Predicting Poor Readers' Response to a Multi-Component Reading Comprehension Intervention

By

Emma Lu Hendricks

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Special Education

May, 10 2019

Nashville, Tennessee

Approved:

Douglas Fuchs, Ph.D.

Lynn Fuchs, Ph.D.

Jeanne Wanzek, Ph.D.

Amanda P. Goodwin, Ph.D.

To my husband Liam, always supportive and loving. In my darkest moments you encouraged me, made me laugh, and never let me give up.

Acknowledgements

This work would not have been possible without the financial support of the Vanderbilt University Peabody College Department of Special Education, the U.S. Department of Education Leadership Training Grant, the Graduate Honor Scholarship, the Peabody Dean's Fellowship, and the Institute of Educational Sciences grant (R324D130003). I would like to thank all of my Dissertation Committee members, who have provided invaluable feedback and encouragement to me during my graduate school career. I would like to extend special thanks my advisor, Dr. Douglas Fuchs, for providing me the opportunity to come to Vanderbilt and to work and learn under his tutelage. Thank you for always pushing me to achieve my potential and encouraging me to pursue my goals.

I am also thankful to my colleagues and peers at from the Special Education department whose support and camaraderie I will never forget. Finally, I would like to thank my family without whose support this achievement would not be possible. Thank you to my parents, Eric and Michelle, for pushing me to be my best, but doing so with love and understanding. Thank you to my sister, Kaitlin for providing perspective and being a wonderful role model. Thank you to my brother Gregory, for supporting me from afar and always grounding me. Most importantly, thank you to my amazing husband Liam, who was there for me through it all. You believed in me, supported me, and encouraged me and I'm so lucky to have you in my life.

Table of Contents

Page

List of Tables

Table Page

List of Figures

Introduction

Since the reauthorization of the Individuals with Disabilities Act (IDEA; U.S. Department of Education, 2004), many educators have used response-to-intervention (RTI) for two purposes: to identify students with learning disabilities (LD) and to provide tiers of increasingly intensive intervention to struggling students. The benefits and challenges of using RTI to make identification and eligibility decisions have been discussed in depth (Fuchs & Fuchs, 2006; Fuchs, Fuchs, & Stecker, 2010; Fuchs, Mock, Morgan, & Young, 2003; Gerber, 2005; Mastropieri & Scruggs, 2005; Vaughn & Fuchs, 2003). The current study focuses on RTI's second purpose; that is, the provision of appropriately intensive intervention to the children who require it. This aspect of RTI, referred to by Fuchs, Fuchs, and Compton (2012) as the "intervention-prevention" dimension, comes with its own unique set of implementation challenges and technical issues regarding how best to match students with interventions and to determine whether those students are sufficiently responsive to those interventions.

A well-functioning RTI framework allows school systems to utilize resources more efficiently through early identification and prevention (i.e. avoiding the wait-to-fail model) and ensuring that struggling students receive intervention support that is matched to their individual level of need. However, this system can only be effective when responsiveness – the mechanism by which students move through the tiers – is defined appropriately. When the criteria for response to intervention are too high, resources may be wasted providing intensive interventions to students who would flourish without them. If criteria are set too low, students may not receive instructional support that is truly needed. Because there are many ways of operationalizing RTI (e.g., Compton, 2006; Frijters, Lovett, Sevcik & Morris, 2013; L.S. Fuchs, 2003), any study of it, including any use of it by practitioners, must be interpreted in light of how adequate or inadequate response has been defined and measured and how this may have influenced findings of who responds to a particular intervention.

Operationalizing Response to Intervention

Methods

The operationalization of “response” in RTI frameworks has been the focus of ongoing discussion in the educational community (Frijters et al, 2013; Fuchs, Fuchs, & Compton, 2004; Schatschneider, Wagner & Crawford, 2008; Tolar, Barth, Fletcher, Francis & Vaughn, 2013). Response operationalization may be considered in two ways, the first of which is the method used. For example, should student response be determined by whether his or her scores rise to average-level performance (“final status,”) or should it be defined on the basis of improvement alone (“growth”). If growth is chosen over final status as the index of response, then how much growth represents meaningful change?

“Normalization” (Torgesen et al., 2001) is a widely used final-status method of indexing treatment response. Normalization defines adequate response to intervention as a post-treatment score at or above 90 on a standardized test. In contrast to normalization, there are alternative methods of operationalizing of response that use growth. These include within-individual gains replicated over tests (WIGROT; Scarborough et al., 2013); reliable change index (RCI) scores (Jaconbson & Truax, 1991; Frijters et al., 2013); growth curve estimates (Compton, 2000; Vadasy, Sanders & Abbot, 2008); and curriculum-based measurement (CBM) slope (Fuchs, Fuchs & Compton, 2004). Although growth curve estimates and CBM are used to assess change in research and practice, both methods require multiple data points beyond pre- and post-treatment assessment. As such, it was not possible to assess the utility of these two methods in the current study. Therefore, they will not be discussed in depth.

Normalization. Normalization is the desired result of many interventionists because affecting change such that a child with an initial "at-risk" label completes an intervention by achieving a score within the average range of a normative population is a reasonable signal of meaningful change and intervention success. However, many evaluations of reading comprehension interventions for older students do not find significant effects on standardized measures (Edmonds et al., 2009; Roberts, Torgesen, Boardman & Scammacca, 2008). Thus, the identification of adequate responders via the normalization approach tends to be conservative and is linked to initial levels of reading problems (i.e. students with higher incoming standardized scores are more likely to be identified as responsive). A commonly used criterion to identify at-risk readers is a reading score 1 standard deviation (SD) below the mean, or a standard score of 85. This means that students may be considered responsive if they improved performance by 5 standard score points. In some cases, this change may be less than the standard error of the reading measure (Frijters et al., 2013).

Growth. The within-individual-gains-replicated-over-tests (WIGROT) approach for identifying adequate response applies a growth criterion to multiple outcome measures. To be identified as responsive, this method requires students to demonstrate positive change across multiple measures of reading comprehension. One limitation of the WIGROT method is that the criterion for the magnitude of positive change is arbitrary. In their study of responsiveness, Fritjers et al. (2013) set the criterion as positive change from pre- to post-treatment that exceeded the standard error of measurement.

Another measure of growth is the reliable change index (RCI) (Jaconbson & Truax, 1991; Frijters, et al., 2013). This growth criterion is more commonly used in psychology than in education. An RCI score is calculated by dividing the change in a student's pre-to-post-treatment score by the standard error of the difference score for that measure. With regard to the RCI method, Frijters et al. (2013) state, “significant change is the degree of gain necessary to exceed the unreliability of the outcome measure” (p. 542).

Measures

Studies of response to intervention must also be considered through the lens of measures, especially in the area of reading comprehension. Various nationally normed, standardized measures of reading comprehension have been shown to tap different component abilities and to identify different groups of students as good or poor readers (Cutting & Scarborough, 2006). While not always feasible, multiple measures should be used to best capture the complex and multi-dimensional nature of the reading comprehension construct (Fletcher, 2006; Scarborough, 2001).

An additional consideration is that of near- versus far-transfer measures. By definition, measures of near transfer (NT) are closely aligned to the reading intervention, whereas far transfer (FT) measures are less aligned with the intervention. Studies of reading comprehension interventions, especially involving older children, often find effects on NT measures but not effects of a similar magnitude on FT measures (Edmonds et al., 2009, Fuchs et al., 2018). As Fuchs et al. (2018) argued, NT and FT measures of reading comprehension should be viewed as complementary methods of assessing change attributable to intervention. NT measures may be more sensitive to change than FT measures and could prove to be useful indicators of change, similar to CBM or criterion-referenced measures. In summary, findings from studies of individual differences in response to intervention must be carefully considered. The operationalization of response (i.e. choices about methods and measures) as well as demographic features of the sample (e.g., age and severity of initial reading deficit) can affect the interpretation of results.

Review of the Literature

I conducted a literature search to identify studies of predictors and moderators of upper elementary students' response to reading comprehension interventions. The purpose of the search was to gather information on the methods and measures used by researchers to define response and to identify patterns in findings about the predictors of response in this population. I searched for studies involving at risk students. From this initial pool, I eliminated studies with relatively younger (3rd grade and below) and older (6th grade and above) participants. Finally, I eliminated studies involving samples that were very dissimilar to the sample in the current study (i.e. very poor readers) and studies which presented only results of overall efficacy analyses.

There were two exceptions to the just-mentioned exclusion considerations. I did *not* eliminate one study with adolescent participants. In this study, the authors' analyses were similar to my own (Frijters et al., 2013). I did *not* eliminate a second study in which the authors conducted only an overall efficacy analyses (Vaughn, Solis, Miciak, Taylor & Fletcher, 2016) because the data from this study were shared by two additional studies I reviewed. Therefore, I reviewed a total of six recent studies on the differential effectiveness of multicomponent reading interventions for struggling upper elementary and middle-school students. Findings will be highlighted regarding for whom these interventions were effective and how response was operationalized.

Three research teams reported studies on the effectiveness of multicomponent reading interventions for struggling 4th grade students (Ritchey, Silverman, Montanaro, Speece & Schatschneider, 2012; Wanzek et al., 2016; Vaughn, Solis, Miciak, Taylor & Fletcher, 2016). A fourth study was conducted by Frijters, Lovett, Sevcik and Morris (2013) with a sample of 6th, 7th, and 8th grade students. While students in this study were older, Frijters et al.'s (2013) comparative investigation of methods of identifying change is noteworthy. Cho et al. (2015) and Miciak, Cirino, Ahmed, Reid, and Vaughn (2018) performed additional analyses focused on individual differences using the data collected by Vaughn et al. (2016).

Participants

In three of the four intervention studies, participants were 4th grade students who were identified as at-risk based on their performance on the Gates-MacGinitie Reading Comprehension subtest (GMRT; MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2006). Wanzek et al. (2016) selected students scoring at or below the 30th percentile; Vaughn et al. (2016) chose students at or below the 16th percentile; and the average performance of the children in Ritchey et al.'s (2012) study was the 18th percentile. However, Ritchey et al. (2012) selected students based on risk level, which was determined by entering their raw scores from the GMRT, the Test of Silent Word Reading Fluency (Mather, Hammill, Allen & Roberts, 2004), and teacher ratings of reading problems into a previously developed logistic regression equation (a full description of the selection procedures can be found in Speece et al., 2010). Frijters et al. (2013) selected 6th-8th grade students who scored at or below the 16th percentile on the Broad or Basic Reading Cluster Score of the Woodcock Johnson - III Tests of Achievement (WJ-III; Woodcock, McGrew & Mather, 2001).

Intervention

Each of the intervention programs combined multiple components to meet the complex needs of struggling intermediate-grade or middle-school readers, and each program was delivered in groups of 2 to 8 students. Vaughn et al. (2016) and Wanzek et al. (2016) used narrative and information passages, whereas the reading program evaluated by Ritchey et al. (2012) used only science texts. All four intervention programs involved activities to build word reading and decoding skills as well as oral reading fluency. The *PHAST* reading program tested by Frijters et al. (2013) emphasized decoding strategies. The four programs taught vocabulary words as well as variety of comprehension skills and strategies, including previewing, activating prior knowledge, comprehension monitoring, question asking and answering, and summarizing. For a more detailed comparison of treatment duration and components see Figure 1.

Analyses

Ritchey et al. (2012), Wanzek et al. (2016), and Vaughn et al. (2016) conducted analyses of the overall effectiveness of their programs (e.g. ANCOVA or multiple regression). Ritchey et al. (2012) and Wanzek et al. (2016) conducted additional analyses (e.g. moderation and quantile regression) to explore the individual differences via differential effectiveness of their programs for subgroups of students. Authors of the remaining three studies focused on individual differences in responsiveness by using analyses which classify students into groups of adequate or inadequate responders (e.g. logistic regression or discriminant function analysis) (Cho et al., 2015; Miciak et al., 2018; Frijters et al., 2013). The results of the three categories of analyses (overall efficacy, moderation, and classification) will be discussed separately below.

Overall efficacy. Ritchey et al.'s (2012) treatment group significantly outperformed controls on two post-treatment-only NT measures ($g = 0.65, 0.56$), but not on other outcome measures, including the GMRT. One of the NT measures assessed content acquisition, the other measure addressed strategy use and reading comprehension. In contrast, Wanzek et al. (2016) found small to medium effects on two FT, norm-referenced measures of reading comprehension, the WJ-III Passage Comprehension subtest and the GMRT Reading

Comprehension subtest (g = 0.14, 0.28), but both effects were non-significant. Vaughn et al. (2016) conducted ANCOVA analyses for each of their outcome measures, which included WJ-III Passage Comprehension and GMRT reading comprehension subtests. However, all effects were non-significant. The authors noted that both the experimental and control groups made strong normative gains over the course of the study.

Moderation. Ritchey et al. (2012) found a significant moderation effect on the NT reading comprehension measure, such that students who received more services (e.g. school provided interventions, special education, Title 1, or speech and language services) and the intervention outperformed comparable controls (g = 1.01). Ritchey et al. (2012) tested whether tutor ratings of attention predicted responsiveness for the outcome variables. They found a marginally significant result for the GMRT and CBM maze, suggesting that students with stronger attention may have benefitted more.

Wanzek et al. (2016) tested for moderation effects on the reading comprehension outcomes as well. Results indicated treatment effects on WJ-III Passage Comprehension were moderated by students' pre-treatment scores, such that students with scores at or above the 60th percentile on the WJ-III Passage Comprehension subtest significantly outperformed comparable controls. The authors also conducted an exploratory quantile regression analysis on the post-treatment GMRT scores, which indicated the intervention program was most effective for students with post-treatment GMRT scores between the .40 and .70 quantiles. In contrast to Ritchey et al.'s (2012) findings, Wanzek et al. (2016) found that their intervention was *least effective* for the students with the lowest levels of comprehension ability and most effective for students with higher scores on the WJ-III Passage Comprehension subtest at pre-treatment.

Classification. Both Cho et al. (2015) and Miciak et al. (2018) conducted individual difference analyses which relied on the classification of students into groups based on a binary outcome variable (i.e. adequate versus inadequate response). First the authors identified groups of adequate and inadequate responders. Adequate response was defined by using the normalization method with two FT, norm-referenced measures, the WJ-III and GMRT Passage Comprehension subtests. Cho et al. (2015) conducted profile analyses to determine whether adequate and inadequate responders differed on cognitive attributes (e.g. verbal knowledge, working memory, and listening comprehension) and on teacher ratings of attention and self-efficacy. Cho et al. found that inadequate responders scored significantly lower on verbal knowledge and listening comprehension. Discriminant function analysis indicated that verbal knowledge best discriminated between the two groups. Miciak et al. (2018) found that, while an EF factor score best discriminated between the two groups, the effect was not statistically significant.

Frijters et al. (2013) compared four methods (normalization, RCI, growth curves, and WIGROT) of identifying adequate and inadequate responders. The authors used binary logistic regression to investigate predictors of adequate response (as defined by each method) for each outcome measure, including word reading, fluency and reading comprehension tests. The following were the results for the reading comprehension outcome: The five predictor variables were phonological blending, phonological loop working memory, rapid letter naming, verbal IQ, and nonverbal IQ. For the normalization method, gender, verbal IQ, CTOPP phoneme reversal, and rapid letter naming were all significant predictors of response, with odds ratios above 1, suggesting that female students and students with higher scores on the predictors were more likely identified as responsive.

For the RCI method, only nonverbal IQ was a significant predictor. For the growth curve method, only rapid letter naming was a significant predictor. For WIGROTS, which accounted for performance on all outcome measures, only age was a significant predictor. Interestingly rapid letter naming speed was inconsistent predictor – for normalization, the odds ratio was above 1, indicating that students who were faster tended to be identified as responsive. For the growth curve method, the odds ratio for rapid letter naming was less than 1, indicating that students who were slower tended to be identified as responsive.

Frijters et al. (2013) noted that the agreement between normalization and the growth curve methods was at chance (k = .04, ns), indicating the two methods identified almost completely different subsamples of responders. Across outcome measures, the growth curve approach tended to identify responders with lower cognitive and language skills while the normalization method tended to identify responders with higher pretreatment skills.

Summary. Taken together, these studies fail to offer a clear and consistent description of which students are likely to benefit from multicomponent reading interventions. The moderation analyses conducted by Ritchey et al. (2012) and Wanzek et al. (2016) suggested different conclusions about which children are likely to respond. That is, higher-risk students appeared to benefit most from Ritchey et al.’s (2012) intervention while lower-risk students benefited most from Wanzek et al.’s (2016) intervention. A few critical differences in the intervention implementation could shed light on these disparate findings. For example, Ritchey et al.'s (2012) intervention was delivered in smaller groups (2-4 versus 4-6) and had fewer components than Wanzek et al.’s (2016) program. The results of classification analyses conducted by Cho et al. (2012), and Miciak et al. (2018) provided important information about the differences between responders and non-responders to a multi-component reading comprehension intervention.

However, the authors used only one method of operationalizing response: normalization with standardized and norm-referenced reading comprehension assessments. Results showed that responders and non-responders could be distinguished by listening comprehension and verbal knowledge measures, but not by measures of executive function. Interpretation of these results are complicated by findings that various methods of operationalizing response identify very different groups of students as adequate responders (Frijters et al., 2013).

Present Study

Fuchs, Fuchs and Compton (2012) argued that, for older students, there is less need for universal screening, since academic deficits are more obvious. Fuchs et al. (2012) suggested more resources should be allocated to determine which students are unlikely to respond to a tier 2 supplementary intervention and these children should be fast tracked into a more intensive intervention. To achieve this goal, more research is necessary on individual differences and predictors of response to intervention for upper elementary students.

Specifically, the use of classification analyses such as cognitive profile analysis, discriminant function analysis, and logistic regression may provide deeper and more accurate understanding of individual differences in response to intervention. Reading comprehension is a multidimensional construct and older struggling readers typically have multiple areas of weakness (Cirino et al., 2013). More research on the predictors of response in this population could lead to more accurate and efficient screening procedures to determine which students are likely to respond to different interventions. Additionally, the contrast of NT and FT tests of reading comprehension may provide important information about how researchers and practitioners may use growth in addition to final status as indicators of response.

I conducted logistic regression analyses using data from two consecutive years of intervention research (Years 4 and 5 of the Accelerating Academic Achievement [A3] of Children with Severe and Persistent Learning Disabilities research program). My purpose was to investigate the influence of individual differences in students' response to a multi-component reading comprehension program addressing informational text. I also investigated how the use of various combinations of reading measures and methods affected (a) the variables that predicted response and (b) which students were identified as responsive or unresponsive. The study questions follow:

1. Which child-level variables (i.e., grade and pre-treatment word reading, expressive vocabulary, non-verbal IQ, and working memory scores; as well as pre-treatment score on the reading comprehension measure and teacher ratings of student attention) best predict response to a multicomponent reading comprehension intervention?
2. Do the predictors change as a function of reading comprehension measures or method used to define responsiveness?
3. What proportion of tutored students are identified as responsive by each combination of measure and method?
4. To what extent do various operationalizations of indexing response agree (indexed by Cohen's kappa) regarding who is identified as responsive?

Method

Participants

Student selection and eligibility. As indicated, the student data came from two consecutive years of research on the efficacy of a multi-component reading comprehension program for fourth and fifth graders with poor comprehension. Selection procedures and criteria were similar in both years. However, there were differences. In Year 4 of the project, we conducted whole-class screening with a standardized reading comprehension assessment. We administered additional measures to students who met the screening criteria. This process was resource intensive. So, in Year 5, we discontinued whole-class screening and depended instead on teachers to nominate their low-performing students whom we then tested on various reading measures. In both years, we excluded students if they were frequently absent, were disruptive in class, or were not proficient in English (as measured by the English Language Development Assessment used by the school district). See Table 1 for eligibility criteria across the 2 years.

Student demographics. A total of 249 students completed the intervention program, on whom we had pre- and post-treatment data. Only the tutored students (not control students) were participants in the current study. Table 2 shows student demographic information. In Year 5, a slightly larger proportion of the sample was Hispanic and a smaller proportion was African American. There were fewer students in Year 5 who received free/reduced lunch and who had an Individualized Education Plan (IEP).

The purpose of the tutoring program was to teach reading comprehension strategies in expository text to students with adequate decoding but poor reading comprehension. At pre-treatment, the mean standard score on the Test of Word Reading Efficiency Sight Word Efficiency subtest (TOWRE SWE; Torgesen, Wagner & Rashote, 2012) was 95.04 (SD=7.47), indicating that students had adequate word reading skills. In contrast, the mean normal curve equivalent on the GMRT at pre-test was 36.82 (SD=10.39), which is equivalent to the 25th percentile or a standard score of 90. Student performance on the pre-treatment measures is shown in Table 3.

Staff. Research assistants (hereafter, RAs) were masters and doctoral students attending Peabody College of Vanderbilt University. In both years, 22 RAs were hired as tutors and testers. In both years, at least one special education doctoral student and two full time staff members assisted with tutoring and testing. RAs participated in extensive training before working with students. For more information see Procedures section.

Measures

We used many measures in both cohort years. The TOWRE Sight Word Efficiency subtest and Working Memory Test Battery Backward Digit Recall subtest (WMTB BDR; Pickering & Gathercole, 2001) were administered both years. The GMRT (MacGinitie, MacGinitie, Maria, Dreyer, & Hughes, 2006) and the Wechsler Individual Achievement Test Reading Comprehension subtests (WIAT; Wechsler, 2009) were also administered in both years. For technical information regarding the commercially-available, standardized, normed tests just mentioned, see Appendix A. In addition to these tests, the research team developed reading comprehension measures designed to assess a wider spectrum of learning transfer.

Near transfer. We designed a near-transfer (NT) test of reading comprehension. Students read four informational passages and answered 24 multiple choice questions about each. The questions assessed students' ability to identify paragraph and passage level main ideas and answer factual and inferential questions. The questions and passages were similar in presentation and topic to those used in tutoring. Whereas the passages had not been seen before by students, they were aligned thematically with topics from the tutoring program. Sample-based Cronbach's alpha for the NT test of reading comprehension at pre-and post-treatment was .70 and .72.

Mid transfer. A mid-transfer (MT) test of reading comprehension was also developed, which consisted of two informational passages on topics *not* related to those covered in tutoring. However, the presentation format (e.g. layout and design) and question types were similar to what the students experienced in tutoring. The questions required students to identify paragraph and passage level main ideas and answer factual and inferential questions. For purposes of the current analyses, both the NT and MT tests of reading comprehension will be considered criterion-referenced measures. Sample-based Cronbach's alpha for the MT test of reading comprehension at pre-and post-treatment was .64 and .66.

Tutoring

The multi-component tutoring program targeted students with adequate word reading but relatively poor reading comprehension. It was designed to teach strategies so students would read informational text with better understanding. Because there are no doubt many reasons why such students struggle to read with understanding (Cirino et al., 2013), the tutoring program included numerous evidence-based strategies for addressing a variety of weaknesses, including limited background knowledge and inadequate inference making, summarizing, and comprehension monitoring.

Tutoring occurred 3 times per week for 40-50 minutes per session for 14-15 weeks. Tutoring lessons were provided to pairs of students matched as closely as possible in terms of their reading skills. Lessons were scripted to support fidelity of treatment implementation and included standard correction procedures for incorrect responses. The tutoring program included a peer mediation component, where students worked together as Coach and Reader on various comprehension activities, as well as a motivational dimension by which students earned points for effort and accuracy.

In both years, two active treatments were compared to a control group. In Year 4, the two active treatments were Comprehension Only (COMP) and COMP plus working memory training ([WM]COMP). In Year 5, COMP was contrasted with COMP plus transfer training ([T]COMP). Few statistically significant differences were found between the two active treatments in either study year. Yet, in both years, children in the two treatments together outperformed controls on many reading measures. So, in this study, children in the two active treatments in Year 4 were combined and regarded as participating in a single "COMP" treatment and, similarly, children in the two active treatments in Year 5 were combined to form a COMP treatment.

While components of the base COMP treatment changed somewhat during development of the program, its primary strategies and structure have not changed. Each year, students learned to preview the passage, check their own background knowledge, make main ideas and answer factual and inferential questions (see Figure 2).

Base COMP program. The base COMP program was designed by combining strategies and activities that had been shown by previous research to be effective in promoting reading comprehension in older children reading informational text. In the active treatment groups, strategies and activities were presented to students as occurring either before-, during-, or after-reading.

Before reading strategies. In both years, students learned vocabulary words by reading the definitions in a glossary. For more abstract or difficult words, the tutor led a short discussion about the meaning of the word and provided examples. Students learned to identify text features, (e.g. titles, headings, maps, pictures and captions) as well as text structures (including descriptive, sequential, compare-contrast and problem-solution). Students checked their background knowledge about the day's topic and watched videos to build their background knowledge. In Year 5, students selected and watched videos every lesson, whereas in Year 4 students watched videos only occasionally and did not choose the video. Lastly, before students began reading, they made a prediction about the most important idea in the passage.

During reading strategies. In Year 4, the children were not taught during reading strategies. In Year 5, they were encouraged to "think while reading," and they learned to stop and clarify ideas that confused them. They could choose among five clarification methods, including re-reading, using background knowledge, and asking for help. They also learned to make connections between their own lives or previously read material and the ideas in the passage.

After reading strategies. Across both years, the after reading strategies remained the most stable. Students used a three-step strategy based on the paragraph shrinking strategy (Fuchs, Fuchs & Burish, 2000) to create the main idea for each paragraph. The same three step strategy was used to create the big idea, or the most important idea in the entire passage. At the end of each lesson, students used the five-step In or Out strategy to determine whether a question was factual (answer found in the passage) or inferential (answer required a connection to background knowledge) and answer the question appropriately

Procedures

Prior to administering test batteries at pre-treatment, RAs were trained to administer and score all assessments in a standardized way. RAs were trained approximately 11 hours during 5 weeks. After training, but before administering a particular measure, they were required to demonstrate at least 90% adherence to the standard administration and scoring rules during a fidelity check. If the RAs did not pass a fidelity check, they were required to retake the check for that measure until 90% fidelity was achieved. Fidelity checks were conducted by PCs and doctoral students, who used a checklist to determine the fidelity score and to provide feedback.

Similarly, before tutoring students the RAs were trained to administer lessons in standard fashion. RAs received 8 hours of tutoring training in two days, not including 2 hours of mandatory practice with other RAs. Each RA was also required to earn a score of 90% or higher on a tutoring fidelity check before tutoring began. During the time the tutoring program was conducted in the schools, two live fidelity checks were conducted on each RA. These checks occurred during a real tutoring session and were conducted by PCs or doctoral students. Before post-treatment testing, the RAs received an additional one to two hours of test training and were required to pass another round of fidelity checks. The 90% criteria remained the same and RAs were required to pass a fidelity check on each of the measures given at post-treatment.

Analytic Approach

I conducted binary logistic regression analyses, varying the method and measures used to classify students as responsive and non-responsive. The reading comprehension measures were (a) the WIAT-III Reading Comprehension subtest, (b) the GMRT Reading Comprehension subtest and (c) the NT and MT tests of reading comprehension. For each reading comprehension outcome measure, I used two methods to define student responsiveness; that is, *final status* and *growth*.

In each logistic regression, I tested the predictive value of seven student level variables, (grade and pretreatment word reading, expressive vocabulary, non-verbal IQ, working memory, pre-treatment score on the outcome measure, and teacher ratings of student attention). The researchers whose work was reviewed in the introduction all found statistically significant or marginally significant effects on variables representing these constructs, with the exception of word reading. However, the word reading eligibility criterion in Year 4 was much higher than in Year 5, and might have affected responsiveness.

The effect size produced by logistic regression is an odds ratio, which can be transformed into probabilities to simplify interpretation. I also plan to compute Cohen's kappa, which quantifies the chance-corrected agreement between methods in classifying students as responsive or not.

Final status method. For the standardized measures of reading comprehension, age-normed standard scores were calculated as described in the test manuals. A student with a post-treatment standard score of 100 or greater was classified as responsive to treatment. Such a score corresponds to the 50th percentile. Traditionally, as previously mentioned, the criterion for, or definition of, "normalization" is a standard score of 90 (25th percentile; cf. Torgesen et al., 2001). However, students in the Torgesen et al. studies often had greater reading deficits than students in the current sample. Table 3 shows that students' mean pre-treatment scores on the GMRT were the equivalent of a standard score of 90, and the average WIAT reading comprehension standard score was slightly higher. Using the conventional normalization criterion (25th percentile), approximately 60% of students in our sample could be classified as responsive based on their *pre-treatment* scores. Thus, in this study, it was more meaningful for me to set a higher normalization criterion. With the proposed 50th percentile normalization criterion, 24% of the student sample (59 of 249) could be classified as responsive based on their WIAT reading comprehension pre-treatment scores. In contrast, we used the GMRT to screen students scoring below the 50th percentile into our study. Therefore, none of our students could be classified as responsive based on their GMRT pre-treatment scores.

Moreover, the normalization method was modified in additional ways for use with the criterion-referenced NT and MT reading comprehension measures created by the A3 research team. Insufficient resources precluded our ability to administer these measures to a representative sample (only students already identified as poor in reading comprehension during sample selection were assessed). So, no normative distribution was available with which to compare students' post-treatment performance. The NT and MT reading comprehension measures are respectively more and less aligned with content and strategies taught in the intervention program. Thus, strong post-treatment performance on those measures indicated that students learned the strategies and applied them to passages and questions with content and format similar to what they experienced in tutoring. As with most criterion-referenced measures, a cut-off was required to determine whether students had performed adequately or not.

The final-status method I selected for the NT and MT reading comprehension measures were necessarily arbitrary, so I explored the utility of setting the criteria at various cut-off points, rather than at just one point. Table 4 shows the effect of different cut-off scores on the number (and percentage) of students at pre- and post-treatment who would be considered responsive on NT and MT reading comprehension measures. A cut-off of 50% items correct was not very informative: at both pre- and post-treatment, a majority of students would be identified as responsive. Similarly, setting the cut-off too high, at about 95% of items correct, identified very few students as responsive at pre-and post-treatment. Therefore, I used 75% and 87.5% cut-off points as criteria for normalized response for our researcher-developed criterion-referenced measures.

The 75% and 87.5% correct cut-off points, while arbitrary, are also meaningful because achievement at this level indicates that students were able to apply the strategies they learned to answer most of the questions on the measure correctly. Specifically, 87.5% items correct on the NT and MT reading comprehension measures corresponds to 21 out of 24 and 14 out of 16 items correct, respectively. In a classroom setting, a 75% grade, while not indicative of full mastery, typically represents a passing grade. Therefore, the selected cut-off points represent the level at which students needed to perform to demonstrate they learned the strategies and could apply them to novel passages and questions with familiar format and content.

Growth method. Responsiveness was also classified according to the level of growth demonstrated from pre- to post-treatment. Reliable change index (RCI) scores were calculated for each student on the commercial reading comprehension measures. To accomplish this, I used the Jacobson-Truax formula (Jacobson, Follette, & Revenstorf, 1984) with Maassen's (2004) modification[1]. Students were classified as responsive if the difference between their pre- and post-treatment scores on a reading comprehension measure were statistically greater than would be expected after accounting for the measure's reliability, unequal pre- and post-treatment variance, and practice effects (Maassen, 2004). The RCI criterion can theoretically be used with both commercially available, normed, and standardized reading comprehension measures as well as with researcher-created measures without normative data. However, the RCI criterion formula requires a "high-quality" (Maassen, 2004, p.889) estimate of the test-retest reliability of the measure, preferably derived from an independent normative sample. When possible for the commercial measures, I entered into the RCI formula the values found in the measures' technical manuals. The GMRT technical manual did not provide test-retest reliability data. The most

[1] $RCI = \frac{X_2 - X_1}{SEM_{diff}}$, where x_2 and x_1 are the student's post- and pre-treatment scores, respectively. The standard error of measurement of the difference score (SEM_{diff}) was calculated using the following formula from Maassen (2004): $\sqrt{(s_x^2 + s_y^2)(1 - r_{xy})}$ where s_x^2 and s_y^2 are the variances of pre-, and post-treatment scores, respectively, and r_{xy} is the test-retest reliability of the measure.

similar statistic provided in the manual was the correlation between the fall and spring administrations of the measure, so I used that in the RCI calculation. This estimate was more conservative and was likely lower than the true test-retest reliability for the GMRT.

For researcher developed measures, a sample specific reliability was calculated, but it was not as trustworthy as reliabilities derived from a large and representative normative sample. Additionally, due to logistic constraints, a true test-retest reliability estimate could not be obtained. Similar to the GMRT, the reliability of the criterion-referenced measures was estimated by calculating the Pearson's r correlation between control students' scores at pre- and post-treatment. Control students' scores, rather than the full samples' scores, were used to estimate the reliability of the measure as well as the pre- and post-treatment variances because the scores of that group were less likely to have changed due to treatment. The limitations of this approach are obvious, however, the NT and MT measures were designed to be more sensitive to intervention change than commercially available tests. Using the control students' scores, while also accounting for practice effects (by including in the RCI calculation the pre- and post-treatment variance), most closely replicates how RCI is calculated for the commercially available measures.

Ultimately, the use of the more conservative reliability statistic (combined with the less-than-ideal psychometric properties of the criterion referenced measures) resulted in few cases of positive reliable change on these measures (i.e. 4% and 12% of the sample for mid- and near-transfer, respectively). As such, logistic regressions using RCI for the criterion referenced measures as an outcome was not very informative. Instead, I used what L.S. Fuchs (2003) referred to as a limited norm criterion. The limited norm criterion is based only on a sample of tutored students, and compares each students' amount of growth to the other tutored students in the sample. I calculated the average score change on the NT and MT reading comprehension measures from pre- to post-treatment for all tutored students. Any student who met or exceeded the average score improvement was classified as responsive using this method.

Power analysis. Power analysis is typically performed prior to conducting an experiment to ensure that a large enough sample is recruited to detect a true significant effect of the expected magnitude. However, because I used extant data, sample size and measures have already been determined. Usually, estimates of the size of an effect size for power analysis are based on previous findings in literature. However, in logistic regression, this can be challenging because published studies rarely provide the information required for a power analysis (i.e. the probability of success at both the mean and one standard deviation above the mean on any predictor variable).

I ran power analyses on the extant data in Stata using the 'powerlog' command. For each criterion (final status or growth) and outcome measure (e.g. WIAT or NT reading comprehension), the sample size required to detect a significant effect on a particular predictor variable at the .80 level was calculated. The magnitude of the expected effect for each predictor variable was estimated using the following procedure. Two logistic regression models with only the focus predictor as an independent variable were run for each criterion-measure pairing. Stata's 'quietly' command was used to hide most of the results. In the first logistic regression model for each operationalization of response (e.g. WIAT growth method; Gates final status method) the predictor variable was held at the sample mean. In the second model, the predictor variable value was held at 1 standard deviation (SD) above the sample mean. The two resulting coefficients represented the probabilities of a student with a score at the mean and at 1 SD above the mean being classified as responsive according to that operationalization (i.e. criterion and measurement combination).

When the two probabilities (at the mean and at 1 SD above the mean of the focus predictor variable) were very similar, it meant the effect of that variable on the probability of a student being classified a responder was very small in the current sample. If the two probabilities were more divergent, it was more likely that a true effect existed. For the normalization outcomes, the sample size was large enough to detect a significant effect with .80 power in approximately 75% of the pairings (reading comprehension measure and predictor variable) where the two probabilities differed by at least 0.05.

For some operationalizations, especially those using RCI scores and the unstandardized measures, where few students were identified as "responsive," it was difficult to achieve adequate power. Logistic regression works best when the proportion of successes (responders) and failures (non-responders) are roughly equal in number. In particular, for the unstandardized reading comprehension measures, the low reliability resulted in relatively fewer students identified as responders. For the growth outcomes, only about 25% of the pairings (outcome measure and predictor variable) indicated the sample size was large enough to detect a significant effect of the expected magnitude with .80 power. Most of the other pairings where the two probabilities indicated a probable effect required sample sizes from 260-500 to achieve .80 power.

Results

Logistic regression analyses were conducted to investigate the predictors of responder status using 10 operationalizations of response (i.e. combinations of measure and method). In each analysis, the overall model was statistically significant and, in most analyses, the pseudo R-squared ranged from 0.25 to 0.15. However, GMRT-growth model had a poorer fit, which was indicated by a pseudo R-squared value of 0.09. This was likely due to the especially low number of students identified as responders using this method of operationalizing response. Of the 249 students in the sample, 17 were missing data from the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale (SWAN; Swanson et al., 2012), two were missing data from the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler & Hsiao-pin, 2011) Vocabulary subtest, and one was missing data on WMTB BDR. Cases with missing data were excluded list-wise from the analyses. Therefore, 229 cases with complete data were used in all analyses. To assist in interpretation of results, continuous predictors were converted to z-scores before analysis. Results in Table 5 can be interpreted as the increase (odds ratios greater than 1) or decrease (odds ratios less than 1) in the probability of being classified as a responder given a 1 SD increase on that variable relative to the sample mean.

Research Question 1: Which Child-Level Variables Best Predict Response to a Multicomponent Reading Comprehension Intervention?

The variables that best predicted response varied depending on the method and measure used to operationalize response (Table 5). However, some variables were more consistent predictors of response. Students' pre-treatment performance on the outcome measures were always significant predictors. For final status methods, the odds ratios for students' pre-treatment performance on reading comprehension measures were greater than one, indicating that students scoring higher on these measures at pre-treatment were more likely classified as responders at post-treatment. For example, a student who scored 1 SD above the sample mean on the WIAT at pre-treatment was 2.77 times more likely to be identified as a responder at post-treatment when response was defined as normalization (e.g. a standard score of 100 or more). Conversely, for each of the growth methods, the odds ratios for students' pre-treatment performance on the reading comprehension measures were less than one. This indicated that *lower* reading comprehension scores at pre-treatment were associated with a *higher* likelihood of being classified as a responder at post-treatment via the growth method, regardless of measure.

In summary, for final-status methods, students with higher pre-treatment scores on the reading comprehension measure were consistently more likely to be identified as responders. For growth methods, students with *lower* pre-treatment scores on the reading comprehension measure were consistently more likely to be identified as responders. Overall, students were more likely to be identified as responders across operationalizations of response (i.e. combinations of methods and measures) when their pre-treatment scores on variables such as expressive vocabulary, non-verbal IQ, and teacher ratings of attention were higher.

Research Question 2: Do the Predictors Change as a Function of Outcome Measures or Methods Used to Define Responsiveness?

Several variables performed well as predictors of responder status across operationalizations of response, including WASI Vocabulary and Matrix Reasoning subtests as well as the SWAN attention rating scale. In each case, the odds ratios for these predictor variables were greater than one, which indicated that students with higher pre-treatment scores on these variables were more likely to be identified as responders. Interestingly, WASI Vocabulary and the SWAN were significant predictors of response when the reading comprehension measures were NT reading comprehension and WIAT, but they were not significant predictors of performance on MT reading comprehension or GMRT. Student performance on WASI Matrix Reasoning was a significant predictor of response when the reading comprehension measure was the MT reading comprehension and the GMRT, but not NT reading comprehension or the WIAT.

WMTB BDR was identified as a significant predictor only on the NT reading comprehension measure and only when adequate response was defined as 75% of items correct. However, the effect approached significance when the cut-off was set as 87.5% correct. TOWRE SWE was only identified as a significant predictor when MT reading comprehension was the measure and adequate response was defined as (a) 87.5% correct or (b) limited norm. TOWRE SWE was the only variable (aside from students' pre-treatment scores on reading comprehension measures in growth models) where odds ratios indicated students with *lower* pre-treatment scores were more likely to be identified as responders.

Figure 3 provides a visual representation of these patterns. The figure contains five graphs depicting the predicted probability of being classified as a responder at various points in the distributions of the predictor variables. For example, Figure 3 (a) shows that the predicted probability of being classified as a responder via the MT reading comprehension measure final status method (87.5% cut off) and MT reading comprehension measure growth method both *increased* as the students' pre-treatment TOWRE SWE score *decreased*. In other words, students with lower scores on pre-treatment the TOWRE SWE subtest were more likely to be identified as responders via both final status and growth methods when the measure was the mid-transfer reading comprehension test. In contrast, the other graphs in Figure 3 illustrate the positive relationships between students' pre-treatment scores on the predictor variables and the probability of being a responder, such that students with higher pre-treatment scores on these variables are more likely to be classified as responders by the various operationalizations of adequate response.

Research Question 3: What Proportion of Tutored Students Are Identified as Responsive by Each Combination of Measure and Method?

Between 19% and 70% of the sample were identified as responders at post-treatment by the various operationalizations of response. The combination that identified the highest proportion of students (approximately 70% of the sample) was, as expected, NT reading comprehension measure final status method (cut-off of 75% of items correct). The operationalization that identified the lowest proportion of students as responders (approximately 19% of the sample) was the GMRT growth method (i.e. adequate response was defined using a RCI score). Across operationalizations, the average proportion of the sample identified as responders was 0.38. When adequate response is defined as meeting *either* the final status or the growth criterion for any given reading comprehension measure, a similar pattern appeared, albeit with higher proportions of "responsive" students. The GMRT measure still identified the smallest proportion of students as responders, about 30% of the sample, and the NT reading comprehension measure the largest, about 80% of the sample.

Research Question 4: To What Extent Do Various Operationalizations of Response Agree Regarding Who is Identified as Responsive?

Overall, the chance corrected agreement between operationalizations ranged from negative or chance agreement (-0.05, ns) to fair (< 0.40) (see Appendix B). Moderate agreement was only found between the GMRT final status and GMRT growth methods (k=0.47). The rates of agreement between various measures with the final status method ranged from poor to fair. The NT reading comprehension measure (87.5% correct cut off), MT reading comprehension measure (75% correct cut off) and the WIAT showed the highest rates of agreement with other final status methods. However, the magnitude of the kappa statistics indicated only fair agreement (Landis & Koch, 1977). Agreement between various reading comprehension measures with growth methods was extremely poor, each falling into the negative or at chance range. This finding indicates that each of the growth operationalizations identified almost a completely different group of students as responders.

Discussion

The primary objectives of this study were to determine which variables best predicted response to a multi-component reading comprehension intervention for at-risk 4th and 5th grade students and to explore the utility of various methods of determining response. The finding that lower risk students (i.e. those with higher pre-treatment scores on the reading comprehension measures) were more likely to be identified as responders across measures when response was determined using a final status method is in line with the findings of Frjiters et al. (2013) and of Wanzek et al. (2016), whose moderation analyses and classification analyses indicated that lower-risk students appeared to benefit more from the intervention.

In contrast, higher risk students (i.e. those with lower pre-treatment scores on the reading comprehension measures) were more likely to be identified as responders across measures when response was determined using a growth method. Ceiling effects may partially explain this pattern on the NT and MT reading comprehension measures. This is because students who scored higher on these measures at pre-treatment were likely unable to improve their score as much as students who scored lower at pre-treatment. The growth method I used to classify responders for these measures was the limited norm. This meant that students were classified as responders if the amount of growth in raw score points from their pre- to post-treatment performance was more than the average growth of all tutored students in the sample. The average amount of growth from pre-to post treatment for all tutored students in the sample was 3.5 raw score points. A student who scored 22 out of 24 on the NT reading comprehension measure at pre-treatment would be unable to be classified as a responder, even if they achieved a perfect score at post-treatment. However, the same pattern was also found on the standardized norm-referenced measures of reading comprehension, on which ceiling effects were less likely to occur. Additionally, the growth methods used to determine response were different for the criterion-referenced versus norm-referenced measures of reading comprehension, yet the pattern of findings was similar (i.e. higher risk students were more likely to be identified as responders when the method was growth). This suggests that this finding is not simply an artifact of the limitations of the criterion referenced measures and the limited norm method of assessing response.

The utility of using a growth method to determine responsiveness with a standardized, norm-referenced measure is complicated however by the findings presented in Table 6. These proportions indicate that for the WIAT and GMRT, most of the students who met the growth criterion also met the final status criterion. This finding calls into question the usefulness of a growth indicator of response in addition to final status, specifically for the commercial, standardized norm-referenced measures (Schatschneider, Wagner & Crawford, 2008). These findings suggest a distinction between students who are likely to meet a high criterion (using final status or growth method) on a standardized norm-referenced measure of reading comprehension and students whose response is subtler and should be viewed in terms of growth on more proximal measures.

One of the goals of this study was to provide more information to researchers and practitioners about how to efficiently allocate resources by matching students with interventions of appropriate intensity. How can students who are unlikely to respond strongly to an intervention be identified beforehand and fast-tracked to a more appropriate intervention? The findings of this study indicate that students with higher scores on pre-treatment variables such as expressive vocabulary, non-verbal IQ, teacher ratings of attention, and reading comprehension would be appropriately matched to this multi-component reading comprehension intervention. Students with this profile are more likely to demonstrate growth and meet a high final status or growth criterion across various measures after receiving this intervention.

In contrast, students who scored lower on the aforementioned variables and on reading comprehension measures may only show subtler signs of response (i.e. growth on proximal measures) after 13 weeks and approximately 33 hours of instruction with this intervention. In order to use resources most efficiently, practitioners should consider alternative options for students with a higher-risk profile. These higher-risk students may need to be fast tracked to a program of higher intensity (i.e. longer duration, more frequent sessions, or 1:1 teacher to student ratio). Or, maybe students with a higher-risk profile need an intervention program with a different approach to instruction. For example, the reading comprehension intervention in this study contained nine to ten components (see Figure 2). It is possible higher-risk students, who had weaker performance on a variety of cognitive and language variables, were overwhelmed by the number of strategies and components they were required to learn in this intervention program, such that few of the strategies were learned effectively. Perhaps for these students, a program focused on fewer components with more opportunities to practice each component would be more effective.

However, the same recommendation (i.e. higher-risk students should be fast tracked to a more intense intervention) was not supported by findings regarding students' word reading performance. The negative relationship between pre-treatment TOWRE SWE score and the likelihood of responder status on the MT reading comprehension measure was therefore an especially interesting finding. In the larger study, comparisons between treatment and control students on word reading at pre- and post-treatment showed no significant differences, indicating that the higher performance of the treatment groups on reading comprehension measures was not due to an improvement in word reading ability. The finding in the current study that students with lower incoming word reading performance were more likely to be identified as responders on the MT reading comprehension measure (both final status and growth methods) is encouraging. It suggests that teaching reading comprehension strategies should not be restricted to only students with grade-level reading or better. In our sample, students with relatively poorer word reading ability demonstrated successful use of comprehension strategies on the MT reading comprehension measure. Various readability formulas estimated the MT passages to be between a high 3rd grade and a low 5th grade level, which suggests that these passages are at an appropriate level for struggling 4th and 5th grade students.

The very poor agreement between methods for identifying individual students as responders can be interpreted as an encouraging finding, rather than a disappointing one. Reading comprehension is a complex construct and comprehension measures vary considerably in the skills they address, and the students they reveal as poor and good readers (Cutting & Scarborough, 2006; Keenan & Meenan, 2014; Keenan, Hua, Meenan, Pennington, Willcutt & Olson, 2014). Consider a student who receives an intervention but fails to achieve normalization on a standardized norm-referenced measure. The same student might have acquired new skills and demonstrated response to the intervention in a subtler way, which could be identified by looking at growth on a more proximal measure. The same student might have acquired new skills that were not assessed by the particular reading comprehension measure used to determine response. The use of a different standardized norm-referenced measure might reveal reliable growth or normalized achievement levels for that same student. Without considering the utility of proximal and multiple measures of reading comprehension, the nuances of students' response to intervention can be overlooked.

The use of proximal measures could also provide practitioners with more specific information about the skills in which students are proficient, and the skills for which they need more instructional support. However, classroom teachers cannot and should not be expected to develop tests proximal to the various interventions they use with their students. Intervention researchers and developers should provide proximal measures (as well as information on their usage) as part of the treatment package so practitioners can incorporate them into decisions regarding whether an intervention may be appropriate for a particular student and whether that student has responded to the intervention. By conducting classification analyses similar to those in this study before making a treatment program available to the public, intervention researchers could provide more information to practitioners about the children who are most likely to benefit from that specific treatment program.

Researchers and practitioners alike are struggling to define, measure, and improve students' reading comprehension. The view of proximal measures as less useful than FT measures (i.e. standardized, norm-referenced tests of reading comprehension) because they are overly aligned with the intervention should be reconsidered. Proximal measures of reading comprehension may shed light on whether students have improved their performance on tasks and texts similar to what they have been exposed to in the intervention. According to Catts & Kamhi (2017), the multidimensional nature of reading comprehension and the variability of performance of the same students across different measures, tasks, and passages means that "instruction will be more effective when tailored to students' abilities with specific texts and tasks" (p. 2). If this is the case, then the determination of student response to a reading comprehension intervention should include evaluations of growth and final status on both NT and FT measures of reading comprehension, resulting in a more nuanced and complete view of response.

Limitations

Findings from the current study provide important information about the predictors of response to a multi-component reading comprehension intervention and the nuances of varying the methods and measures of determining responsiveness. Nevertheless, there are important study limitations. First, findings apply only to the current study sample and the specific intervention implemented. As evidenced by the literature review, various multi-component reading comprehension interventions (including the current program) find very different results regarding the overall and differential effectiveness of the program. This could be caused by differences in sample characteristics, intervention duration, or components, and how much teaching experience the interventionists who implemented the programs had.

Second, the analyses conducted in the present study were somewhat underpowered, which probably undermined my ability to detect significant predictors of response. While the sample was relatively large, logistic regression works best when the proportions of successes (responders) and failures (non-responders) are relatively equal. The proportions of responders and non-responders was variable in this study, so some of the models (i.e. GMRT growth) fit the data more poorly than others. Third, the near- and mid-transfer reading comprehension measures did not have adequate psychometric properties for use with the RCI method, so I used an alternative growth method to determine response on these criterion-referenced measures. This criterion for the growth method used these measures was whether the student demonstrated greater than average growth from pre- to post-treatment. This resulted in approximately half of the sample identified as responsive using this method. Additional testing, development and refinement of these measures should be done in order to improve their use with a growth method of determining response.

References

Cain, K. and Oakhill, J. (2006) Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology, (76),* 683-696.

Catts, H. W., & Kamhi, A. G. (2017). Prologue: Reading comprehension is not a single ability. *Language, Speech, and Hearing Services in Schools*, *48*(2), 73-76.

Cho, E., Roberts, G. J., Capin, P., Roberts, G., Miciak, J., & Vaughn, S. (2015). Cognitive Attributes, Attention, and Self-Efficacy of Adequate and Inadequate Responders in a Fourth Grade Reading Intervention. *Learning Disabilities Research & Practice*, *30*(4), 159-170.

Christensen, L., & Mendoza, J. L. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17*, 305–308.

Cirino, P. T., Romain, M. A., Barth, A. E., Tolar, T. D., Fletcher, J. M., & Vaughn, S. (2013). Reading skill components and impairments in middle school struggling readers. *Reading and Writing*, *26*(7), 1059-1086.

Compton, D. L. (2000). Modeling the response of normally achieving and at-risk first grade children to word reading instruction. *Annals of Dyslexia*, *50*(1), 53-84.

Compton, D. L., Fuchs, D., Fuchs, L. S., & Bryant, J. D. (2006). Selecting at-risk readers in first grade for early intervention: A two-year longitudinal study of decision rules and procedures. *Journal of Educational Psychology*, *98*(2), 394.

Compton, D. L., Gilbert, J. K., Jenkins, J. R., Fuchs, D., Fuchs, L. S., Cho, E., ... & Bouton, B. (2012). Accelerating chronically unresponsive children to tier 3 instruction: What level of data is necessary to ensure selection accuracy? *Journal of Learning Disabilities*, *45*(3), 204-216.

Cutting, L. E., & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific studies of reading*, *10*(3), 277-299.

Edmonds, M. S., Vaughn, S., Wexler, J., Reutebuch, C., Cable, A., Tackett, K. K., & Schnakenberg, J. W. (2009). A synthesis of reading interventions and effects on reading comprehension outcomes for older struggling readers. *Review of Educational Research*, *79*(1), 262-300.

Fletcher, J. M. (2006). Measuring reading comprehension. *Scientific Studies of Reading*, *10*(3), 323-330.

Frijters, J. C., Lovett, M. W., Sevcik, R. A., & Morris, R. D. (2013). Four methods of identifying change in the context of a multiple component reading intervention for struggling middle school readers. *Reading and writing*, *26*(4), 539-563.

Fuchs, D., Fuchs, L. S., & Burish, P. (2000). Peer-assisted learning strategies: An evidence-based practice to promote reading achievement. *Learning Disabilities Research & Practice*, *15*(2), 85-91.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2004). Identifying reading disabilities by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disability Quarterly*, *27*(4), 216-227.

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

Fuchs, D., Fuchs, L. S., & Stecker, P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Exceptional Children*, *76*(3), 301-323.

Fuchs, D., Hendricks, E., Walsh, M. E., Fuchs, L. S., Gilbert, J. K., Zhang Tracy, W., ... & Peng, P. (2018). Evaluating a Multidimensional Reading Comprehension Program and Reconsidering the Lowly Reputation of Tests of Near-Transfer. *Learning Disabilities Research & Practice*, *33*(1), 11-23.

Fuchs, D., Mock, D., Morgan, P. L., & Young, C. L. (2003). Responsiveness- to- intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice*, *18*(3), 157-171.

Fuchs, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research & Practice*, *18*(3), 172-186.

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of learning disabilities*, *45*(3), 195-203. Use caps

Gerber, M. M. (2005). Teachers are still the test: Limitations of response to instruction strategies for identifying children with learning disabilities. *Journal of Learning Disabilities*, *38*(6), 516-524.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior therapy*, *15*(4), 336-352. Use caps

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2006). Gates-MacGinitie Reading Tests (4th ed.). Rolling Meadows, IL: Riverside Publishing.

Jacobson, N. S., & Truax, P. (1991). Clinical significance: a statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*(1), 12.

Keenan, J. M., Hua, A. N., Meenan, C. E., Pennington, B. F., Willcutt, E., & Olson, R. K. (2014). Issues in identifying poor comprehenders. *L'annee psychologique*, *114*(4), 753.

Keenan, J. M., & Meenan, C. E. (2014). Test differences in diagnosing reading comprehension deficits. *Journal of learning disabilities*, *47*(2), 125-135.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Maassen, G. H. (2004). The standard error in the Jacobson and Truax Reliable Change Index: The classical approach to the assessment of reliable change. *Journal of the International Neuropsychological Society*, *10*(6), 888-893.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). TOSWRF: *Test of silent word reading fluency: Examiner's manual*. Austin, TX: Pro-Ed.

Mastropieri, M. A., & Scruggs, T. E. (2005). Feasibility and consequences of response to intervention: Examination of the issues and scientific evidence as a model for the identification of individuals with learning disabilities. *Journal of Learning Disabilities*, *38*(6), 525-531.

Pickering, S., & Gathercole, S. E. (2001). *Working memory test battery for children (WMTB-C)*. Psychological Corporation.

Ritchey, K. D., Silverman, R. D., Montanaro, E. A., Speece, D. L., & Schatschneider, C. (2012). Effects of a tier 2 supplemental reading intervention for at-risk fourth-grade students. *Exceptional Children*, *78*(3), 318-334.

Roberts, G., Torgesen, J. K., Boardman, A., & Scammacca, N. (2008). Evidence-based strategies for reading instruction of older students with learning disabilities. *Learning Disabilities Research & Practice*, *23*(2), 63-69.

Scarborough, H. S., Sabatini, J. P., Shore, J., Cutting, L. E., Pugh, K., & Katz, L. (2013). Meaningful reading gains by adult literacy learners. *Reading and Writing*, *26*(4), 593-613.

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning and Individual Differences*, *18*(3), 308-315.

Speece, D. L., Ritchey, K. D., Silverman, R., Schatschneider, C., Walker, C. Y., & Andrusik, K. N. (2010). Identifying children in middle childhood who are at risk for reading problems. *School Psychology Review*, *39*(2), 258.

Storch, S. A., & Whitehurst, G. J. (2002). Oral language and code-related precursors to reading: evidence from a longitudinal structural model. *Developmental psychology*, *38*(6), 934.

Stuebing, K. K., Barth, A. E., Trahan, L. H., Reddy, R. R., Miciak, J., & Fletcher, J. M. (2015). Are child cognitive characteristics strong predictors of responses to intervention? A meta-analysis. *Review of Educational Research*, *85*(3), 395-429.

Swanson, J. M., Schuck, S., Porter, M. M., Carlson, C., Hartman, C. A., Sergeant, J. A., ... & Wigal, T. (2012). Categorical and dimensional definitions and evaluations of symptoms of ADHD: history of the SNAP and the SWAN rating scales. *The International journal of educational and psychological assessment*, *10*(1), 51.

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.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2012). *TOWRE-2 Examiner's Manual*. Austin, TX: Pro-Ed.

U.S. Department of Education. (2004). Individuals with Disabilities Improvement Act of 2004, Pub. L. 108-466. Federal Register, 70, 35802–35803.

Vadasy, P. F., Sanders, E. A., & Abbott, R. D. (2008). Effects of supplemental early reading intervention at 2-year follow up: Reading skill growth patterns and predictors. *Scientific Studies of Reading*, *12*(1), 51-89.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities rResearch & Practice*, *18*(3), 137-146.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Small, S. G., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*, *88*(4), 601.

Vellutino, F. R., Scanlon, D. M., Small, S., & Fanuele, D. P. (2006). Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: Evidence for the role of kindergarten and first-grade interventions. *Journal of Learning Disabilities*, *39*(2), 157-169.

Vellutino, F. R., Tunmer, W. E., Jaccard, J. J., & Chen, R. (2007). Components of reading ability: Multivariate evidence for a convergent skills model of reading development. *Scientific studies of reading*, *11*(1), 3-32.

Wanzek, J., Petscher, Y., Al Otaiba, S., Kent, S. C., Schatschneider, C., Haynes, M., ... & Jones, F. G. (2016). Examining the average and local effects of a standardized treatment for fourth graders with reading difficulties. *Journal of Research on Educational Effectiveness*, *9*(sup1), 45-66.

Wechsler, D. (2009). *Wechsler Individual Achievement Test* (3rd ed.). San Antonio, TX: Pearson

Wechsler, D., & Hsiao-pin, C. (2011). WASI II: *Wechsler Abbreviated Scale of Intelligence*. 2nd. Psychological Corporation.

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. (2001). *Woodcock-Johnson R III NU Tests of Achievement*. Itasca, IL: Riverside.

Figure 1. Comparison of Program Components in Reviewed Interventions

	Ritchey et al. (2012)	Wanzek et al. (2016)	Vaughn et al. (2016)	Frijters et al. (2013)
Decoding	X	X	X	X
Fluency	X	X	X	X
Vocabulary	X	X	X	X
Goal Setting		X		
Previewing	X	X		
Prediction				X
Background Knowledge				X
Text Structure		X		X
Clarifying / Monitoring	X	X	X	X
Text Based Questions	X	X	X	X
Inference Making	X	X		
Summarizing	X	X	X	X
Total Intervention Time	16 hours	60 hours	~47 hours	125 hours

Table 1. Eligibility Criteria

Domain	Year 4	Year 5
Word Reading	TOWRE SWE >20th percentile	TOWRE SWE >10th percentile (4th) or >12th percentile (5th)
Reading Comprehension	GMRT < 50th percentile	GMRT < 50th percentile
IQ	T-Score > 37 on either WASI MR or WASI V	T-Score > 37 on either WASI MR or WASI V
English Proficiency	ELDA > 4	ELDA > 4

Note: TOWRE SWE is the Test of Word Reading Efficiency Sight Word Efficiency subtest; GMRT is the Gates-MacGinitie Reading Test; WASI MR is the Wechsler Abbreviated Scale of Intelligence Matrix Reasoning subtest; WASI V is the Wechsler Abbreviated Scale of Intelligence Vocabulary subtest.

Table 2. Student Demographics (n=249)

	N	%
Grade 4	124	49.80
Male	111	47.03
Student Race		
Black or African American	104	44.44
Hispanic	71	30.34
Caucasian	44	18.80
Other	15	6.41
Free/Reduced Lunch	137	42.68
IEP	10	4.31
Retained	2	0.85

Note. Percentages are based on the number of students with reported demographic data.

Table 3. Student Pre-test Scores

Measure	Mean (SD)
WASI 2 - Matrix Reasoning [a]	12.84 (3.73)
WASI 2 – Vocabulary [a]	24.81 (4.60)
WMTB Backward Digit Recall [a]	13.86 (3.88)
TOWRE SWE [b]	95.03 (7.47)
GMRT Reading Comprehension [c]	36.82 (10.39)
WIAT III Reading Comprehension [b]	93.43 (7.56)
Mid Transfer Reading Comprehension [a]	9.30 (3.14)
Near Transfer Reading Comprehension [a]	15.17 (3.78)

Note: TOWRE SWE is the Test of Word Reading Efficiency Sight Word Efficiency subtest; GMRT is the Gates-MacGinitie Reading Test; WASI 2 is the Wechsler Abbreviated Scale of Intelligence Second Edition; [a] *= raw score;* [b] *= standard score;* [c] *= normal curve equivalent.*

Table 4. Effect of Various Raw Score Criteria

Measure	Raw score criterion	# (%) of students Responsive	
		Pre	Post
Near Transfer	50% correct	207 (83%)	242 (97%)
	75% correct	73 (29%)	174 (70%)
	87.5% correct	14 (5.6%)	82 (33%)
	95% correct	3 (1%)	8 (8%)
Mid Transfer	50% correct	180 (72%)	215 (86%)
	75% correct	68 (27%)	124 (50%)
	87.5% correct	17 (7%)	50 (20%)
	94% correct	7 (3%)	30 (12%)

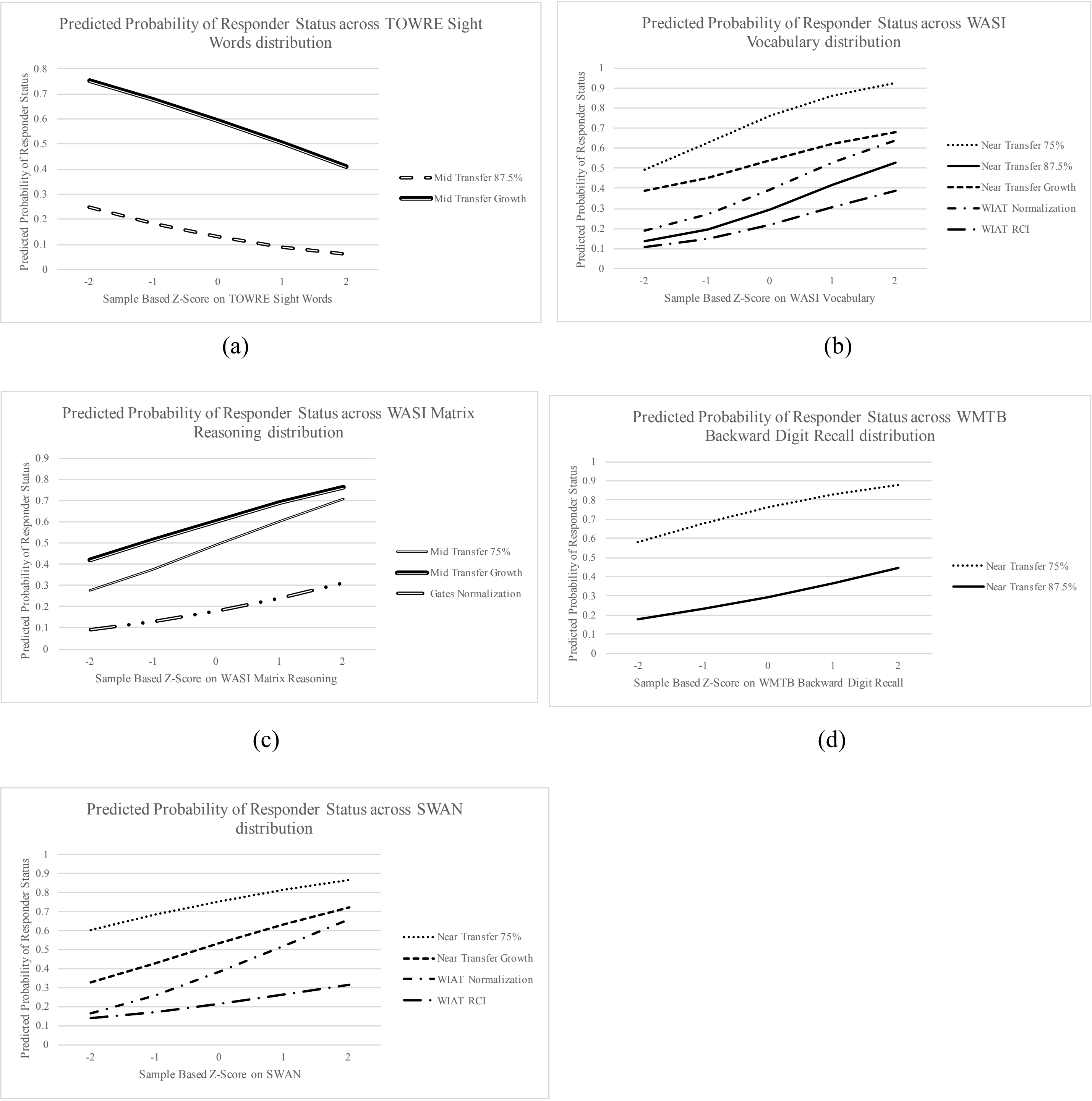
Figure 2. Comp Only Program Components Across Years

Program Component		Year 4	Year 5
Before	Text Features	√	√
	Text Structure	√	√
	Vocabulary	√	√
	Prediction	√	√
	BK Media	√	√
During	Clarify & Connect	✗	√
After	Main Idea	√	√
	Big Idea	√	√
	Factual & Inference Questions	√	√

Table 5. Odds Ratios for Predictors of Responder Status

	Grade	TOWRE SWE	WASI Vocab.	WASI Matrix Reasoning	WMTB Backward Digit Recall	SWAN	Pre-treatment Near Transfer	Pre-treatment Mid Transfer	Pre-treatment WIAT	Pre-treatment Gates
Near Transfer 75% correct			1.83**		1.49*	1.42~	2.55***	--	--	--
Near Transfer 87.5% correct			1.64**		1.37~		2.05***	--	--	--
Near Transfer Growth			1.36~			1.50*	0.19***	--	--	--
Mid Transfer 75% correct				1.54**			--	2.84***	--	--
Mid Transfer 87.5% correct		0.68*				1.45~	--	3.04***	--	--
Mid Transfer Growth		0.70*		1.42*			--	0.20***	--	--
Normalization WIAT			1.67**			1.74**	--	--	2.77***	--
RCI WIAT			1.52*				--	--	0.52**	--
Normalization Gates				1.42*			--	--	--	2.84***
RCI Gates							--	--	--	0.56**

*Note: TOWRE SWE is Test of Word Reading Efficiency, Sight Word Efficiency subtest; WASI is Wechsler Abbreviated Scale of Intelligence; WMTB is Working Memory Test Battery; SWAN is Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale; WIAT is Wechsler Individual Achievement Test, Reading Comprehension subtest; Gates is Gates MacGinitie Reading Comprehension subtest; * $p \leq 0.05$; ** $p \leq 0.01$; *** $p \leq 0.001$; ~ $0.06 \leq p \leq 0.08$*


Predicted Probability of Responder Status across TOWRE Sight Words distribution

(a)

Predicted Probability of Responder Status across WASI Vocabulary distribution

(b)

Predicted Probability of Responder Status across WASI Matrix Reasoning distribution

(c)

Predicted Probability of Responder Status across WMTB Backward Digit Recall distribution

(d)

Predicted Probability of Responder Status across SWAN distribution

(e)

Figure 3. Graphs depicting predicted probabilities of responder status for various operationalizations of adequate response across the distributions of five predictor variables. The predicted probabilities were calculated while holding all other variables at their means.

Table 6. Proportion of Sample Identified as Responders

	Final status	Growth	Final Status or Growth
Near Transfer (75% correct)	.70	.50	.80
Near Transfer (87.5% correct)	.33		.61
Mid Transfer (75% correct)	.50	.54	.72
Mid Transfer (87.5% correct)	.20		.60
WIAT	.42	.24	.49
Gates	.23	.19	.30

Appendix A: Measures

Reading Comprehension. Two standardized tests of reading comprehension were administered: The Reading Comprehension subtest of the Wechsler Individual Achievement Tests-III (WIAT; Wechsler, 2009) and the Gates-MacGinitie Reading Tests-4 (Gates-MacGinitie; MacGinitie, MacGinitie, Maria, & Dreyer, 2000). On the WIAT, students read a selection of texts (typically 3) and answer open-ended factual and inferential questions about them. Questions are read aloud by the tester and students may view the texts as they answer them. On the Gates-MacGinitie, students read 11 short passages and answer multiple-choice questions about them. Students are given 35 minutes for the test.

IQ. Two subtests from the *Wechsler Individual Scale of Intelligence-2* (Wechsler, 2011) were used to obtain a brief estimate of students IQ. The Vocabulary subtest evaluates expressive vocabulary and verbal knowledge. For each item, students see a picture or hear a word read aloud by the tester and must identify the picture or provide a definition. The Matrix Reasoning subtest assesses non-verbal reasoning with tasks that require pattern completion, classification, analogy, and serial reasoning. For each item, students select one of five options that best completes a visual pattern.

Word Reading. Word reading was assessed using the Sight Word Efficiency subtest of the Test of Word Reading Efficiency-2 (TOWRE; Torgesen, Wagner, & Rashotte, 2012) The Sight Word Efficiency subtest requires students to read as many sight words as possible in 45 seconds from a list of words that gradually increase in difficulty.

Working Memory. Working memory was assessed at pretreatment only using the Backward Digit Recall subtest of the Working Memory Test Battery for Children (WMTB; Pickering & Gathercole, 2001) Students are required to recall in backwards order a set of numbers read aloud by the tester. The test is divided into spans of six items of increasing difficulty, ranging from 2 to 7 digits. We modified the standard administration of this test by discontinuing it when a student incorrectly answered four instead of three items within a span. Therefore, from this measure, only raw scores are used in analyses.

Appendix B: Additional Tables

Table 7. Agreement between Final Status Criteria and Growth Criteria by Measure

	Cohen's Kappa
Near Transfer (75% correct)	0.19*
Near Transfer (87.5% correct)	0.21*
Mid Transfer (75% correct)	0.19*
Mid Transfer (87.5% correct)	0.11*
WIAT	0.33*
Gates	0.47*

*Note: * p≤0.05*

Table 8. Agreement between Measures using Final Status Criteria

	Mid Transfer (75% correct)	Mid Transfer (87.5% correct)	WIAT	Gates
Near Transfer (75% correct)	0.33*	0.14*	0.24*	0.15*
Near Transfer (87.5% correct)	0.28*	0.25*	0.30*	0.29*
Mid Transfer (75% correct)	--	--	0.27*	0.21*
Mid Transfer (87.5% correct)	--	--	0.18*	0.27*
WIAT			--	0.22*

*Note: * p≤0.05*

Table 9. Agreement between Measures using Growth Criteria

	Mid Transfer	WIAT	Gates
Near Transfer	0.04	0.06	-0.05
Mid Transfer	--	0.01	-0.03
WIAT		--	-0.04

*Note: * p≤0.05*