. If this was true, findings should have shown that ORF gains in Advanced PA intervention equaled or bested those observed in the fluency instruction group. Because ORF gains attributable to Advanced PA instruction fell significantly short of those attributable to Fluency instruction, it is unlikely that Advanced PA intervention would

meaningfully affect distal text reading outcomes for students who can access connected text. Present findings stand in direct contrast to claims made by proponents of Advanced PA about the relative value of Advanced PA and repeated reading (fluency) instruction:

> "Since repeated reading interventions do not teach the skills required for efficient orthographic mapping, they would not be expected to yield strong, sustained normative results with struggling readers. Likewise, interventions involving large amounts of reading practice (not repeated reading) have similar, limited results (O'Connor et al., 2007; Wexler et al., 2008; …). Ultimately, there is no research evidence to suggest that repeated reading, or similar practice-based interventions, substantially closes the gap between struggling readers and their typical peers." – Kilpatrick & O'Brien, 2019, p. 196.

Not only does ample research exist to support the efficacy of repeated reading instruction (cf. Hudson et al., 2020; Lee & Yoon, 2017), but now preliminary research exists to undermine the claim that adding Advanced PA to practice-based reading fluency interventions will improve text reading in comparison to practice-based fluency intervention only. Students grew more and more quickly when receiving repeated reading instruction than when receiving Advanced PA instruction, making them more likely to "close the gap" to typical peers.

The lack of connection between Advanced PA skills and ORF could signal several potential implications for reading theory above and beyond the possibility that additional training in Advanced PA skills is not necessary for students in second and third grade who need support reading connected text. First, in the order of lowest to highest inference, it is possible that Advanced PA instruction did not improve students' Advanced PA skills. If the intervention used by Reading Corps were ineffective for affecting this mediating skill, it would by nature be ineffective for affecting the outcome of interest. Second, perhaps Advanced PA instruction is

effective but not as effective as the gold standard fluency interventions (Shapiro & Clemens, 2023). However, inferiority to standard practice does not by definition indicate inefficacy, in theory. Investigations of Advanced PA theory may wish to compare Advanced PA instruction to a non-reading instructional control condition. However, for schools that are required to make the most of the limited amount of time for providing intervention, such a test may have little practical significance.

A third possibility is that Advanced PA skills improved by the intervention do not transfer to connected text reading. This conclusion is that reached by Clemens and colleagues (2021), who demonstrated skepticism that Advanced PA skills would meaningfully affect students' ability to orthographically map novel words. If gains in Advanced PA skills indeed fail to transfer to connected text, it may indicate that orthographic mapping skills rely more heavily on (a) phoneme grapheme correspondence (Ehri et al., 2001a, b) or (b) continued, and repeated exposure to novel words (Share, 1995). The possibility that orthographic mapping relies more heavily on phoneme grapheme correspondence ties directly into a fourth possibility. Perhaps Advanced PA skills need to be taught in conjunction with graphemes to effectively improve orthographic mapping and connected text reading. Although proponents of Advanced PA contend that PA instruction with graphemes is instruction in letter sound correspondence or decoding and not PA, the National Reading Panel (Ehri et al., 2001) suggests otherwise and cites evidence that PA instruction is more valuable and efficacious when taught with graphemes.

Two final theoretical accounts for findings in the present study involve higher inferences. The first high-inference account is a prediction offered by Kilpatrick and O'Brien (2019) that reading fluency measures do not accurately capture word reading efficiency. They suggest that standardized measures such as the Test of Word Reading Efficiency (TOWRE-2; Torgesen et al.,

2012) are necessary for capturing gains in text reading efficiency that they expect to follow Advanced PA instruction. Despite this possibility, evidence investigating the relative roles of word reading efficiency and text reading fluency has shown text reading fluency to mediate the relationship between word reading and reading comprehension (Kim & Wagner, 2015; Kim et al., 2021). Meaningful improvements in word reading efficiency, then, should have been observable in reading fluency outcomes.

A final high-inference theoretical account of these results comes again from Clemens and colleagues' 2021 review. Clemens and colleagues reviewed data that suggest many poor readers have strong Advanced PA skills while many strong readers have poor or underdeveloped Advanced PA skills. They argued that that Advanced PA skills may not be necessary for connected text instruction. However, it's possible that participants in this study (i.e., second and third grade students needing support in text reading fluency) already had sufficient Advanced PA skills. Further, improvements in this area will not benefit more advanced reading outcomes. Lonigan and colleagues (2018) show that the impact of decoding instruction is less effective for improving reading comprehension once students are more limited by language comprehension than by word reading skills. Similarly, students who continue to struggle reading connected text after developing some level of Advanced PA skill likely have the source of their challenges located in other component skills.

In sum, findings from the present study undermine predictions that Advanced PA instruction is a useful method for improving text reading fluency for struggling second and third grade students who can access text. Given that hypotheses about a link between Advanced PA and connected text reading rely on orthographic mapping as a mediating skill (Kilpatrick & O'Brien, 2019; Kilpatrick, 2020b), it is unlikely that Advanced PA meaningfully affects

students' orthographic mapping. Had Advanced PA instruction affected orthographic mapping capabilities, students would have improved their word reading efficiency which would have been reflected in their oral reading fluency (Shapiro & Clemens, 2023).

Implications for Program Evaluation

Crucially, results from the present study have implications for practice and policy that are more immediately applicable than their implications for theory. The most proximal implication of these results is that Reading Corps may consider pausing implementation of Advanced PA intervention with second- and third-grade students. Continuing to select Advanced PA intervention for these students harms students materially; their ORF improvements are diminished by the equivalent of between two-months and half of a year's growth based on normative growth rates (Christ et al., 2019). More concerningly, these diminished improvements lead to significantly longer persistence in Reading Corps services and reduced likelihood of returning to Tier 1 services alone. This is counter to the logic of Tier to interventions in general (Fuchs & Fuchs, 2017). These more systemic outcomes, persistence in Tier 2 services, affect Reading Corps as much as they affect students. If students in this sample who received Advanced PA had graduated at the same rate as students who received fluency, Reading Corps would have been able to provide services to more than 300 additional students in the 2018-2019 school year.

Because Reading Corps aims to expand service capacity for Tier 2 intervention delivery, takeaways extend to schools providing Tier 2 services internally. Interventions for second- and third-grade students with some risk in reading fluency should focus on skills other than Advanced PA skills. More specifically, results of the present study show that fluency instruction focused on repeated opportunities to read connected text is more effective in isolation than when

combined with Advanced PA skill training. It is important, however, to note that prior research has shown instructional match to be a critical component of intervention success and not all students will benefit from fluency alone. Many students with Tier 2 needs who fail to respond to an evidence-based fluency intervention may benefit more from intervention focused on decoding instruction with graphemes (cf.; Parker & Burns, 2014; Shapiro & Clemens, 2023; Szadokierski et al., 2017).

Many schools also use Advanced PA instruction as a Tier 1 or universal instructional component (Literary Resources, 2023). I did not evaluate the effect of Advanced PA (a) as a universal curriculum, nor (b) for students who respond to Tier 1 instruction. These findings offer no insight to the utility of Advanced PA for these contexts. The sole *empirical* study to evaluate Advanced PA (Coyne et al., 2021) did examine effects for universal, Tier 1 contexts, finding no significant effects for word reading and connected text outcomes. Advocates clarified that Advanced PA instruction is most necessary for struggling students who experience challenges with efficient orthographic mapping and typically developing readers have less need for these skills (Kilpatrick et al., 2022). Results from the present study broaden the range of reading abilities at which students are unlikely to need Advanced PA instruction to be successful.

As states continue to adopt legislation aiming to systematize reading instruction, policymakers are faced with the difficult task of sifting through an ever-growing body of instructional materials marketed as ‘aligned with the Science of Reading’ (Covington, 2023; Shanahan, 2020). The science of reading (lowercase) is a fluid and growing body of research based on rigorous empirical evidence across fields. As such, the chances of students receiving instruction focused on Advanced PA skills should be lowered based on the available data including the present study, Coyne et al., (2021), and Clemens et al., (2021). Should future

research find circumstances in which Advanced PA is *causally* beneficial for students, it would be appropriate to increase the chances that certain students receive Advanced PA instruction in those circumstances. The links between the Science of Reading (uppercase) movement and the science of reading (lowercase) must continue to be constructed of evidence and data, rather than of philosophy. The unwarranted influence of a persuasive educational philosophy curated an environment in which whole-word reading instruction remained popular far after the evidence indicated it should (Castles et al., 2018; Rayner et al., 2002). To avoid the same being true for the Science of Reading (uppercase) movement, systematic literature reviews, meta-analyses, and policy briefs should be undertaken to summarize the fluid state of the science of reading (lowercase) in ways that are direct and digestible for policymakers.

A final, broad policy takeaway involves the significant equity considerations underlying the present study. As outlined in the descriptive statistics above, students were not randomly assigned to Advanced PA intervention. Students were more likely to receive Advanced PA when reading slower and less accurately at baseline, but also based on demographic factors such as race. Previous data suggest that Reading Corps services are equitably efficacious across racial and ethnic groups (Markowitz et al., 2022). Despite this, students from diverse racial and ethnic backgrounds were more likely to receive the less effective Advanced PA intervention. Students of color and those from other marginalized backgrounds already face significantly disproportionate reading outcomes (Fien et al., 2021). As MTSS is intended to be a preventative model for ameliorating moderate deficits before they become intractable (Jimerson et al., 2016), providing less effective Advanced PA interventions could deepen inequality by disproportionately offering racially diverse students diminished reading outcomes. Survival analyses indicate these diminished outcomes affect graduation from the program, meaning that

the more racially homogenous Fluency group returned to Tier 1 instruction alone at higher rates. In essence, continuing to provide a less effective intervention specifically for students who are lower performing and more diverse, compounds the very inequity that Reading Corps has been notably successful at addressing previously (Markowitz et al., 2022).

Adequacy of Sampling Size and Sampling Validity

The original analytic dataset obtained from ServeMinnesota included all 12,576 Reading Corps participants from the 2018-2019 academic year. After trimming to include only schools with students in both treatment (FL and FLPA) conditions, the final sample included about one-quarter of all Reading Corps students (n = 3,236) and schools (k = 218). Had I been able to acceptably weight all 12,576 participants, the sample would have been equal to the population of Reading Corps participants and decisively representative of Reading Corps outcomes in Minnesota. The question of sampling adequacy then would have been the question of whether Reading Corps participants could reasonably represent a sample of a broader population of students receiving Tier 2 interventions regardless of service provider (school or Americorps tutors). Before assessing this question, the pre-requisite question for the present study involves investigating whether the trimmed sample adequately represents the original population of Reading Corps schools and students. Characteristics of excluded students, as discussed previously, are presented in Table 5. Overall, minimal differences emerged on measured covariates between trimmed and included schools. Although sample trimming was nonrandom, the instrument used to identify schools for trimming (presence of only one intervention group) appeared to function approximately randomly (Angrist & Pischke, 2014). That is, inclusion status T_j (t = 1 if included, t = 0 if excluded) for schools is uninformative of student and school level covariate and outcome values X_{ij} and Y_{ij}.

Generalizability

The present study focused intensely on precisely estimating causal effects for students who received Advanced PA intervention in addition to fluency intervention. The estimand of interest, then, was the Sample Average Treatment effect on the Treated (SATT), indicating that results apply namely to the actual students who received intervention. A different approach prioritizing generalizability over precision in internal validity would have been to estimate the Population Average Treatment Effect (PATE). PATE describes the estimated effect of an intervention for a hypothetical superpopulation from which sampled students were drawn (Hirano et al., 2003). As such, assumptions for estimating PATE are more rigorous and precision is diminished. A key barrier to estimating PATE in the present study regards the original dataset; both students (Level 1) and schools (Level 2) were non-randomly selected for intervention and therefore less likely to be representative of a superpopulation of students and schools implementing Tier 2 interventions.

Nevertheless, research advances have described increasingly efficient ways of describing the generalizability of educational impact evaluations. The Generalizer (Tipton & Miller, 2024) provides point-and-click tools for planning (a priori) for generalizability in designing evaluations and investigating (post hoc) generalizability of completed evaluations. The R package 'generalizeR' (Nixon et al., 2024) offers similar, but code-based, tools using information from the same referent national datasets. Tipton and Olsen (2018) offered several options for improving statistical estimates of generalizability including propensity score methods, model-based approaches, and bounding estimates. The present study has yet to complete these additional estimates of generalizability. Considering this, and the focus on SATT over PATE, it is important to report generalizability responsibly (Tipton & Olsen, 2022). Findings from the

present study are *likely* to generalize to Reading Corps schools who were excluded from the final dataset and had a non-zero probability of inclusion in the final sample. Descriptive characteristics of both students and schools excluded from the present study vary little from those of included students and schools, as described above. To more conclusive estimate generalizability, future directions for this project include (a) using The Generalizer (Tipton & Miller, 2024) to statistically identify the generalizability index, and (b) estimating weighted treatment effects for excluded students using single-level propensity score and outcome models (ignoring school blocks) to compare estimates with those derived from included students.

Limitations and Future Directions

Several limitations affect the present research and provide opportunities for future directions. First, there may still be students for whom Advanced PA is beneficial. The present study did not include students who received only Advanced PA intervention with no fluency practice; it is possible that subgroups of students who respond more to Advanced PA instruction exist in the population of students assigned to Advanced PA alone or Advanced PA with other decoding interventions. It's further possible that outside the scope of this study, students with more intensive instructional needs (i.e., those with Tier 3 or special educational needs) Advanced PA. Future research should examine effect heterogeneity as a function of a wide range of baseline reading performance.

Second, tutor records do not control for the amount of time spent on each skill focus. The analytic dataset included tutor records that described which interventions were provided during each intervention session. These records do not specify the share of time taken up by Advanced PA relative to fluency (e.g., 5 + 15 minutes, 10 +10 minutes). Results from the present study seem to indicate that any amount of Advanced PA instruction attenuates gains that otherwise

would be observed with fluency instruction. However, future evaluations should examine whether time spent in Advanced PA instruction meaningfully impacts intervention effects. It is possible that students who spent more time in Advanced PA gained fewer WCPM due to diminished time in the evidence-based fluency intervention, or students who spent less time in Advanced PA gained fewer WCPM due to rushed Advanced PA intervention procedures.

Third, the sample trimming process, while necessary for causal inference and internal validity, did decrease the power and generalizability of the present study. As described above, the excluded- and included schools differed insignificantly on observed covariates and the instrument used to exclude schools (share of students in each intervention) did not appear to correlate meaningfully with observed measures. Methodological demonstrations have also shown earnest sample trimming procedures to be valid methods for improving unbiased causal estimates (e.g., Sturmer et al., 2021). Regardless, future evaluations could plan for generalizability in experimental design (Tipton & Olsen, 2018).

Fourth, Hall and colleagues (2024) demonstrate convincingly that school level (i.e., level two) factors explain a non-trivial amount of variance observed in reading intervention growth. It is possible that school level demographic variables such as whole-school rates of FRL uptake may have sustained different environments in which interventions took place. It is most likely that any school level effects would influence both interventions similarly. However, this would be worthy of investigation if for no other reason than to increase estimate precision.

Finally, a Randomized Controlled Trial (RCT) may still be warranted to evaluate whether Advanced PA training, as a component, meaningfully affects reading outcomes. Although the present study used causally rigorous methodology to create unbiased estimates of treatment effects, RCTs remain the gold standard for establishing an evidence base for (or against) an

intervention (What Works Clearinghouse, 2022). Propensity weighting approaches like the one used in the present study may meet What Works Clearinghouse's research standards 'with reservations' (What Works Clearinghouse, 2022). An RCT may meet standards without reservations, assuming it is well designed, and no significant attrition occurs. A further benefit to conducting an RCT also remedies many of the limitations mentioned previously. A well designed RCT may systematically replicate these findings by standardizing the amount of fluency instruction at 20 minutes and adding a fixed amount of Advanced PA training for one group. More specifically, future RCT's may wish to compare Advanced PA instruction to (a) instruction in an unrelated skill (e.g., math) to examine effects relative to natural, uninstructed growth, (b) Basic PA, (c) PA instruction with the addition of graphemes, or (d) decoding instruction to precisely identify how any effects may relate to theoretically similar, but evidence-supported, interventions.

Conclusion

Advanced Phonemic Awareness (Advanced PA) skills are popular instructional targets (Literary Resources, 2023). Despite this popularity, critics have noted an absence of evidence supporting the efficacy of Advanced PA for improving text reading. The lone study empirically evaluating Advanced PA skills found that KG and Grade 1 students who received universal Advanced PA instruction improved on measures of Advanced PA skills, but not on measures of word reading and text reading. Proponents of Advanced PA have defended their contentions by arguing that (a) Advanced PA is beneficial for students who struggle to read more than those who respond to universal instruction, and (b) theory suggests that Advanced PA skills are necessary to improve orthographic mapping (Kilpatrick & O'Brien, 2019; Kilpatrick, 2020b). The present study evaluated these competing claims about the role of Advanced PA by

empirically examining the effect of any amount of Advanced PA training on text reading fluency among struggling second- and third-grade readers. Findings from the fixed- and random-effect models indicate that this group of students experienced significantly diminished reading fluency gains when receiving any amount of Advanced PA, compared to receiving only text reading fluency instruction (standard practice). Survival analysis indicates that students in both grades who received Advanced PA instruction were significantly less likely to exits Tier 2 instruction (Hazard Ratio = 0.22 in Grade 2, 0.35 in Grade 3) and those who did graduate took significantly longer periods of time to do so. Taken together, results show that Advanced PA does not appear effective for improving connected text reading for struggling readers. The present study provided evidence that directly undermines the theory-based hypothesis that Advanced PA skills relate meaningfully to orthographic mapping and therefore should improve connected text reading. Schools providing Advanced PA intervention to students with Tier 2 needs may consider replacing programming with interventions supported by broad research evidence and well matched to students' instructional needs.

References

Aaron, P. G., Joshi, R. M., Gooden, R., & Bentum, K. E. (2008). Diagnosis and treatment of reading disabilities based on the component model of reading: An alternative to the discrepancy model of LD. *Journal of Learning Disabilities, 41*, 67–84. https://doi.org/10.1177/0022219407310838

Amir-Behghadami, M., & Janati, A. (2020). Population, Intervention, Comparison, Outcomes and Study (PICOS) design as a framework to formulate eligibility criteria in systematic reviews. *Emergency Medicine Journal 37*(6), 387-387 https://doi.org/10.1136/emermed-2020-209567

Angrist, J. D., & Pischke, J. S. (2014). *Mastering' metrics: The path from cause to effect.* Princeton University Press.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. https://doi.org/10.1080/00273171.2011.568786

Baumann, J. F., Hoffman, J. V., Moon, J., & Duffy-Hester, A. M. (1998). Where are teachers' voices in the phonics/whole language debate? Results from a survey of US elementary classroom teachers. *The Reading Teacher, 51*(8), 636-650.

Bailey, P., Webb, B., Kelley, C., Nguyen, T., Huo, H. (2023). *Package 'WeMix'*. Weighted Mixed-Effects Models Using Multilevel Pseudo Maximum Likelihood Estimation. R package version 4.0.3, <https://CRAN.R-project.org/package=WeMix>.

Begeny, J. C., Laugle, K. M., Krouse, H. E., Lynn, A. E., Tayrose, M. P., & Stage, S. A. (2010). A control-group comparison of two reading fluency programs: The Helping Early Literacy with Practice Strategies (HELPS) program and the Great Leaps K-2 reading program. *School Psychology Review, 39*(1), 137-155. https://doi.org/10.1080/02796015.2010.12087795

Betts, E. A. (1946). Foundations of reading instruction, with emphasis on differentiated guidance. *American Book Co.*

Blair, G., Cooper, J., Coppock, A., Humphreys, M., Sonnet, L. (2022). _*estimatr*: Fast Estimators for Design-Based Inference_. R package version 1.0.0, <https://CRAN.R-project.org/package=estimatr>.

Bowey, J. A., & Hansen, J. (1994). The development of orthographic rimes as units of word recognition. *Journal of Experimental Child Psychology, 58*(3), 465-488. https://doi.org/10.1006/jecp.1994.1045

Burns, M. K. (2007). Reading at the instructional level with children identified as learning disabled: Potential implications for response-to-intervention. *School Psychology Quarterly*, *22*(3), 297. https://doi.org/10.1037/1045-3830.22.3.297

Burns, M. K., VanDerHeyden, A. M., & Boice, C. H. (2008). Best practices in delivery of intensive academic interventions. In A. Thomas & J. Grimes (Eds.). *Best Practices in School Psychology V* (pp. 1151-1162). National Association of School Psychologists.

Burns, M. K., Duke, N. K., & Cartwright, K. B. (2023). Evaluating components of the active view of reading as intervention targets: Implications for social justice. *School Psychology*, *38*(1), 30. https://doi.org/10.1037/spq0000519

Byrne, B., & Fielding-Barnsley, R. (1989). Phonemic awareness and letter knowledge in the child's acquisition of the alphabetic principle. *Journal of Educational Psychology*, *81*(3), 313. https://doi.org/10.1037/0022-0663.81.3.313

Byrne, B., & Fielding-Barnsley, R. (1993). Evaluation of a program to teach phonemic awareness to young children: A 1-year follow-up. *Journal of Educational Psychology, 85*(1), 104–111. https://doi.org/10.1037/0022-0663.85.1.104

Carlisle, J. F., & Stone, C. A. (2005). Exploring the role of morphemes in word reading. *Reading Research Quarterly*, *40*(4), 428-449. https://doi.org/10.1598/rrq.40.4.3

Castles, A., & Nation, K. (2010). How does orthographic learning happen?. *In 'From Inkmarks to Ideas' (Eds. B. Blachman; pp. 181-209)*. Psychology Press. https://doi.org/10.4324/9780203841211

Castles, A., Coltheart, M., Wilson, K., Valpied, J., & Wedgwood, J. (2009). The genesis of reading ability: What helps children learn letter–sound correspondences?. *Journal of Experimental Child Psychology, 104*(1), 68-88. https://doi.org/10.1016/j.jecp.2008.12.003

Castles, A., Rastle, K., & Nation, K. (2018). Ending the reading wars: Reading acquisition from novice to expert. *Psychological Science in the Public Interest*, *19*(1), 5-51. https://doi.org/10.1177/1529100618772271

Cho, E., Ju, U., Kim, E. H., Lee, M., Lee, G., & Compton, D. L. (2023). Relations among motivation, executive functions, and reading comprehension: Do they differ for students with and without reading difficulties?. *Scientific Studies of Reading, 27*(4), 289-310. https://doi.org/10.1080/10888438.2022.2127357

Christ, T. J., Arañas, Y. A., Johnson, L., Kember, J. M., Kilgus, S., Kiss, J. K., McCarthy, Trentman, A. M., Monaghen, B. D., Nelson, G., Nelson, P., Newell, K. W., Van Norman, E. R., White, M. J., Windram, H.(2018). *Formative assessment system for teachers technical manual*. Minneapolis, MN: Author and FastBridge Learning.

Clemens, NH., Solari, E. J., Kearns, D. M., Fien, H., Nelson, N. J., Stalega, M. V., Burns, M. K., St. Martin, K., & Heoft, F. (2021). *They say you can do phonemic awareness instruction "in the dark", but should you? A critical evaluation of the trend toward advanced phonemic awareness training.* PsyArXiv. https://doi.org/10.31234/osf.io/ajxbv

Closson, T. (2023). New York is Forcing Schools to Change How They Teach Children to Read. *The New York Times.* https://www.nytimes.com/2023/05/09/nyregion/reading-nyc-schools.html

Covington, N. (2023). *Unsettling the science of reading: who is being sold a story?* [Web Article]. Accessed via: https://www.humanrestorationproject.org/writing/who-is-being-sold-a-story-unsettling-the-science-of-reading on April 6th, 2024.

Coyne, M.D., McCoach, D., B., Santoro, L. E., & Kastner, P. (2021). *Evaluating the effects of advanced phonemic awareness instruction in first grade.* Manuscript in preparation.

Cummings, K. D., Kaminski, R. A., Good III, R. H., & O'Neil, M. (2011). Assessing phonemic awareness in preschool and kindergarten: Development and initial validation of first sound fluency. *Assessment for Effective Intervention, 36*(2), 94-106. https://doi.org/10.1177/1534508410392209

Daly III, E. J., Lentz Jr, F. E., & Boyer, J. (1996). The Instructional Hierarchy: A conceptual model for understanding the effective components of reading interventions. *School Psychology Quarterly, 11*(4), 369. https://doi.org/10.1037/h0088941

Daly III, E. J., Bonfiglio, C. M., Mattson, T., Persampieri, M., & Foreman-Yates, K. (2005). Refining the experimental analysis of academic skills deficits: Part I. An investigation of variables that affect generalized oral reading performance. *Journal of Applied Behavior Analysis, 38*(4), 485-497. https://doi.org/10.1901/jaba.2005.113-04

Denckla, M. B., & Rudel, R. G. (1976). Rapid 'automatized'naming (RAN): Dyslexia differentiated from other learning disabilities. *Neuropsychologia*, *14*(4), 471-479. https://doi.org/10.1016/0028-3932(76)90075-0

Department for Education. (2023). *The Reading Framework*. London, UK. Department for Education

Duke, N. K., & Cartwright, K. B. (2021). The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly, 56*, S25-S44. https://doi.org/10.1002/rrq.411

Ehri, L. C., & Wilce, L. S. (1985). Movement into reading: Is the first stage of printed word learning visual or phonetic?. *Reading Research Quarterly, 20*(2), 163-179. https://doi.org/10.2307/747753

Ehri, L.C., & Wilce, L.S. (1987a). Cipher versus cue reading: An experiment in decoding acquisition. *Journal of Educational Psychology, 79*(1), 3–13. https://doi.org/10.1037/0022-0663.79.1.3

Ehri, L. C. (1997). Sight word learning in normal readers and dyslexics. In B. A. Blachman (Ed.). *Foundations of reading acquisition and dyslexia: Implications for Early Intervention* (pp. 163-189). https://doi.org/10.4324/9781410601230

Ehri, L. C., Nunes, S. R., Stahl, S. A., & Willows, D. M. (2001). Systematic phonics instruction helps students learn to read: Evidence from the National Reading Panel's meta-analysis. *Review of Educational Research*, *71*(3), 393-447. https://doi.org/10.3102/00346543071003393

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly*, *36*(3), 250-287. https://doi.org/10.1598/rrq.36.3.2

Ehri, L. C. (2005). Development of sight word reading: Phases and findings. *The Science of Reading: A Handbook, 135–154.* https://doi.org/10.1002/9780470757642.ch8

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of Reading, 9*(2), 167-188. https://doi.org/10.1207/s1532799xssr0902_4

Ehri, L. C. (2014). Orthographic mapping in the acquisition of sight word reading, spelling memory, and vocabulary learning. *Scientific Studies of Reading*, *18*(1), 5-21. https://doi.org/10.1080/10888438.2013.819356

Ehri, L. C. (2020). The science of learning to read words: A case for systematic phonics instruction. *Reading Research Quarterly, 55*, S45-S60. https://doi.org/10.1002/rrq.334

Every Student Succeeds Act, Pub. L. No. 114-95, 129 Stat. 1802 (2015).

Fien, H., Chard, D.J., & Baker, S.K. (2021). Can the Evidence Revolution and Multi-Tiered Systems of Support Improve Education Equity and Reading Achievement?. *Reading Research Quarterly*, *56*(S1), S105–S118. https://doi.org/10.1002/rrq.391

Florit, E., & Cain, K. (2011). The simple view of reading: Is it valid for different types of alphabetic orthographies?. *Educational Psychology Review, 23*, 553-576. https://doi.org/10.1007/s10648-011-9175-6

Foulin, J., N. (2005). Why is letter-name knowledge such a good predictor of learning to read?. *Reading and Writing, 18*, 129-155. https://doi.org/10.1007/s11145-004-5892-2

Fountas, I. C., & Pinnell, G. S. (2012). Guided reading: The romance and the reality. *The Reading Teacher*, *66*(4), 268-284. https://doi.org/10.1002/trtr.01123

Freeman, R., Miller, D., & Newcomer, L. (2015). Integration of academic and behavioral MTSS at the district level using implementation science. *Learning Disabilities: A Contemporary Journal, 13*(1), 59-72.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance, 13*(1), 104. https://doi.org/10.1037/0096-1523.13.1.104

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. https://doi.org/10.1177/0014402917693580

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2012). Smart RTI: A next-generation approach to multilevel prevention. *Exceptional Children, 78*(3), 263-279. https://doi.org/10.1177/001440291207800301

Girard, S. (2023) Phonics mandate: What to know about a new Wisconsin reading bill. *The Cap Times. https://captimes.com/news/education/phonics-mandate-what-to-know-about-a-new-wisconsin-reading-bill/article_6d4faeb5-753a-5eaf-9cc0-3d8a2bb6c64e.html*

Goodman, K. S. (1967). Reading: A psycholinguistic guessing game. *Literacy Research and Instruction, 6*(4), 126-135. https://doi.org/10.1080/19388076709556976

Gough, P. and Tunmer, W. (1986). Decoding, reading, and reading disability. *Remedial and Special Education*, *7*,6-10. https://doi.org/10.1177/074193258600700104

Gough, P. B., Juel, C., & Griffith, P. L. (1992). Reading, spelling, and the orthographic cipher. *Reading Acquisition*, 35-48. https://doi.org/10.4324/9781351236904-2

Greifer, N. (2023). *WeightIt*: Weighting for Covariate Balance in Observational Studies (version 0.14.2). https://CRAN.R-project.org/package=WeightIt.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis, 20*(1), 25-46. https://doi.org/10.1093/pan/mpr025

Hammerschmidt-Snidarich, S. M., Wagner, D., Parker, D. C., & Wagner, K. (2021). Reading Tutors' Interpretation of Curriculum-Based Measurement Graphs. *Assessment for Effective Intervention, 47*(1), 26-36. https://doi.org/10.1177/1534508420963193

Hanford, E. (2022). *Sold a Story.* American Public Media. https://features.apmreports.org/sold-a-story/

Haring N. G., Eaton M. D. (1978). Systematic instructional procedures: An instructional hierarchy. In Haring N. G., Lovitt T. C., Eaton M. D., Hansen C. L. (Eds.), *The Fourth R: Research in the Classroom* (pp. 23–40). Merrill.

Hernán, M.A., & Robins, J.M. (2020). *Causal Inference: What If.* Chapman & Hall/CRC. https://www.hsph.harvard.edu/miguelHernan/causal-inference-book/

Hess, C. (2023) Gov. Tony Evers signs sweeping reading literacy bill into law. *Wisconsin Public Radio*. https://www.wpr.org/evers-signs-science-reading-literacy-bill-law

Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161-1189. https://doi.org/10.1111/1468-0262.00442

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2,* 127-160. https://doi.org/10.1007/bf00401799

Hoover, W. A., & Tunmer, W. E. (2018). The simple view of reading: Three assessments of its adequacy. *Remedial and Special Education, 39*(5), 304-312. https://doi.org/10.1177/0741932518773154

Hoover, W. A., & Tunmer, W. E. (2022). The primacy of science in communicating advances in the science of reading. *Reading Research Quarterly, 57*(2), 399-408. https://doi.org/10.1002/rrq.446

Houwelingen, H. C., Stijnen, T. (2013) Cox regression model. In Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (Eds.) *Handbook of Survival Analysis.* (pp. 5-25). CRC Press. https://doi.org/10.1201/b16248-8

Hudson, T. M., & McKenzie, R. G. (2016). Evaluating the use of RTI to identify SLD: A survey of state policy, procedures, data collection, and administrator perceptions. *Contemporary School Psychology, 20*, 31-45., https://doi.org/10.1007/s40688-015-0081-7

Hudson, A., Koh, P. W., Moore, K. A., & Binks-Cantrell, E. (2020). Fluency interventions for elementary students with reading difficulties: A synthesis of research from 2000–2019. *Education Sciences, 10*(3), 52. https://doi.org/10.3390/educsci10030052

Hund, A. M., Bove, R. M., & Van Beuning, N. (2023). Cognitive flexibility explains unique variance in reading comprehension for elementary students. *Cognitive Development, 67*, 101358. https://doi.org/10.1016/j.cogdev.2023.101358

Imbens, G. W., & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press. https://doi.org/10.1017/cbo9781139025751.001

Jimerson, S. R., Burns, M. K., & VanDerHeyden, A. M. (Eds.). (2016). *Handbook of response to intervention: The science and practice of multi-tiered systems of support*. Springer. https://doi.org/10.1007/978-1-4899-7568-3_1

Kendeou, P., Van den Broek, P., White, M. J., & Lynch, J. S. (2009a). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology, 101*(4), 765. https://doi.org/10.1037/a0015956

Kendeou, P., Savage, R., & van den Broek, P. (2009b). Revisiting the simple view of reading. *British Journal of Educational Psychology, 79*(2), 353-370. https://doi.org/10.1348/978185408x369020

Kilgus, S. P., Methe, S. A., Maggin, D. M., & Tomasula, J. L. (2014). Curriculum-based measurement of oral reading (R-CBM): A diagnostic test accuracy meta-analysis of evidence supporting use in universal screening. *Journal of School Psychology*, *52*(4), 377-405. https://doi.org/10.1016/j.jsp.2014.06.002

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

Kilpatrick, D. A., & O'Brien, S. (2019). Effective prevention and intervention for word-level reading difficulties. In D.K. Kilpatrick, R.M. Joshi, & R.K. Wagner (Eds.) *Reading Development and Difficulties* (pp. 179-210). Springer. https://doi.org/10.1007/978-3-030-26550-2_8

Kilpatrick, D. A. (2020a). *Equipped for reading success: A comprehensive, step-by-step program for developing phoneme awareness and fluent word recognition*. Casey & Kirsch Publishers.

Kilpatrick, D. A. (2020b). How the phonology of speech is foundational for instant word recognition. *Perspectives on Language and Literacy, 46*(3), 11-15. https://literacyhow.org/wp-content/uploads/2020/09/The-Phonology-of-Speech-in-WR-Kilpatrick.pdf

Kilpatrick, D. A., Ashby, J., Naftel, S., Moats, L. C. (2022) *A Response to the Clemens et al., (2021) Manuscript Entitled, "They Say You Can Do Phonemic Awareness Instruction "In the Dark," But Should You? A Critical Evaluation of the Trend Toward Advanced Phonemic Awareness Training"* *https://osf.io/csu8h?view_only=65f6f1a6099349a49b7697de00c7a77b*

Kim, Y. S., Park, C. H., & Wagner, R. K. (2014). Is oral/text reading fluency a "bridge" to reading comprehension?. *Reading and Writing, 27*, 79-99. https://doi.org/10.1007/s11145-013-9434-7

Kim, Y. S. G., & Wagner, R. K. (2015). Text (oral) reading fluency as a construct in reading development: An investigation of its mediating role for children from grades 1 to 4. *Scientific Studies of Reading, 19*(3), 224-242. https://doi.org/10.1080/10888438.2015.1007375

Kim, Y. S. G. (2017). Why the simple view of reading is not simplistic: Unpacking component skills of reading using a direct and indirect effect model of reading (DIER). *Scientific Studies of Reading, 21*(4), 310-333. https://doi.org/10.1080/10888438.2017.1291643

Kim, Y. S. G. (2020a). Toward integrative reading science: The direct and indirect effects model of reading. *Journal of Learning Disabilities, 53*(6), 469-491. https://doi.org/10.1177/0022219420908239

Kim, Y. S. G. (2020b). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667. https://doi.org/10.1037/edu0000407

Kim, Y. S. G., Quinn, J. M., & Petscher, Y. (2021). What is text reading fluency and is it a predictor or an outcome of reading comprehension? A longitudinal investigation. *Developmental Psychology, 57*(5), 718. https://doi.org/10.1037/dev0001167

Kirby, J. R., & Savage, R. S. (2008). Can the simple view deal with the complexities of reading?. *Literacy, 42*(2), 75-82. https://doi.org/10.1111/j.1741-4369.2008.00487.x

Klein, J. P., Van Houwelingen, H. C., Ibrahim, J. G., & Scheike, T. H. (Eds.). (2016). *Handbook of Survival Analysis.* CRC Press.

Klingbeil, D. A., Nelson, P. M., Van Norman, E. R., & Birr, C. (2017). Diagnostic accuracy of multivariate universal screening procedures for reading in upper elementary grades. *Remedial and Special Education, 38*(5), 308-320. https://doi.org/10.1177/0741932517697446

Klingbeil, D. A., Latham, A. D., Schmitt, M. C., & Kim, J. S. (2022, August 4-6). *Toward an evidence-based assessment approach to academic screening in schools* [Poster presentation]. American Psychological Association Annual Convention, Minneapolis, MN, United States.

Kovaleski, J. F., VanDerHeyden, A. M., Runge, T. J., Zirkel, P. A., & Shapiro, E. S. (2022). *The RTI approach to evaluating learning disabilities.* Guilford.

Kullback, S. (1959). *Information Theory and Statistics.* John Wiley & Sons.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*(2), 293-323. https://doi.org/10.1016/0010-0285(74)90015-2

Landerl, K., Freudenthaler, H. H., Heene, M., De Jong, P. F., Desrochers, A., Manolitsis, G., ... & Georgiou, G. K. (2019). Phonological awareness and rapid automatized naming as longitudinal predictors of reading in five alphabetic orthographies with varying degrees of consistency. *Scientific Studies of Reading, 23*(3), 220-234. https://doi.org/10.1080/10888438.2018.1510936

Lee, J., & Yoon, S. Y. (2017). The effects of repeated reading on reading fluency for students with reading disabilities: A meta-analysis. *Journal of Learning Disabilities, 50*(2), 213-224. https://doi.org/10.1177/0022219415605194

Lin, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics, 7*(1). https://doi.org/10.1214/12-aoas583

Literacy Resources. (2023). *Heggerty. About Us.* [Webpage] Accessed via: *https://heggerty.org/about-us/* on April 4th 2024.

Lonigan, C. J., & Burgess, S. R. (2017). Dimensionality of reading skills with elementary-school-age children. *Scientific Studies of Reading, 21*(3), 239-253. https://doi.org/10.1080/10888438.2017.1285918

Lonigan, C. J., Burgess, S. R., & Schatschneider, C. (2018). Examining the simple view of reading with elementary school children: Still simple after all these years. *Remedial and Special Education, 39*(5), 260-273. https://doi.org/10.1177/0741932518764833

Machin S., McNally S., & Viarengo M. (2016). "*Teaching to teach" literacy* (Centre for Economic Performance Discussion Paper No. 1425). http://cep.lse.ac.uk/pubs/download/dp1425.pdf

Maki, K. E., Zaslofksy, A. F., Knight, S., Ebbesmeyer, A. M., & Chelmo-Boatman, A. (2021). Intervening with multiplication fact difficulties: Examining the utility of the instructional hierarchy to target interventions. *Journal of Behavioral Education, 30*, 534-558. https://doi.org/10.1007/s10864-020-09388-0

Maki, K. E., & Adams, S. R. (2020). Specific learning disabilities identification: Do the identification methods and data matter?. *Learning Disability Quarterly, 43*(2), 63-74. https://doi.org/10.1177/0731948719826296

Markovitz, C. E., Hernández, M. W., Hedberg, E. C., & Whitmore, H. W. (2022). Evaluating the effectiveness of a volunteer one-on-one tutoring model for early elementary reading intervention: A randomized controlled trial replication study. *American Educational Research Journal, 59*(4), 788-819. https://doi.org/10.3102/00028312211066848

Marsh, G., Friedman, M., Desberg, P., & Saterdahl, K. (1981). Comparison of reading and spelling strategies in normal and reading disabled children. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.). *Intelligence and Learning* (pp. 363-367). https://doi.org/10.1007/978-1-4684-1083-9_33

Melby-Lervåg, M., Lyster, S. A. H., & Hulme, C. (2012). Phonological skills and their role in learning to read: a meta-analytic review. *Psychological Bulletin, 138*(2), 322-340. https://doi.org/10.1037/a0026744

Metsala, J. L., & Ehri, L. C. (Eds.). (2013). *Word recognition in beginning literacy*. Routledge. https://doi.org/10.4324/9781410602718

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific*

research literature on reading and its implications for reading instruction: Reports of the subgroups. National Institute of Child Health and Human Development, National Institutes of Health.

Neuman, S. B., Quintero, E., & Reist, K. (2023). *Reading Reform Across America: A Survey of State Legislation.* https://www.shankerinstitute.org/sites/default/files/2023-07/ReadingReform%20ShankerInstitute%20FullReport.pdf

Nixon, N., Ruel, T., Ackerman, B., Coburn, K., Chao, B., Tipton, B. (2024*). generalizeR: Design a Sample Recruitment Plan and Assess Its 'Generalizability' to Broader Populations*. R package version 0.0.1, https://github.com/NUstat/generalizeR

Parker, D. C., & Burns, M. K. (2014). Using the instructional level as a criterion to target reading interventions. *Reading & Writing Quarterly, 30*(1), 79-94. https://doi.org/10.1080/10573569.2012.702047

Parker, D. C. (2023) *Personal Communication.* May 19, 2023.

Parker, S. (2022). *The Kilpatrick Conundrum*. https://www.parkerphonics.com/post/the-kilpatrick-conundrum

Pearson, P. D., Palincsar, A. S., Biancarosa, G., & Berman, A. I. (2020). *Reaping the Rewards of the Reading for Understanding Initiative.* National Academy of Education. https://doi.org/10.31094/2020/2

Peng, P., & Goodrich, J. M. (2020). The cognitive element model of reading instruction. *Reading Research Quarterly, 55*, S77-S88. https://doi.org/10.1002/rrq.336

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading,11*(4), 357-383. https://doi.org/10.1080/10888430701530730

Perfetti, C. A., & Hart, L. (2002). The lexical quality hypothesis. *Precursors of Functional Literacy, 11,* 67-86. https://doi.org/10.1075/swll.11.14per

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review, 114*, 273–315. https://doi.org/10.1037/0033-295X.114.2.273.

Petscher, Y., Cabell, S., Catts, H. W., Compton, D., Foorman, B., Hart, S. A., Lonigan, C., Phillips, B., Schatschneider, C., Steacy, L. M., Terry, N. P., & Wagner, R. (2020). How the Science of Reading Informs 21st Century Education. *Reading Research Quarterly. 55*(1), S267-S282. https://doi.org/10.31234/osf.io/yvp54

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review, 103*(1), 56. https://doi.org/10.7551/mitpress/1888.003.0029

Rayner, K., Foorman, B. R., Perfetti, C. A., Pesetsky, D., & Seidenberg, M. S. (2002). How should reading be taught?. *Scientific American, 286*(3), 84-91. https://doi.org/10.1038/scientificamerican0302-84

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology, 47*(6), 427-469. https://doi.org/10.1016/j.jsp.2009.07.001

Roberts, T. (2003). Effects of alphabet letter instruction on young children's word recognition. *Journal of Educational Psychology, 95*(1), 41–51. https://doi.org/10.1037/0022-0663.95.1.41

Rose J. (2006). *Independent review of the teaching of early reading final report.* U.K. Department for Education and Skills. http://dera.ioe.ac.uk/5551/2/report.pdf

Samuels, S. J. (1976). Automatic decoding and reading comprehension. *Language Arts, 53*(3), 323-325.

Scarborough, H. S., Ehri, L. C., Olson, R. K., & Fowler, A. E. (1998). The fate of phonemic awareness beyond the elementary school years. *Scientific Studies of Reading, 2*(2), 115-142. https://doi.org/10.1207/s1532799xssr0202_2

Schwartz, S. (2019). *A Comparative Analysis of Student Achievement of First Grade Students Using Fundations vs. Heggerty and Words Their Way. https://www.proquest.com/docview/2334214249?pq-origsite=gscholar&fromopenview=true*

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*(4), 523. https://doi.org/10.1037/0033-295x.96.4.523

Seidenberg, M. (2017). *Language at the speed of sight: How we read, why so many cannot, and what can be done about it.* Basic Books.

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

Shanahan, T. (2020). What constitutes a science of reading instruction?. *Reading Research Quarterly*, *55*, S235-S247. https://doi.org/10.1002/rrq.349

Shapiro, E. S., & Clemens, N. H. (2023) *Academic Skills Problems.(Fifth Edition)*. Guilford.

Share, D. L. (1995). Phonological recoding and self-teaching: Sine qua non of reading acquisition. *Cognition*, *55*(2), 151-218. https://doi.org/10.1016/0010-0277(94)00645-2

Skinner, C. H., & Daly, E. J. (2010). Improving generalization of academic skills: Commentary on the special issue. *Journal of Behavioral Education, 19*, 106-115. https://doi.org/10.1007/s10864-010-9100-y

Snowling, M. J., & Melby-Lervåg, M. (2016). Oral language deficits in familial dyslexia: A meta-analysis and review. *Psychological Bulletin, 142*(5), 498-510. https://doi.org/10.1037/bul0000037

Stanovich, K. E. (1981). Relationships between word decoding speed, general name-retrieval ability, and reading progress in first-grade children. *Journal of Educational Psychology*, *73*(6), 809. https://doi.org/10.1037/0022-0663.73.6.809

Stoiber, K. C., & Gettinger, M. (2016). Multi-tiered systems of support and evidence-based practices. In S.R. Jimerson, M. K. Burns, & A. M. VanDerHeyden (Eds.). *Handbook of response to intervention: The science and practice of multi-tiered systems of support* (pp. 121-141). Springer. https://doi.org/10.1007/978-1-4899-7568-3_9

Student Achievement Partners. (2020). *Comparing Reading Research to Program Design. An Examination of Teachers College Units of Study.* Accessed via

https://achievethecore.org/page/3240/comparing-reading-research-to-program-design-an-examination-of-teachers-college-units-of-study on November 30, 2023.

Stürmer, T., Webster-Clark, M., Lund, J. L., Wyss, R., Ellis, A. R., Lunt, M., Rothman, K. J., & Glynn, R. J. (2021). Propensity Score Weighting and Trimming Strategies for Reducing Variance and Bias of Treatment Effect Estimates: A Simulation Study. *American Journal of Epidemiology. 190*(8), 1659-1670. https://doi.org/10.1093/aje/kwab041

Szadokierski, I., Burns, M. K., & McComas, J. J. (2017). Predicting intervention effectiveness from reading accuracy and rate measures through the instructional hierarchy: Evidence for a skill-by-treatment interaction. *School Psychology Review,* 46(2), 190-200. https://doi.org/10.17105/spr-2017-0013.v46-2

Therneau, T. (2022). *survival*: A Package for Survival Analysis in R_. R package version 3.4-0, <https://CRAN.R-project.org/package=survival>.

Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., & Rapp, D. (2009). Simple but complex: Components of the simple view of reading across grade levels. *Journal of Research in Reading, 32*(4), 383-401. https://doi.org/10.1111/j.1467-9817.2009.01401.x

Tipton, E. & Miller, K. (2024). *The Generalizer.* Webtool hosted at https://thegeneralizer.org.

Tipton, E., & Olsen, R. B. (2018). A review of statistical methods for generalizing from evaluations of educational interventions. *Educational Researcher, 47*(8), 516-524. https://doi.org/10.3102/0013189x18781522

Tipton, E., & Olsen, R. B. (2022). Enhancing the Generalizability of Impact Studies in Education. Toolkit. NCEE 2022-003. National Center for Education Evaluation and Regional Assistance.

Torgesen, J. K., Alexander, A. W., Wagner, R. K., Rashotte, C. A., Voeller, K. K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34*(1), 33-58. https://doi.org/10.1177/002221940103400104

Torgesen, J. K. (2002). The prevention of reading difficulties. *Journal of School Psychology, 40*(1), 7-26. https://doi.org/10.1016/s0022-4405(01)00092-9

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *Test of word reading efficiency, second edition*. Austin, TX: PRO-ED, Inc.

Torgerson C., Brooks G., Hall J. (2006). *A systematic review of the research literature on the use of phonics in the teaching of reading and spelling* (Research Report RR711). U.K. Department for Education and Skills. http://dera.ioe.ac.uk/14791/1/RR711_.pdf

Treptow, M. A., Burns, M. K., & McComas, J. J. (2007). Reading at the frustration, instructional, and independent levels: The effects on students' reading comprehension and time on task. *School Psychology Review, 36*(1), 159-166. https://doi.org/10.1080/02796015.2007.12087958

Truch, S. (1994). Stimulating basic reading processes using auditory discrimination in depth. *Annals of Dyslexia, 44*(1), 60-80. https://doi.org/10.1007/bf02648155

Tübbicke, S. (2021). Entropy balancing for continuous treatments. *Journal of Econometric Methods, 11*(1), 71-89. https://doi.org/10.1515/jem-2021-0002

Tunmer, W. E., & Hoover, W. A. (2019). The cognitive foundations of learning to read: A framework for preventing and remediating reading difficulties. *Australian Journal of Learning Difficulties*, *24(*1), 75-93. https://doi.org/10.1080/19404158.2019.1614081

University of Florida Literacy Institute. (2023). *What is UFLI Foundations?*. Webpage accessed via https://ufli.education.ufl.edu/foundations/ on November 30, 2023.

U.S. Department of Education. Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2022 Reading Assessment. *Accessed via https://www.nationsreportcard.gov/reading/nation/achievement/?grade=4 on November 30, 2023*

VanDerHeyden, A. M., & Burns, M. K. (2010). *Essentials of response to intervention (Vol. 79).* John Wiley & Sons

VanDerHeyden, A. M., Kovaleski, J. F., Shapiro, E. S., & Painter, D. T. (2014). Scientifically supported identification of SLD using RTI: A response to Colker. *Journal of Law and Education, 43,* 229.

Van Norman, E. R., Nelson, P. M., & Parker, D. C. (2018). Curriculum-based measurement of reading decision rules: Strategies to improve the accuracy of treatment recommendations. *School Psychology Review, 47*(4), 333-344. https://doi.org/10.17105/spr-2017-0089.v47-4

Van Norman, E. R., Nelson, P. M., & Klingbeil, D. A. (2020). Profiles of reading performance after exiting tier 2 intervention. *Psychology in the Schools, 57*(5), 757-767. https://doi.org/10.1002/pits.22354

Vaughn, S., & Fletcher, J. M. (2012). Response to intervention with secondary school students with reading difficulties. *Journal of Learning Disabilities, 45*(3), 244-256. https://doi.org/10.1177/0022219412442157

Wang, A., Nianogo, R. A., & Arah, O. A. (2017). G-computation of average treatment effects on the treated and the untreated. *BMC Medical Research Methodology, 17*, 1-5. https://doi.org/10.1186/s12874-016-0282-4

Wanzek, J., Vaughn, S., Scammacca, N., Gatlin, B., Walker, M. A., & Capin, P. (2016). Meta-analyses of the effects of tier 2 type reading interventions in grades K-3. *Educational Psychology Review, 28,* 551-576. https://doi.org/10.1007/s10648-015-9321-7

Wayman, M. M., Wallace, T., Wiley, H. I., Tichá, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*(2), 85-120. https://doi.org/10.1177/00224669070410020401

What Works Clearinghouse (2022). Procedures and Standards Handbook. https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5_0-0-508.pdf

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for the developmental dyslexias. *Journal of Educational Psychology, 91*(3), 415-438. https://doi.org/10.1037/0022-0663.91.3.415

Wyse, D., & Styles, M. (2007). Synthetic phonics and the teaching of reading: the debate surrounding England's 'Rose Report'. *Literacy, 41*(1), 35-42. https://doi.org/10.1111/j.1467-9345.2007.00455.x

Young, C., & Rasinski, T. (2009). Implementing readers theatre as an approach to classroom fluency instruction. *The Reading Teacher, 63*(1), 4-13. https://doi.org/10.1598/rt.63.1.1

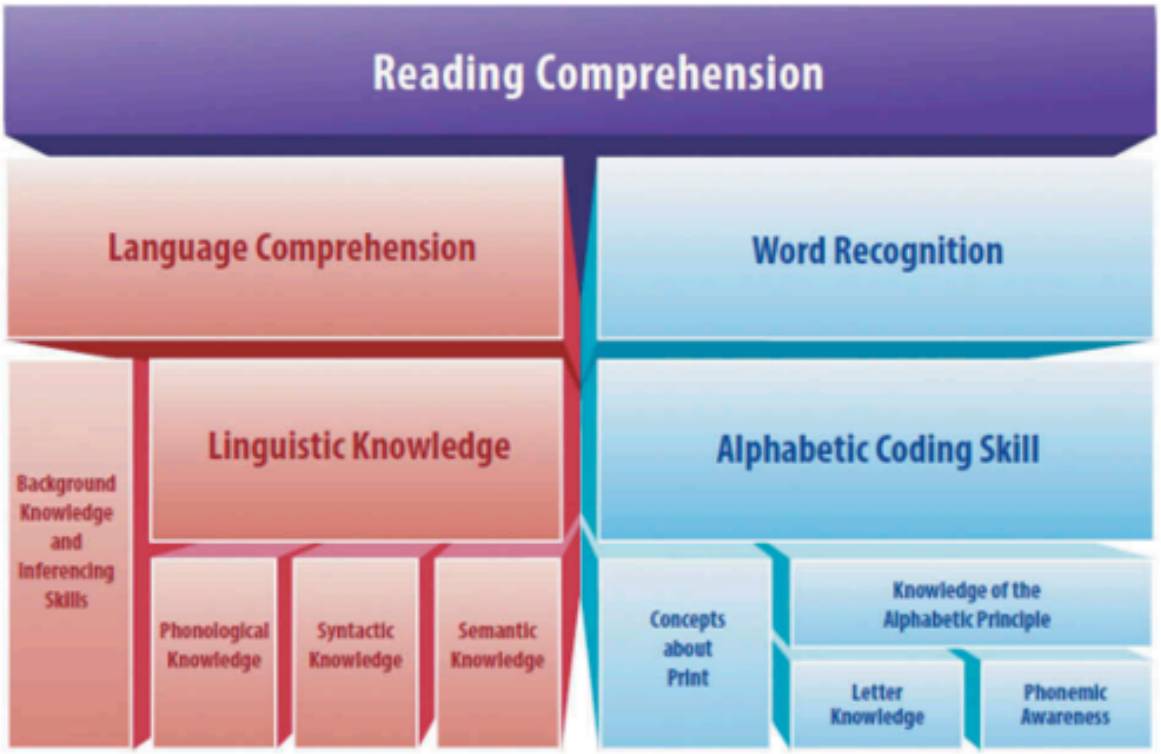


Figure 1. The Cognitive Foundations Framework (Tunmer & Hoover, 2019), an updated framework of the original Simple View of Reading (SVR). Reproduced from Tunmer & Hoover (2019).

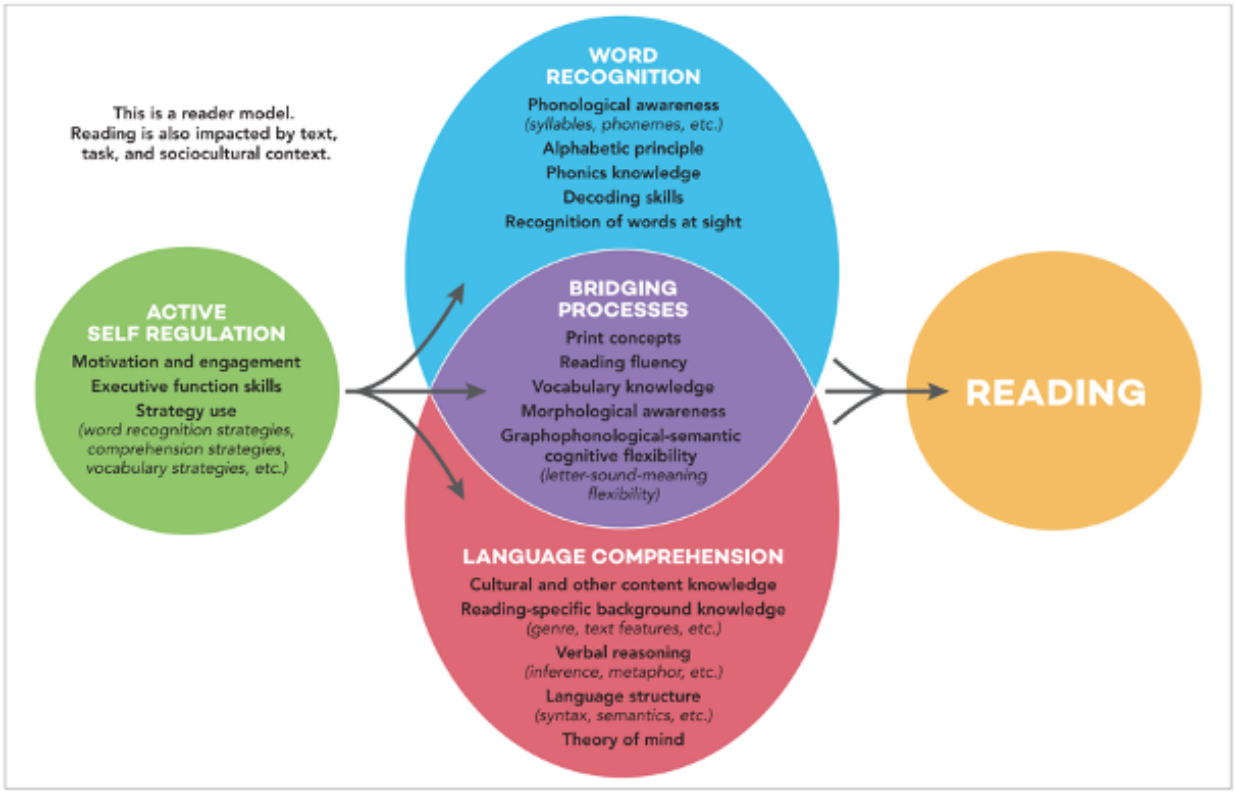


Figure 2. The Active View of Reading (Duke & Cartwright, 2021).

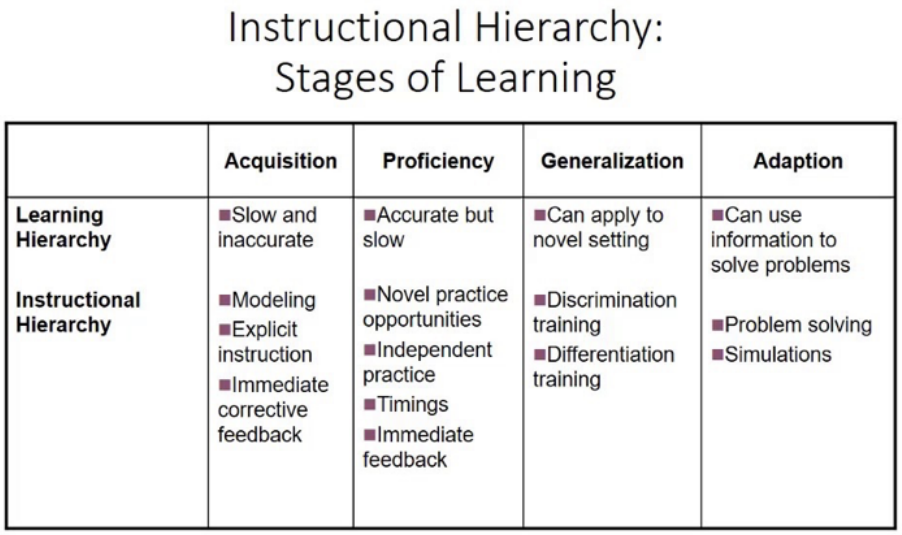


Figure 3. The Instructional Hierarchy (IH; Haring & Eaton, 1978). Bullet points at each stage describe expected demonstration of the target skill as well as indicated instructional methods.

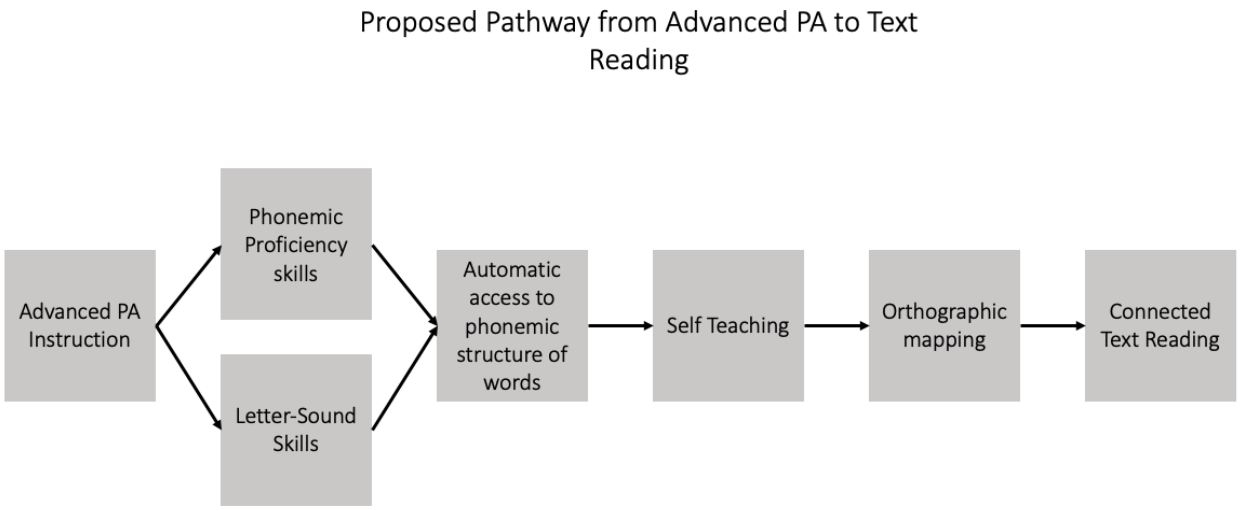


Figure 4. A mechanistic diagram of the theorized influence of Advanced PA instruction on Text Reading. This diagram amalgamates conjectures from Kilpatrick & O'Brien (2019) and Kilpatrick (2020b).

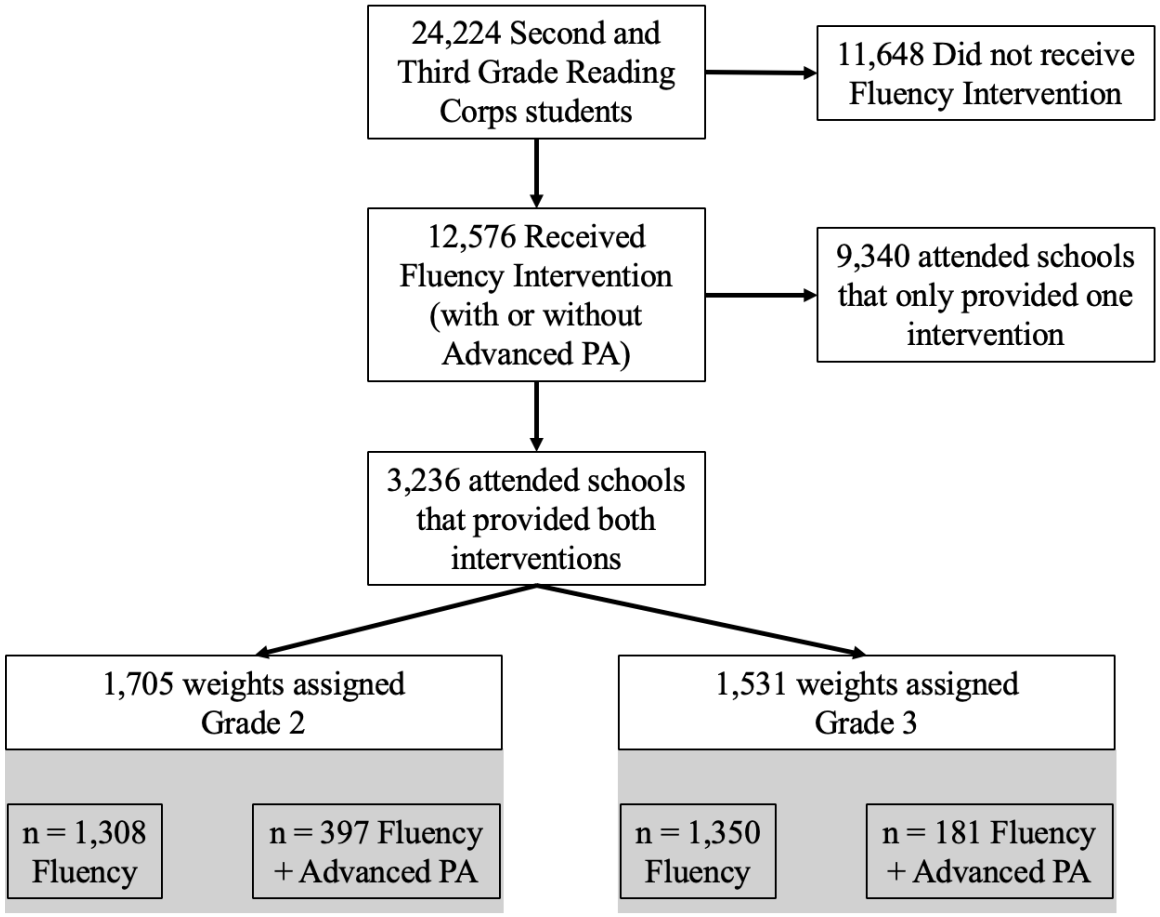


Figure 5. PRISMA diagram outlining the distillation of original sample into final, trimmed, analytic sample.

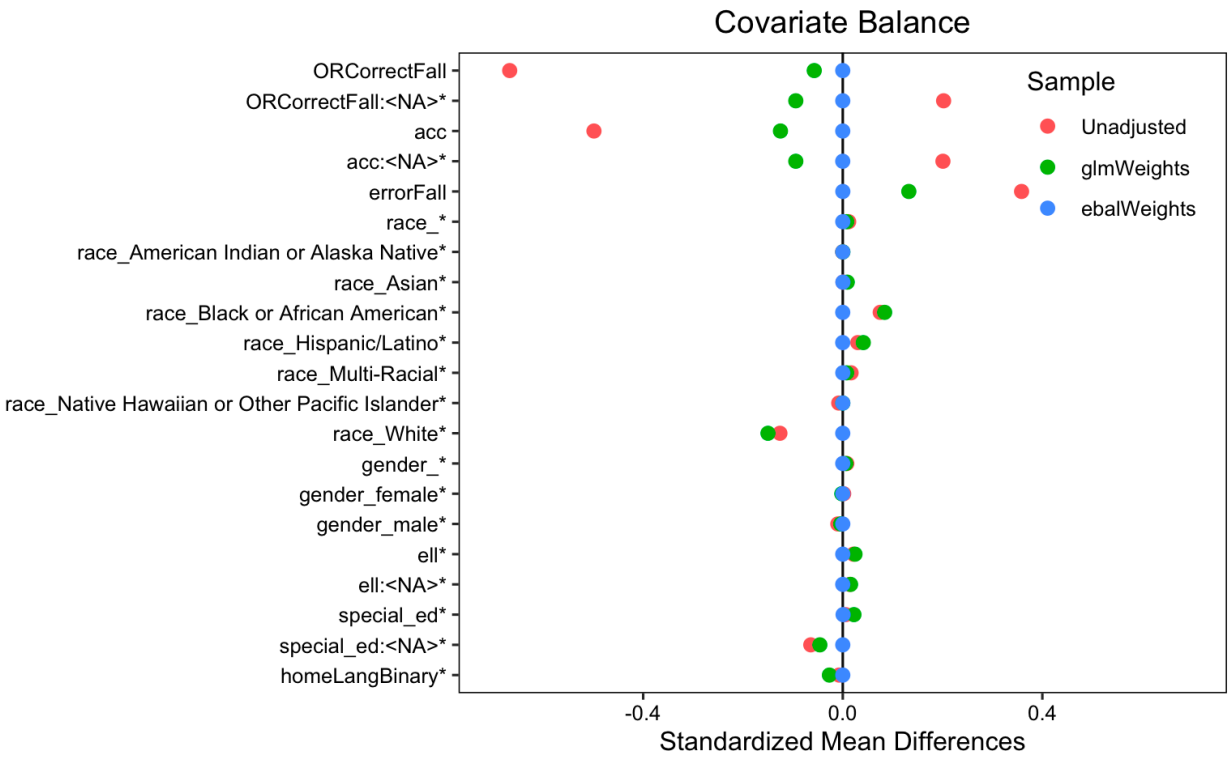
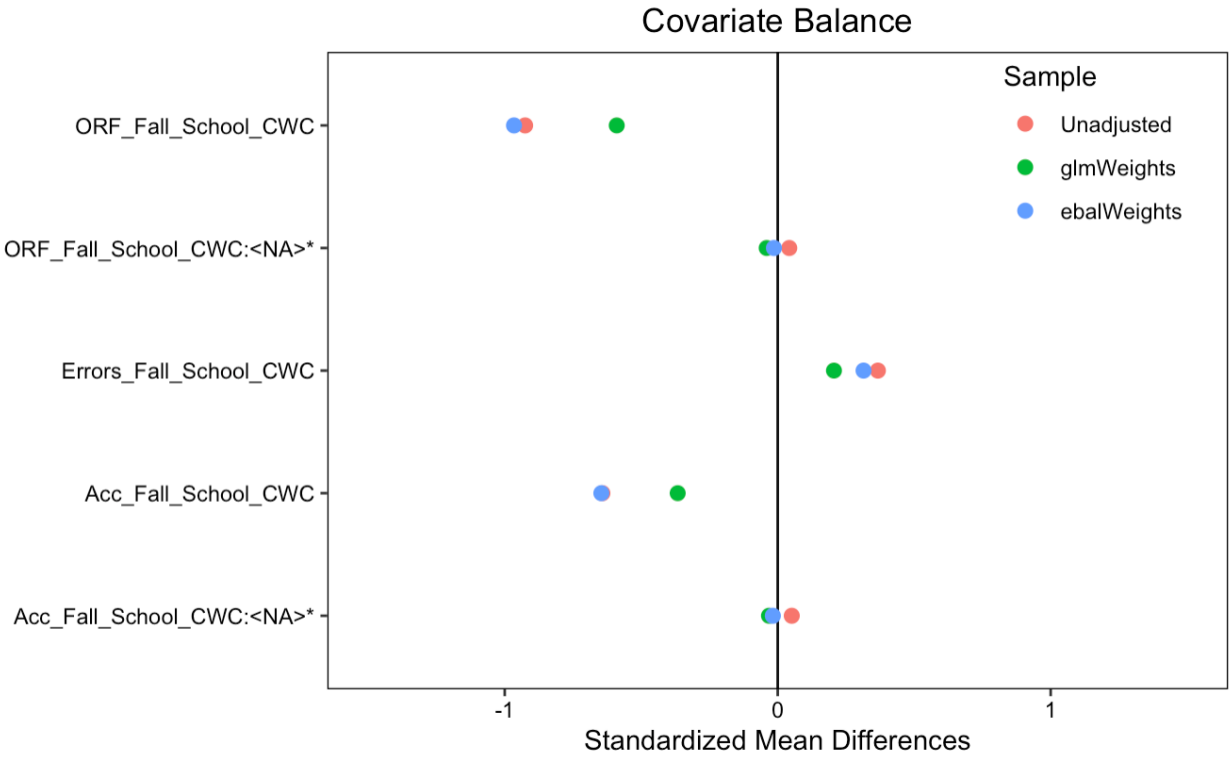


Figure 6a. Love plot detailing standard mean differences between FL and FLPA experimental groups (a) unweighted (red), (b) weighted with GLM weights (green), and (c) weighted with entropy balancing weights (blue).

Figure 6b. Love plot detailing standard mean differences of centered-within-cluster means between experimental groups (a) unweighted (red), (b) weighted with GLM weights (green), and (c) weighted with entropy balancing weights (blue).



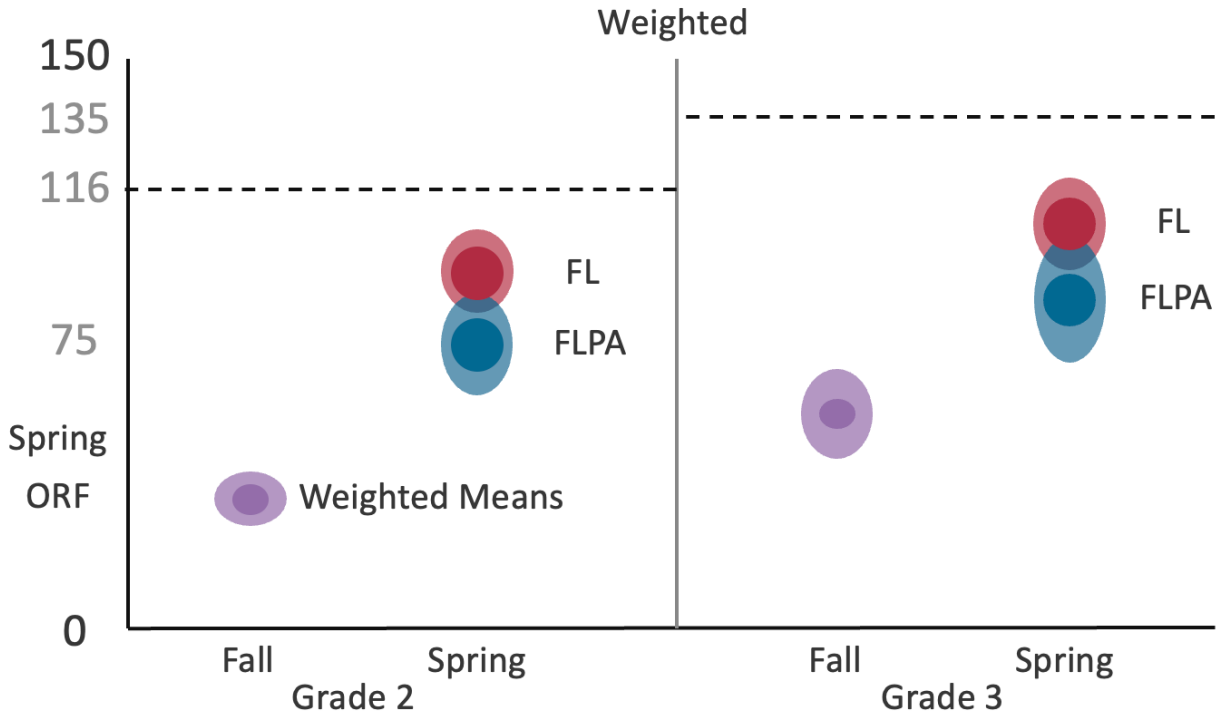


Figure 7. Weighted mean ORF scores by condition in Fall and Spring. Note: Point estimates are represented by the opaque circles and are surrounded standard deviations represented by the translucent ovals.

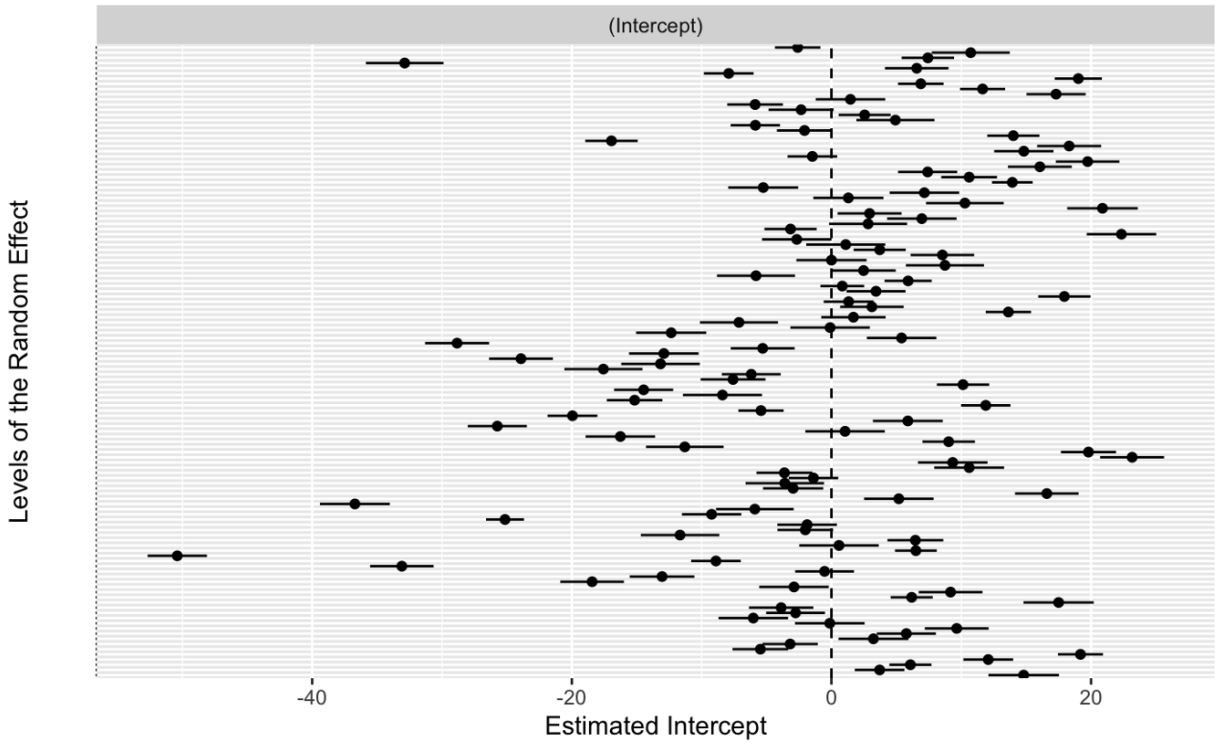


Figure 8. Caterpillar plot depicting estimated random intercepts (relative to overall grand intercept at 0) on the X axis by school site on the Y axis.

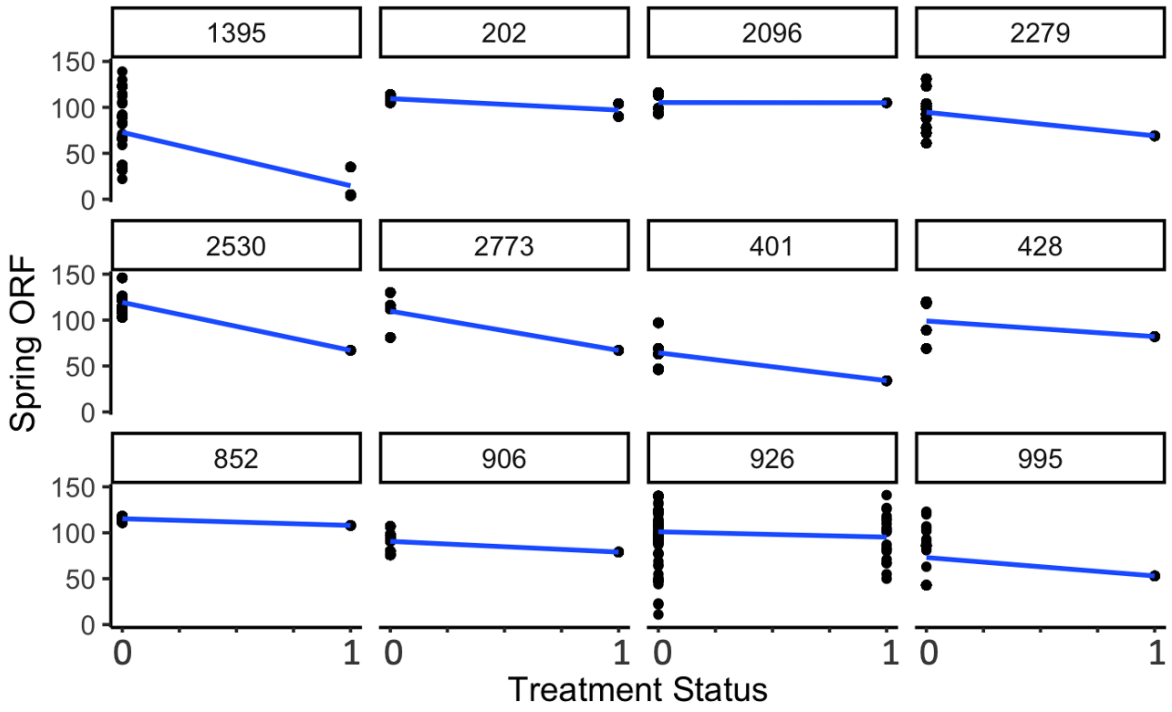


Figure 9. Example site-specific naïve treatment effects for a randomly selected group of 15 schools. Note: Although treatment means and slopes were allowed to vary in this example graph, this is for illustrative purposes only. Only random intercepts were used in the evaluative model, though random slopes were used in the sensitivity analyses.

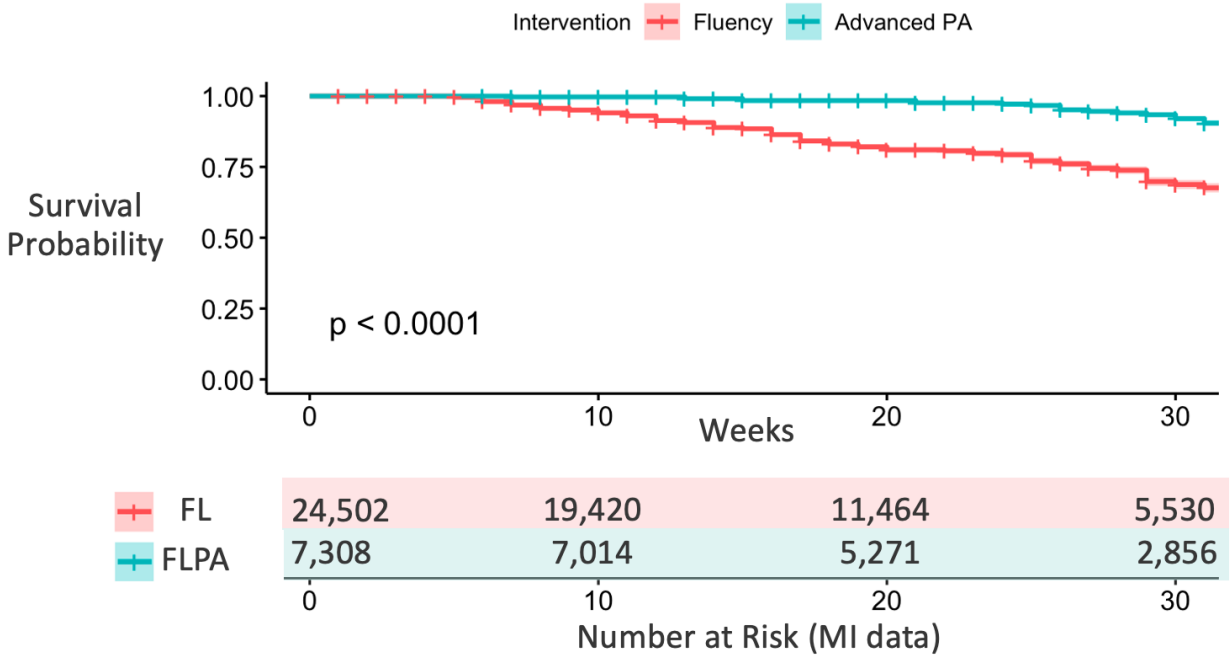


Figure 10a. Survival curves and risk tables for second grade multiply imputed datasets.

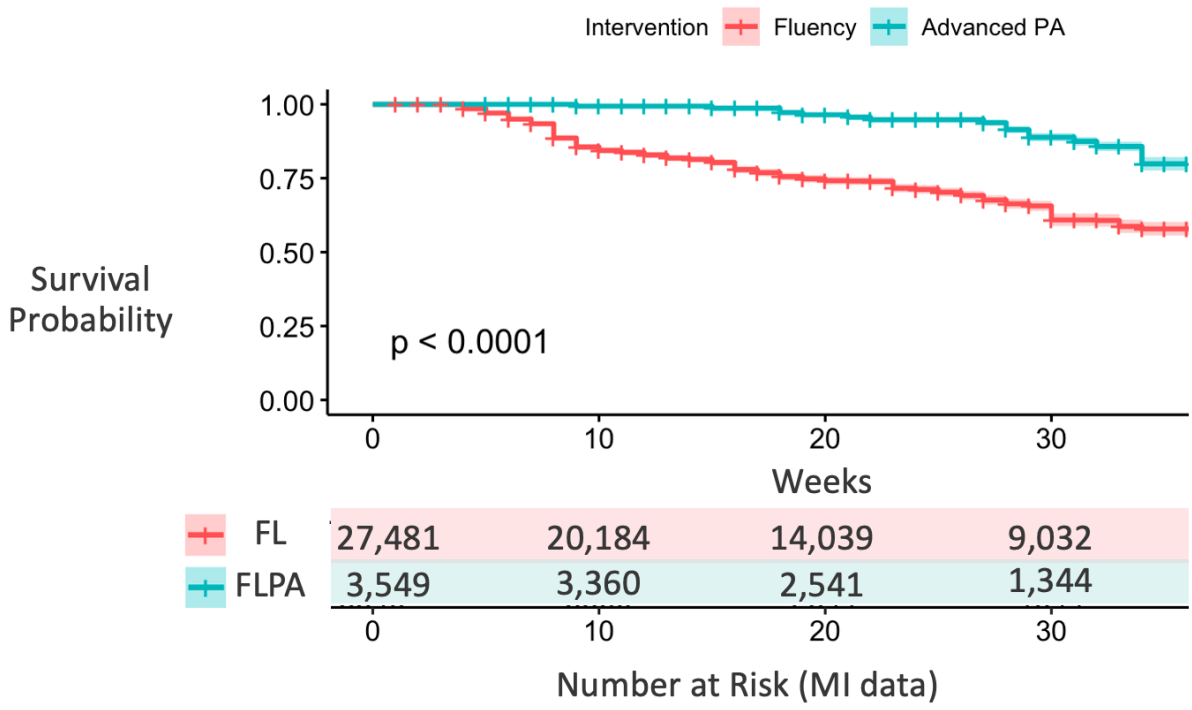


Figure 10b. Survival curves and risk tables for third grade multiply imputed datasets.

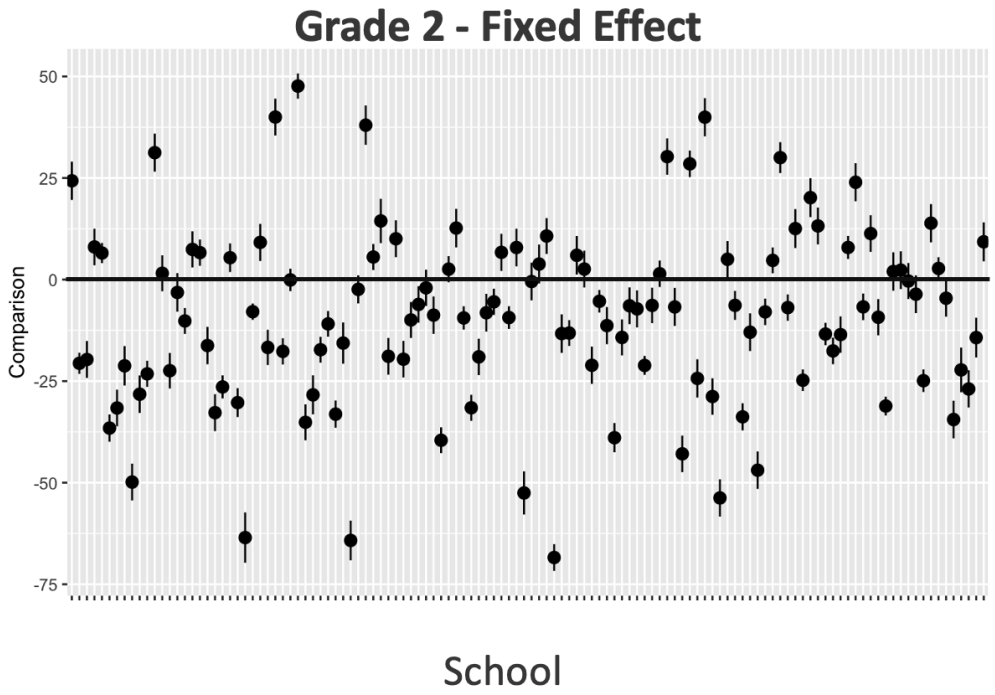


Figure 11a. Site specific marginal effects for Grade 2. Note: School site is coded on the X-axis with estimated treatment effect on the Y axis. More negative values favor FL over FLPA.

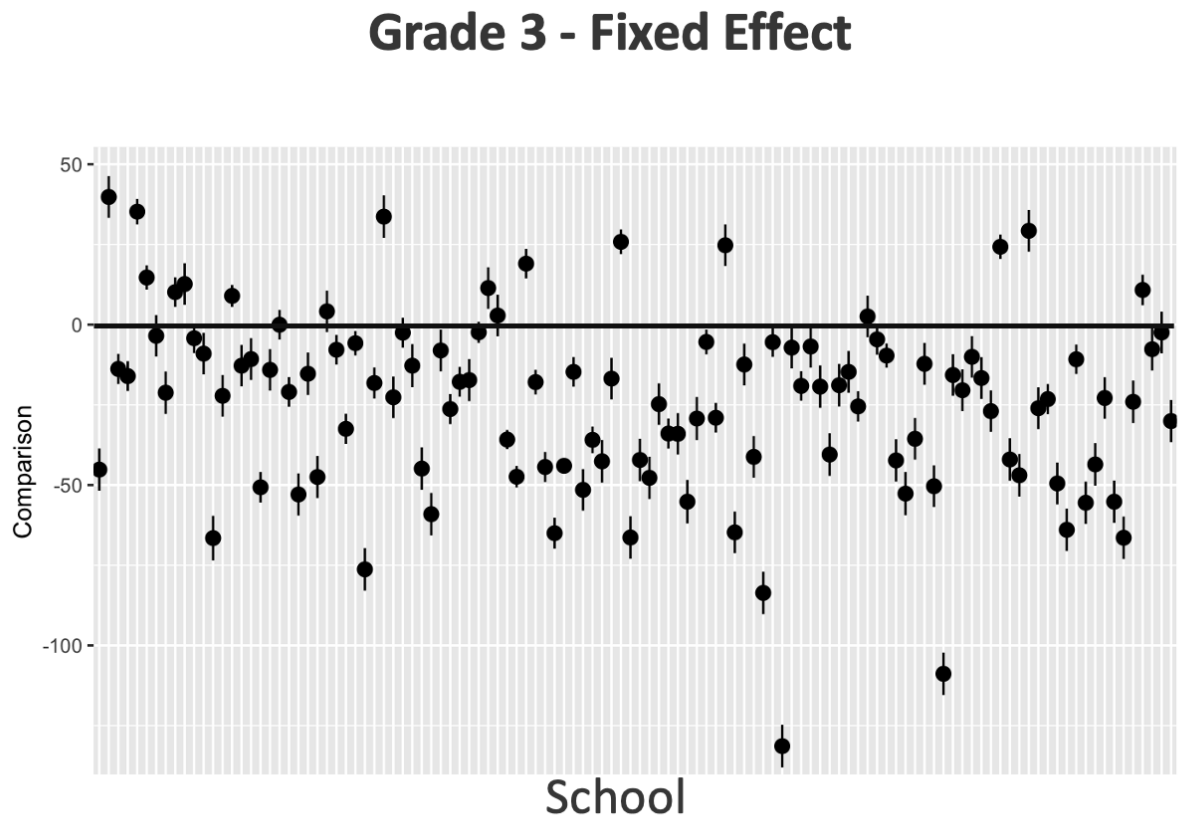


Figure 11b. Site specific marginal effects for Grade 3. Note: School site is coded on the X-axis with estimated treatment effect on the Y axis. More negative values favor FL over FLPA.

Table 1. Participant characteristics at baseline for full sample. Note: Parentheses represent standard deviations of the mean for WCPM, EPM, and Accuracy, but proportions for the remaining variables.

	Treatment Status	
Variable	**Control (Fluency)**	**Treatment (Fluency + Advanced PA)**
n	10,820	2,081
Fall Words Read Correct Per Minute	64.02 (22.60)	34.45 (20.79)
Fall Errors Per Minute	4.34 (4.52)	7.05 (4.63)
Fall Accuracy	0.92 (0.09)	0.78 (0.18)
Gender - Female (%)	5,083 (47.0%)	973 (46.8%)
Gender - Male (%)	5,068 (47.0%)	955 (45.9%)
English Language Learner (%)	12%	13%
Special Education (%)	3%	4%
Race - AI/AN (%)	157 (1.5%)	20 (1.0%)
Race - Asian (%)	286 (2.6%)	26 (1.2%)
Race - Black or African American (%)	2,346 (21.7%)	886 (42.6%)
Race - Hispanic/Latin(x) (%)	1,091 (10.1%)	272 (13.1%)
Race - Multiracial (%)	266 (2.5%)	66 (3.2%)
Race - NH/PI (%)	59 (0.5%)	2 (0.1%)
Race - White (%)	5,821 (52.2%)	643 (30.9%)
Grade 2 (%)	4,890 (76.6%)	1,491 (23.4%)
Grade 3 (%)	5,930 (85.7%)	590 (14.3%)

Table 2a. Unbalanced standardized mean differences (SMD) between treatment groups.

Unbalanced

Variable	Variable Type	SMD - Grade 2	Descriptor - Grade 2	SMD - Grade 3	Descriptor - Grade 3
Fall Words Read Correct Per Minute	Continuous	-0.7476	Not Balanced	-1.1993	Not Balanced
Missingness - Fall WCPM	Binary	0.0735	Not Balanced	0.0744	Not Balanced
Fall Accuracy	Continuous	-0.5454	Not Balanced	-0.7435	Not Balanced
Missingness - Fall Accuracy	Binary	0.0728	Not Balanced	0.0807	Not Balanced
Fall Errors per Minute	Continuous	0.3814	Not Balanced	0.6105	Not Balanced
Race	Binary	0.0223	Balanced	0.004	Balanced
Race - AI/AN (%)	Binary	-0.001	Balanced	-0.0095	Balanced
Race - Asian (%)	Binary	-0.0125	Balanced	-0.56	Balanced
Race - Black or African American (%)	Binary	0.1187	Not Balanced	0.1659	Not Balanced
Race - Hispanic/Latin(x) (%)	Binary	0.0293	Balanced	0.0456	Balanced
Race - Multiracial (%)	Binary	0.0191	Balanced	0.0207	Balanced
Race - NH/PI (%)	Binary	-0.00048	Balanced	-0.0049	Balanced
Race - White (%)	Binary	-0.1711	Not Balanced	-0.2152	Not Balanced
Gender	Binary	0.0262	Balanced	0.0076	Balanced
Gender - Female	Binary	0.0156	Balanced	0.034	Balanced
Gender - Male	Binary	-0.0418	Balanced	-0.0416	Balanced
English Language Learner	Binary	0.0052	Balanced	0.0076	Balanced
Missingness - English Language Learner	Binary	0.0223	Balanced	0.026	Balanced
Special Education	Binary	0.0064	Balanced	0.0361	Balanced
Missingness - Special Education	Binary	-0.0587	Not Balanced	-0.0235	Balanced
Home Language (English vs. Not)	Binary	-0.0112	Balanced	-0.0083	Balanced

Table 2b. GLM balanced standardized mean differences (SMD) between treatment groups. Logit function using Maximum Likelihood Estimation.

GLM

Variable	Variable Type	SMD - Grade 2	Descriptor - Grade 2	SMD - Grade 3	Descriptor - Grade 3
Fall Words Read Correct Per Minute	Continuous	-0.0562	Not Balanced	-0.0431	Balanced
Missingness - Fall WCPM	Binary	-0.0966	Not Balanced	-0.031	Balanced
Fall Accuracy	Continuous	-0.1096	Not Balanced	-0.0883	Not Balanced
Missingness - Fall Accuracy	Binary	-0.0966	Not Balanced	NA	Balanced
Fall Errors per Minute	Continuous	0.1224	Not Balanced	0.0350	Balanced
Race	Binary	0.0279	Balanced	-0.0172	Balanced
Race - AI/AN (%)	Binary	-0.0011	Balanced	-0.0149	Balanced
Race - Asian (%)	Binary	-0.0006	Balanced	-0.0021	Balanced
Race - Black or African American (%)	Binary	0.0618	Not Balanced	0.0294	Balanced
Race - Hispanic/Latin(x) (%)	Binary	0.0391	Balanced	0.0413	Balanced
Race - Multiracial (%)	Binary	0.0025	Balanced	0.0093	Balanced
Race - NH/PI (%)	Binary	0.000	Balanced	-0.0458	Balanced
Race - White (%)	Binary	-0.1296	Not Balanced	-0.0458	Balanced
Gender	Binary	0.0273	Balanced	-0.0172	Balanced
Gender - Female	Binary	-0.020	Balanced	0.0229	Balanced
Gender - Male	Binary	-0.0073	Balanced	-0.0057	Balanced
English Language Learner	Binary	0.0392	Balanced	0.0143	Balanced
Missingness - English Language Learner	Binary	0.0335	Balanced	-0.120	Balanced
Special Education	Binary	0.0081	Balanced	0.000	Balanced
Missingness - Special Education	Binary	-0.0428	Balanced	-0.0111	Balanced
Home Language (English vs. Not)	Binary	-0.0484	Balanced	0.0057	Balanced

Table 2c. Entropy balanced standardized mean differences (SMD) between treatment groups. Logit function maximizing Shannon Entropy (minimizing Kullback divergence).

		Entropy balancing			
Variable	**Variable Type**	**SMD - Grade 2**	**Descriptor - Grade 2**	**SMD - Grade 3**	**Descriptor - Grade 3**
Fall Words Read Correct Per Minute	Continuous	0.000	Balanced	0.000	Balanced
Missingness - Fall WCPM	Binary	0.000	Balanced	0.000	Balanced
Fall Accuracy	Continuous	0.000	Balanced	0.000	Balanced
Missingness - Fall Accuracy	Binary	0.000	Balanced	0.000	Balanced
Fall Errors per Minute	Continuous	-0.0001	Balanced	0.000	Balanced
Race	Binary	0.000	Balanced	0.000	Balanced
Race - AI/AN (%)	Binary	0.000	Balanced	0.000	Balanced
Race - Asian (%)	Binary	0.0002	Balanced	0.000	Balanced
Race - Black or African American (%)	Binary	-0.0001	Balanced	0.000	Balanced
Race - Hispanic/Latin(x) (%)	Binary	-0.0001	Balanced	0.000	Balanced
Race - Multiracial (%)	Binary	0.0001	Balanced	0.0001	Balanced
Race - NH/PI (%)	Binary	0.000	Balanced	0.000	Balanced
Race - White (%)	Binary	0.000	Balanced	0.000	Balanced
Gender	Binary	0.000	Balanced	0.000	Balanced
Gender - Female	Binary	0.000	Balanced	0.000	Balanced
Gender - Male	Binary	0.000	Balanced	0.000	Balanced
English Language Learner	Binary	0.000	Balanced	0.000	Balanced
Missingness - English Language Learner	Binary	-0.0001	Balanced	0.000	Balanced
Special Education	Binary	0.0006	Balanced	0.000	Balanced
Missingness - Special Education	Binary	0.000	Balanced	0.000	Balanced
Home Language (English vs. Not)	Binary	0.000	Balanced	0.000	Balanced

Table 3a. Grade 2 descriptive statistics for the final trimmed sample, clustered at the school level. Note: Descriptive statistics for participants' racial identity were not grade specific and collapsed across grades.

Variable	Mean	SD	IQR
Fall Words Read Correct Per Minute	43.40	10.26	37.59-51.00
Fall Accuracy	87.04%	8.50%	84.40 – 92.30%
Gender - Male (%)	50.40%	-	37.20 – 66.70%
English Learner (%)	10.08%	19.00%	0.00 – 13.39%
Special Education (%)	3.61%	11.63%	0.0 – 0.0%
Race - AI/AN (%)	1.41%	-	0.00 – 0.00
Race - Asian (%)	2.75%	-	0.00 – 0.00
Race - Black or African American (%)	27.28%	-	0.00 – 50.00%
Race - Hispanic/Latin(x) (%)	10.68%	-	0.00 – 14.29%
Race - Multiracial (%)	2.18%	-	0.00 – 0.00%
Race - NH/PI (%)	0.34%	-	0.00 – 0.00
Race - White (%)	50.21%	-	6.67 – 88.89%
Home Language English	86.36%	25.37%	83.00 – 100%

Table 3b. Grade 3 descriptive statistics for the final trimmed sample, clustered at the school level. Note: Descriptive statistics for participants' racial identity are not grade specific and collapsed across grades.

Variable	Mean	SD	IQR
Fall Words Read Correct Per Minute	72.37	13.77	63.70 – 83.47
Fall Accuracy	92.97%	4.71%	91.02 – 96.12
Gender - Male (%)	49.17%	-	40.0 – 57.1%
English Learner (%)	12.80%	19.79%	0.00 – 17.8%
Special Education (%)	2.82%	11.84%	0.00 – 0.00
Race - AI/AN (%)	1.41%	-	0.00 – 0.00
Race - Asian (%)	2.75%	-	0.00 – 0.00
Race - Black or African American (%)	27.28%	-	0.00 – 50.00%
Race - Hispanic/Latin(x) (%)	10.68%	-	0.00 – 14.29%
Race - Multiracial (%)	2.18%	-	0.00 – 0.00%
Race - NH/PI (%)	0.34%	-	0.00 – 0.00
Race - White (%)	50.21%	-	6.67 – 88.89%
Home Language English	81.75%	28.15%	71.43 – 100%

Table 4a. Grade 2 descriptive statistics pooled across 20 multiply imputed data sets. No missingness was observed in observations of students' racial identities.

	Treatment Status	
Variable	**Control (Fluency)**	**Treatment (Fluency + Advanced PA)**
Fall Words Read Correct Per Minute	45.21 (14.49)	36.69 (16.91)
Fall Errors Per Minute	5.15 (4.32)	6.32 (5.17)
Fall Accuracy	88.36% (10.91%)	81.0% (17.3%)
Gender - Male (%)	49.66%	51.93%
English Language Learner (%)	12.29%	12.47%
Special Education (%)	3.67%	6.36%
Home Language English	87.23%	87.7%

Table 4b. Grade 3 descriptive statistics pooled across 20 multiply imputed data sets. No missingness was observed in observations of students' racial identities.

	Treatment Status	
Variable	**Control (Fluency)**	**Treatment (Fluency + Advanced PA)**
Fall Words Read Correct Per Minute	74.38 (19.50)	61.01 (25.84)
Fall Errors Per Minute	4.23 (3.76)	6.05 (4.25)
Fall Accuracy	93.89% (6.06%)	87.30 (12.30%)
Gender - Male (%)	49.19%	50.43%
English Language Learner (%)	15.36%	18.50%
Special Education (%)	2.07%	1.11%
Home Language English	83.04%	79.01%

Table 5a. Grade 2 participant characteristics at baseline for the updated, trimmed sample. Note: Data in this table use listwise deletion, not MI.

Variable	Control (Fluency)	Treatment (Fluency + Advanced PA)
n	1,308	397
Fall Words Read Correct Per Minute	46.19 (13.92)	34.00 (16.64)
Fall Errors Per Minute	5.00 (4.18)	6.77 (5.47)
Fall Accuracy	89.0% (10.0%)	80.0% (17.0%)
Gender - Female (%)	638 (48.8%)	188 (47.4%)
English Language Learner (%)	*9%*	10%
Special Education (%)	*3%*	5%
Race - AI/AN (%)	17 (1.3%)	5 (1.3%)
Race - Asian (%)	38 (2.9%)	6 (1.5%)
Race - Black or African American (%)	322 (24.6%)	118 (29.7%)
Race - Hispanic/Latin(x) (%)	106 (8.1%)	39 (9.8%)
Race - Multiracial (%)	30 (2.3%)	7 (1.8%)
Race - NH/PI (%)	5 (0.4%)	0 (0.0%)
Race - White (%)	732 (56.0%)	209 (52.6%)

Table 5b. Grade 3 participant characteristics at baseline for the updated, trimmed sample. Note: Data in this table use listwise deletion, not MI.

Variable	Control (Fluency)	Treatment (Fluency + Advanced PA)
n	1,350	181
Fall Words Read Correct Per Minute	74.89 (19.05)	56.08 (26.31%)
Fall Errors Per Minute	4.16 (3.70)	6.72 (4.25)
Fall Accuracy	0.94 (0.06)	0.86 (0.13)
Gender - Female (%)	661 (49.0%)	86 (47.5%)
English Language Learner (%)	13%	17%
Special Education (%)	3%	2%
Race - AI/AN (%)	27 (2.0%)	5 (2.8%)
Race - Asian (%)	47 (3.5%)	8 (4.4%)
Race - Black or African American (%)	477 (35.3%)	73 (40.3%)
Race - Hispanic/Latin(x) (%)	157 (11.6%)	26 (14.4%)
Race - Multiracial (%)	33 (2.4%)	2 (1.1%)
Race - NH/PI (%)	8 (0.6%)	0 (0.0%)
Race - White (%)	536 (39.7%)	60 (33.1%)

Table 6a. Excluded participant characteristics for Grades 2 and 3 by condition.

Variable	Control (Fluency)	Treatment (Fluency + Advanced PA)
	Grade 2	
Fall Words Read Correct Per Minute	46.94 (14.76)	28.18 (16.30)
Fall Errors Per Minute	5.02 (4.91)	7.39 (4.36)
Fall Accuracy	89.00% (11.00%)	74.00% (19.00%)
	Grade 3	
Fall Words Read Correct Per Minute	77.74 (18.23)	41.97 (23.33)
Fall Errors Per Minute	3.73 (4.48)	6.55 (4.58)
Fall Accuracy	95.00% (6.00%)	82.00% (15.00%)

Table 6b. Grade 2 descriptive statistics for excluded participants, clustered at the excluded school level. Note: Descriptive statistics for participants' racial identity are not grade specific and collapsed across grades.

Variable	Mean	SD	IQR
Fall Words Read Correct Per Minute	42.77	13.11	35.06 – 51.75
Fall Accuracy	85.85%	10.47%	82.97% - 92.66%
Fall Errors Per Minute	5.48	2.57	3.67 – 6.99
Gender - Male (%)	50.48%	24.00%	33.33% – 66.67%
English Learner (%)	12.71%	24.13%	0.00% - 16.67%
Special Education (%)	3.66%	13.78%	0.00% - 0.00%
Race - AI/AN (%)	1.34%		
Race - Asian (%)	2.54%		
Race - Black or African American (%)	23.50%		
Race - Hispanic/Latin(x) (%)	10.28%		
Race - Multiracial (%)	2.86%		
Race - NH/PI (%)	0.43%		
Race - White (%)	49.69%		
Home Language English	81.50%	32.29%	80.00% - 100.00%

Table 6c. Grade 3 descriptive statistics for excluded participants, clustered at the excluded school level. Note: Descriptive statistics for participants' racial identity are not grade specific and collapsed across grades.

Variable	Mean	SD	IQR
Fall Words Read Correct Per Minute	75.43	15.21	68.00 – 85.64
Fall Accuracy	93.99%	5.30	93.02% - 96.92%
Fall Errors per Minute	3.96	2.43	2.60 – 4.93
Gender - Male (%)	48.32%	23.21%	33.33% – 62.50%
English Learner (%)	11.88%	23.72%	0.00% - 14.29%
Special Education (%)	3.41%	13.19%	0.00% - 0.00%
Race - AI/AN (%)	1.34%		
Race - Asian (%)	2.54%		
Race - Black or African American (%)	23.50%		
Race - Hispanic/Latin(x) (%)	10.28%		
Race - Multiracial (%)	2.86%		
Race - NH/PI (%)	0.43%		
Race - White (%)	49.69%		
Home Language English	82.47%	31.64%	80.00% – 100.00%

Table 7a. Weighted baseline and outcome descriptive statistics by treatment group for Grades 2 and 3.

Variable	Control (Fluency)	Treatment (Fluency + Advanced PA)
	Grade 2	
Fall Words Read Correct Per Minute	33.90 (13.92)	34.00 (16.64)
Fall Errors Per Minute	6.76 (4.18)	6.77 (5.47)
Fall Accuracy	80.22 (10.38%)	80.30 (16.79%)
Spring Words Read Correct per Minute	93.75 (21.83)	74.96 (27.38)
Unweighted Spring WCPM	100.13 (21.83)	74.97 (27.38)
	Grade 3	
Fall Words Read Correct Per Minute	56.07 (19.05)	56.08 (26.31)
Fall Errors Per Minute	6.72 (3.70)	6.72 (4.25)
Fall Accuracy	85.76 (5.90%)	85.77 (12.79%)
Spring Words Read Correct per Minute	106.54 (24.85)	86.28 (33.67)
Unweighted Spring WCPM	116.57 (24.82)	86.28 (33.67)

Table 7b. Centered, weighted baseline and outcome descriptive statistics within school for treatment and control groups.

Variable	***Grade 2***		***Grade 3***	
	Control	**Treatment**	**Control**	**Treatment**
Fall Words Read Correct Per Minute	10.99 (4.13)	0 (5.07)	13.18 (3.96)	0 (9.57)
Fall Errors Per Minute	-1.874 (1.74)	0 (1.83)	-1.67 (1.33)	0 (2.05)
Fall Accuracy	0.08 (0.03)	0 (0.04)	0.07 (0.02)	0 (0.05)

Table 8a. Marginal Effect Estimates for Fixed Effect Models. Note: SE represents cluster robust (CR2) standard errors clustered by school

Model	Estimate	SE	Z	P	CI	CI
Grade 2 - Listwise	-15.8	5.89	-2.69	.007	-27.4	-4.27
Grade 2 - MI	-11.8	2.09	-5.64	<.001	-15.9	-7.69
Grade 3 - Listwise	-8.52	1.79	-4.76	<.001	-12	-5.01
Grade 3 - MI	-13.8	3.72	-3.71	<.001	-21.1	-6.52

Table 8b. Fixed effect model coefficient output for Grade 2 MI data. Adjusted Model $R^2 = 0.8177$. Note: No single beta represents the effect of treatment status due to the treatment by covariate interactions. Marginal effects summarize effects of treatment status by g-computing over all calculated betas. Note: Betas for School ID and Treatment by School ID interactions could not be included in the table due to the number of terms (400+). School specific marginal effects (Treat x Site ID), however, are visualized in Figure 11.

Variable	Beta	SE	t	P
Intercept	86.55	1.31	66.03	<.001
Treatment Status (1 for FLPA, 0 for FL)	20.45	2.39	8.55	<.001
Baseline ORF	0.12	0.01	9.89	<.001
Baseline Accuracy	11.13	1.78	6.24	<.001
Baseline Errors	0.17	0.04	4.27	<.001
EL Status	-4.59	0.70	-6.52	<.001
Home Language English	-22.47	0.89	-25.19	<.001
Treat x Baseline ORF	0.04	0.03	1.489	.137
Treat x Baseline Acc.	114.70	5.01	22.90	<.001
Treat x Baseline Errors	2.21	0.10	21.07	<.001
Treat x EL Status	2.82	1.49	1.90	.057
Treat x Home Language	19.28	1.75	11.00	<.001
Site IDs				
*TX * Site ID's*				

Table 8c. Fixed Effect model coefficient output for Grade 3 MI Data. Adjusted Model R^2 = 0.8051. Note: No single beta represents the effect of treatment status due to the treatment by covariate interactions. Marginal effects summarize effects of treatment status by g-computing over all calculated betas. Note: Betas for School ID and Treatment by School ID interactions could not be included in the table due to the number of terms (400+).

Variable	Beta	SE	t	P
Intercept	125.33	1.28	97.93	<.001
Treatment Status (1 for FLPA, 0 for FL)	-42.33	3.34	-12.67	<.001
Baseline ORF	0.15	0.01	24.348	<.001
Baseline Accuracy	72.54	2.60	27.88	<.001
Baseline Errors	-0.24	0.03	-6.66	<.001
EL Status	3.33	0.51	6.52	<.001
Home Language English	0.59	0.58	1.03	.305
Treat x Baseline ORF	-0.03	0.02	-1.422	.155
Treat x Baseline Acc.	-49.92	4.53	-11.03	<.001
Treat x Baseline Errors	0.77	0.09	8.37	<.001
Treat x EL Status	-7.86	1.88	-4.17	<.001
Treat x Home Language	15.84	2.26	7.00	<.001
Site IDs				
*TX * Site ID's*				

Table 9a. Random Effects model coefficient outputs for Grade 2 MI data. The Level 2 variance component was estimated at 351.81 (SE = 17.76).

Variable	Estimate	SE	T value	P
Intercept	93.01	1.96	47.496	<.001
Treatment Status (1 for FLPA, 0 for FL)	-12.93	1.82	-7.10	<.001
Baseline ORF	0.30	0.06	5.17	<.001
Baseline Accuracy	31.34	15.67	2.00	.023
Baseline Errors	0.50	0.21	2.37	.009
EL Status	-5.28	2.85	-1.86	.031
Home Language English	-7.29	4.89	-1.49	.068
Treatment x Baseline ORF	-0.14	0.18	-0.79	.215
Treatment x Baseline Acc.	94.72	43.10	2.20	.014
Treatment x Baseline Err.	1.89	0.78	2.43	.008
Treatment x EL Status	3.52	5.13	0.69	.245
Treatment x Home Lang.	4.10	6.14	0.67	.251

Table 9b. Random effect model coefficient output for Grade 3 MI data. The Level 2 variance component was estimated at 724.23 (SE = 26.91).

Variable	Estimate	SE	T value	P
Intercept	103.56	3.52	29.43	<.001
Treatment Status (1 for FLPA, 0 for FL)	-8.81	2.60	-3.39	<.001
Baseline ORF	0.24	0.05	4.81	<.001
Baseline Accuracy	113.18	24.69	4.58	<.001
Baseline Errors	0.494	0.26	1.91	.028
EL Status	-0.99	4.42	-0.22	.413
Home Language English	-0.80	4.72	-0.17	.432
Treatment x Baseline ORF	0.63	0.34	1.83	.034
Treatment x Baseline Acc.	-91.89	125.96	-0.73	.233
Treatment x Baseline Err.	0.98	1.29	0.76	.224
Treatment x EL Status	31.24	12.13	2.57	.005
Treatment x Home Lang.	56.81	16.62	3.42	<.001

Table 10a. Cox Proportional Hazards model coefficient output for Grade 2 MI Data

Variable	Beta	Hazard Ratio	SE	P
Treatment Status (1 for FLPA, 0 for FL)	-1.516	0.22	0.05	<.001
Baseline ORF	0.03	1.04	0.01	<.001
Baseline Accuracy	0.05	1.05	0.01	<.001
Baseline Errors	0.01	1.01	0.01	.121
EL Status	-0.52	0.59	0.11	<.001
Home Language English	-0.04	0.96	0.09	.653

Table 10b. Cox Proportional Hazards model coefficient output for Grade 3 MI Data

Variable	Beta	Hazard Ratio	SE	P
Treatment Status (1 for FLPA, 0 for FL)	-1.65	0.19	0.06	<.001
Baseline ORF	0.04	1.04	0.01	<.001
Baseline Accuracy	0.04	1.04	0.01	<.001
Baseline Errors	0.03	1.03	0.01	.008
EL Status	-0.43	0.65	0.07	<.001
Home Language English	-0.10	0.90	0.05	.052

Table 11: Estimates of treatment effect under sensitivity analyses. Note: Fixed effect estimates represent marginal effects, while Random Effect estimates represent raw estimates.

Model	Estimate	SE	Z	P	CI	CI
Grade 2 – No T by X (Fixed Effect)	-12.80	1.42	-9.02	<.001	-15.60	-10.00
Grade 3 – No T by X (Fixed Effect)	-8.89	2.12	-4.20	<.001	-13.00	-4.74
Grade 2 – No T by X (Random Effect)	-12.82	1.80	-7.13	<.001	-16.35	-9.29
Grade 3 – No T by X (Random Effect)	-8.90	2.52	-3.53	<.001	-13.84	-3.96
Grade 2 – Quadratic (Fixed Effect)	-11.00	2.58	-4.26	<.001	-16.00	-5.93
Grade 3 – Quadratic (Fixed Effect)	-3.49	8.46	-0.413	.680	-20.10	+13.1
Grade 2 – Quadratic (Random Effect)	-12.68	1.81	-7.01	<.001	-16.23	-9.13
Grade 3 – Quadratic (Random Effect)	-16.73	2.85	-5.87	<.001	-22.34	-11.12
Grade 2 – Random Slopes	-12.91	2.02	-6.403	<.001	-16.87	-8.95
Grade 3 – Random Slopes	-20.99	2.72	-7.72	<.001	-26.32	-15.66
Grade 2 – 20+ sessions (Fixed Effect)	-11.40	3.92	-2.91	.002	-19.08	-3.72
Grade 3 – 20+ sessions (Fixed Effect)	-12.00	2.74	-4.38	<.001	-17.40	-6.63
Grade 2 – 20+ sessions (Random Effect)	-12.87	1.91	-6.75	<.001	-16.61	-9.13
Grade 3 – 20+ sessions (Random Effect)	-7.68	2.50	-3.07	.002	-12.58	-2.78

Grade 2 – 40+ sessions (Fixed Effect)	-14.1	2.41	-5.86	<.001	-18.90	-9.41
Grade 3 – 40+ sessions (Fixed Effect)	-13.1	3.76	-3.48	<.001	-20.47	-5.73
Grade 2 – 40+ sessions (Random Effect)	-11.99	1.87	-6.40	<.001	-15.66	-8.32
Grade 3 – 40+ sessions (Random Effect)	-9.36	3.16	-2.96	.003	-15.55	-3.17
Mean Sensitivity Analysis Effect Estimate	-11.87					

Table 12. Centered, weighted baseline and outcome descriptive statistics within school for treatment and control groups.

Variable	***Grade 2***		***Grade 3***	
	Control	**Treatment**	**Control**	**Treatment**
Fall Words Read Correct Per Minute	10.99 (4.13)	0 (5.07)	13.18 (3.96)	0 (9.57)
Fall Errors Per Minute	-1.874 (1.74)	0 (1.83)	-1.67 (1.33)	0 (2.05)
Fall Accuracy	0.08 (0.03)	0 (0.04)	0.07 (0.02)	0 (0.05)