Hammill Institute on Disabilities

Journal of Learning Disabilities
2020, Vol. 53(6) 428–443

DOI: 10.1177/0022219420920379
journaloflearningdisabilities.sagepub.com
SAGE

Are Individual Differences in Response to Intervention Influenced by the Methods and Measures Used to Define Response? Implications for Identifying Children With Learning Disabilities

Emma L. Hendricks, PhD[1] and Douglas Fuchs, PhD[1]

Abstract

Response to intervention (RTI) has been promoted for nearly 20 years as a valid supplement to or alternative method of learning disability (LD) identification. Nevertheless, important unresolved questions remain about its role in disability identification. We had two purposes when conducting this study of 229 economically and racially diverse poor readers in Grades 4 and 5 in 28 public elementary and middle schools in Nashville. First, we examined predictors of the children's response to a reading comprehension tutoring program. Second, we explored the utility of different methods (growth vs final status) and measures (near- and mid-transfer vs far-transfer) in operationalizing "response," and whether these contrasting methods and measures identified similar children. Findings indicated students with higher pretreatment scores on expressive vocabulary, nonverbal IQ, teacher ratings of attention, and reading comprehension measures were more likely classified as responsive with final status methods. Students with lower pretreatment comprehension scores were more likely identified as responsive with growth methods. These and other findings suggest "response" is strongly context dependent, raising questions about the validity of RTI as a means of disability identification.

Keywords

RTI, reading comprehension, learning disabilities identification

The 2017 National Assessment of Educational Progress (NAEP) results indicated that 32% of fourth grade students and 68% of fourth grade students with disabilities scored below the Basic level in reading. Children scoring below a Basic level are unlikely to make simple inferences about characters or plot events, provide details to support interpretations, or identify main ideas in information text. The large number of children below this level of performance signifies that educators, researchers, and policymakers in the United States are simply not doing enough. This seems especially true for at-risk readers with and without disabilities in the late elementary and middle grades. However, improving reading comprehension is as complex a task as the act of comprehension itself, which, as is well known, requires the coordination of many cognitive processes, knowledge bases, and strategies and is influenced by text and task demands (Cain et al., 2004; Kintsch & Van Dijk, 1978; RAND Reading Study Group, 2002).

Two Randomized Control Trials

Mindful of the problem and challenge, we developed a comprehensive tutoring program that addresses some of the complexity of reading comprehension for a large group of economically and racially diverse intermediate-grade students with adequate word reading but poor comprehension. The program's purpose is to teach strategies for understanding information (e.g., social studies) texts. We also developed near-transfer (NT) and mid-transfer (MT; proximal) measures of reading comprehension to be used in combination with commercially developed far-transfer (distal) measures to obtain a relatively comprehensive estimate of program effects.

Latent variables—one for the NT and MT measures and one for the far-transfer tests—were created to compare two variations of the tutoring program against each other and against controls in randomized control trials (RCTs)

[1]Vanderbilt University, Nashville, TN, USA

Corresponding Author:
Emma L. Hendricks, PhD, 1600 Court Street, Martinez, CA, 94553, USA.
Email: Hendricks.emma@gmail.com

Douglas Fuchs, PhD, 417C One Magnolia Circle, Peabody College, Vanderbilt University, Nashville, TN 37203, USA.
Email: doug.fuchs@vanderbilt.edu

conducted in two successive years. In the first RCT, a core (or base) comprehension treatment (COMP) was compared with COMP with a working memory (WM) component (WM + COMP) and with controls. Both tutored groups significantly outperformed controls on a test of knowledge acquisition (*ES*: 1.93 and 1.93 for the two groups, respectively) and had reliably stronger scores on the NT and MT latent measure (*ES*: 0. 47 and 0.35 for the groups, respectively). However, the tutored groups did not outperform controls on the latent far-transfer measure (*ES*: 0.20 and 0.04).

In the second RCT, the COMP treatment and COMP with an explicit transfer component (Transfer + COMP) were compared with controls. Again, both tutored groups significantly outperformed controls on the knowledge acquisition test (*ES*: 2.38 and 2.25) and the latent NT and MT reading comprehension measure (*ES*: 0.41 and 0.66), but not on the far-transfer measure (*ES*: −0.08 and 0.14). This is not an uncommon finding among studies in which researchers attempt to strengthen comprehension in older students (e.g., Edmonds et al., 2009; Roberts et al., 2008). Nevertheless, we wondered whether this frequently observed pattern of effects—whereby treated children perform better than controls on proximal but not distal measures—might belie a more nuanced accounting of students' response to this kind of intervention.

In this article, we describe a secondary analysis of the combined data from these two RCTs to explore whether some students benefited from these tutoring programs more than others. There are multiple ways to test for students' differential response, one of which is moderation analysis. Wanzek et al. (2016) used moderator analysis to evaluate the effects of a reading intervention for at-risk intermediate-grade students. Only the treated students with higher pretreatment comprehension scores significantly outperformed controls.

Frijters et al. (2013) used an alternative approach, namely, classification analysis. Frijters et al. operationalized response with multiple methods—normalization, growth curve estimates, and reliable change index scores. Then they used binary logistic regression to investigate predictors of response as defined by these various methods and by a commercially developed test of reading comprehension. An intriguing finding was that methods of response strongly affected predictors of response. When the methods were compared against each other, the researchers identified nearly completely different subsamples of responsive students: Growth curve and normalization tended to identify responders who had lower and higher pretreatment cognitive and reading skills, respectively.

We wondered whether we might find similar results if we conducted our own classification analysis based on Frijters et al.'s (2013) work, but involving younger children who had participated in a tutoring program to strengthen their understanding of information texts and whose performance was indexed by both experimenter-made and commercially developed tests of comprehension. Like Frijters et al., we asked if and how the tutoring program's (presumed) differential effects had been influenced by our choice of methods and measures. Unlike Frijters et al., we used multiple reading comprehension measures to help define "response." Such exploration of the possible importance of both methods and measures has been infrequently investigated, but may have significant implications for the use of response to intervention (RTI) as a means of disability identification and as a process by which children with serious learning problems obtain appropriate instruction. The possible influence of methods and measures on educators' perceptions of who is and is not a responder—and who gets more intensive and costly intervention—seems especially important with regard to reading comprehension, a construct widely recognized as challenging to measure (e.g., Catts & Kamhi, 2017; Clemens & Fuchs, 2019; Keenan & Meenan, 2014).

RTI

In principle, a well-functioning RTI framework allows schools to deploy resources more efficiently through early identification/prevention and by ensuring that the focus and intensity of interventions are matched to struggling students' needs. (This framework has also been described as "multi-tiered systems of support" [MTSS]. We use the RTI designation because it typically addresses both disability identification and multiple levels of intervention, whereas MTSS does not; cf. D. Fuchs et al., 2010.) An RTI system, however, can only be effective and efficient when "response" is meaningfully defined. To illustrate, when the criterion for adequate response is set too high, resources may be squandered by providing intensive and costly attention to students who would develop adequately without it. If the criterion is set too low, students may not receive the instructional supports they truly need. Because there are many ways of operationalizing RTI (e.g., Compton, 2006; Frijters et al., 2013; D. Fuchs et al., 2004; L. S. Fuchs, 2003), its study by researchers and use by practitioners must be understood in light of how response/no-response is defined and how this may influence the identification of who responds to a particular intervention.

Methods of Indexing Response: Final Status Versus Growth

Whereas the operationalization of response in RTI frameworks has been discussed by the educational community (cf. Consortium for Evidence-Based Early Intervention Practices, 2010; Learning Disabilities Association, 2010; Schatschneider et al., 2008), research on it has been infrequent. See Barth et al. (2008) and D. Fuchs et al. (2004, 2008) as examples of such work. Operationalizing response may be considered in two ways, the first of which is the

method used. For example, should response be determined by whether a student's score rises to average or near-average performance ("final status"), or should it be defined as improvement ("growth")? If growth, then how much growth represents meaningful change?

Final status. In the research literature, "normalization" (Torgesen et al., 2001) is a widely used final-status index of response. Those who use it typically define adequate response as a posttreatment standard score of 90 or greater on a commercially developed test. Normalizing an at-risk child's academic performance is the desired result of many interventionists because it is believed to signal meaningful change and intervention success. However, many evaluations of reading comprehension interventions, especially those involving older students, fail to find positive effects on commercial tests of far-transfer (Edmonds et al., 2009; Roberts et al., 2008). Because of this, normalization tends to be viewed as a conservative, or high-bar, approach to identify responders, at least as regards reading comprehension.

Another concern about normalization is that it is often affected by initial levels of performance. That is, students with stronger pretreatment reading scores, say, are more likely to be eventually "normalized." This is partly because 1 *SD* below the mean, or a standard score of 85, is commonly used to identify at-risk readers. So, a student with an initial score of 85 who raises her performance to 90 at posttreatment may be viewed as normalized despite a change of only five standard score points. This 5-point difference may be less than the standard error associated with the reading measure (cf. Frijters et al., 2013). Notwithstanding these concerns, we used normalization as a final-status method of response in our secondary analysis.

Growth. There are at least several methods of defining response in terms of growth. These include *within-individual gains replicated over tests*, which requires students to demonstrate positive change across multiple measures of reading comprehension (Scarborough et al., 2013); *reliable change index* scores (Jacobson & Truax, 1991); *growth curve estimates* (e.g., Compton, 2000; Vadasy et al., 2008); *curriculum-based measurement* (CBM) slope (e.g., D. Fuchs et al., 2004); and *limited norm criterion* (L. S. Fuchs, 2003). Although growth curve estimates and CBM slopes are often chosen to index change, both require a greater number of data points than pre-/posttesting. For this reason, we could not explore their utility.

Similarly, and for the same reason, we could not explore *dual discrepancy* (e.g., L. S. Fuchs & Fuchs, 1998), which calls for comparing students' (typically CBM) slope and final status to their classmates' performance or to a normative population. Students are identified as nonresponsive when both their slope and final status are at least 1 *SD* below the mean of the referent group. Unsurprisingly, this method identifies fewer nonresponders than growth alone or final status alone (McMaster et al., 2005).

One of the two growth methods we indeed studied was the just-mentioned reliable change index (Jacobson & Truax, 1991). It is calculated by dividing the change in a student's pre-to-posttreatment score by the standard error of the difference, reflecting a belief that "significant change is the degree of gain necessary to exceed the unreliability of the outcome measure" (Frijters et al., p. 542). We also used a "limited norm criterion" (L. S. Fuchs, 2003), which was exclusively based on our tutored sample, and which compares each student's growth to that of all other tutored students.

Measures of Response: Commercially Developed Versus Experimenter-Made

Research on response to instruction should also consider measures—perhaps especially so with regard to reading comprehension instruction. Commercially developed measures of reading comprehension vary in numerous ways. This variation includes genres, length of passages, question types, response formats, and more. Cutting and Scarborough (2006), Keenan and Meenan (2014), and others have shown that such tests probe different dimensions of the construct, correlate weakly-to-moderately with each other, and sometimes identify different groups as adequate (or inadequate) readers. Therefore, the choice of reading comprehension measure, apart from the nature or strength of an intervention, may be expected to influence findings about which students are responsive or not.

Most commercially developed tests of reading comprehension are considered far-transfer measures because their developers usually assess a small and arbitrarily chosen subset of skills and strategies with texts and tasks unfamiliar to many children. By definition and design, these measures do not align with most program developers' interventions. Unsurprisingly, and as mentioned, evaluators of such interventions who rely on far-transfer tests often find small to null effects. Larger effects tend to be obtained on experimenter-made measures deliberately aligned—more or less closely—with instruction (e.g., Edmonds et al., 2009; D. Fuchs et al., 2018). Whereas one might expect program developers and others to view near- and far-transfer measures as complementary and important, one or the other is usually selected for use.

Present Study

As indicated, this study is a secondary analysis of data from two consecutive years of intervention research. The aim of the research was to develop and validate a multicomponent reading comprehension program for fourth and fifth graders with adequate word-reading but inadequate comprehension. In both years, the two treatment groups together

outperformed controls, but performed similar to each other, on a knowledge acquisition test and NT and MT tests of reading comprehension. The secondary analysis was conducted on the merged databases of these studies with two purposes in mind. First, we applied logistic regression analyses to the tutored students' performance to identify pretreatment predictors of response. Such exploration may eventually lead to more efficient screening and a better understanding of those likely to benefit from the intervention.

Second, we investigated whether various combinations of final status and growth methods, and commercially developed and experimenter-made measures, change the variables that predict response and influence the identification of which children are responsive. Making use of NT, MT, and far-transfer measures of comprehension might lead to more nuanced and accurate identification of students with learning disabilities (LDs), an identification that accounts for both how much children learn from instruction and whether they transfer that knowledge to situations that vary in degree of similarity to, or familiarity with, the actual instruction. Implicit in this exploration is recognition of the possibility that response may be more profitably understood as a "graduated response" than as a binary "response/no-response."

As explained, the intent and methods of our study were influenced by Frijters et al.'s (2013) work. Nevertheless, our sample was younger (fourth and fifth graders vs. sixth and eighth graders) and better word-level readers. Our tutoring program was more narrowly focused (only comprehension vs. comprehension plus word-level reading) and briefer (33 instructional hours vs. 125 hr). Perhaps most importantly, we systematically explored the influence of experimenter-made NT and MT measures and commercially developed far-transfer tests on an understanding of response. We believe becoming more knowledgeable about the effects of methods and measures used to define response is necessary for more successful implementations of policies like RTI. Our research questions were as follows:

Research Question 1: Which child-level variables best predict response?
Research Question 2: How do methods and measures of response influence the predictors?
Research Question 3: What proportion of tutored students were identified as responsive by each combination of measure and method?
Research Question 4: Do these different combinations identify different responsive students?

Method

Participants

Student selection and eligibility. Student data came from Years 4 and 5 of a 5-year program of research to develop an efficacious multicomponent reading comprehension intervention for fourth and fifth graders. Its aim, as indicated, was to improve understanding of information texts among students with adequate word reading but weak comprehension. Selection criteria and procedures were similar in both years, but there were also several differences.

Word reading was assessed with the Sight Word Efficiency subtest of the *Test of Word Reading Efficiency* (TOWRE; Torgesen et al., 2012). In Year 4, students were eligible for study participation if they scored above the 20th percentile on the TOWRE. However, in Year 5, we lowered the criterion to find a sample of necessary size. Fourth-grade eligible students scored above the 10th percentile; fifth grade eligible students performed above the 12th percentile. Otherwise, the selection criteria across the two years were the same. Students scored below the 50th percentile on the Reading Comprehension subtest of the *Gates-MacGinitie Reading Test* (GMRT; MacGinitie et al., 2006), and above a *T*-score of 37 on either the Matrix Reasoning or Vocabulary subtests of the *Wechsler Abbreviated Scale of Intelligence* (WASI; Wechsler & Hsiao-pin, 2011). Children were excluded from study participation if they were frequently absent, disruptive in class, or not proficient in English (as measured on the district's English Language Development Assessment).

Student demographics. Complete pre- and posttreatment data were collected on 229 students. Table 1 shows demographic information. In Year 5, in comparison with Year 4, a slightly larger proportion of the sample was Hispanic and a smaller proportion was African American. Fewer students in Year 5 received free/reduced lunch and had an Individualized Education Plan.

As mentioned, the tutoring program's purpose was to teach comprehension strategies for information texts to students with adequate word reading skills but weak comprehension, an intention that influenced our eligibility criteria. At pretreatment, our sample's mean standard score on the Sight Word Efficiency subtest of the TOWRE was 95.24 ($SD = 7.53$), or the 37th percentile. This suggests students on average had low, but arguably adequate, word reading. In contrast, their mean pretreatment normal curve equivalent on the GMRT was 36.98 ($SD = 10.44$), or a standard score of 90 and percentile score of 26. Table 1 shows the sample's pretreatment means and *SDs* on cognitive, language, and reading measures for Years 4 and 5 combined.

Student assignment to study groups. In each year, children eligible for study participation (Ns = 203 and 204 in Years 4 and 5, respectively) were assigned randomly within their schools to three study groups: two variants of the reading comprehension program and controls. Of these 407 students, 249 were assigned to treatment groups and completed the study. The final sample consisted of 229 treated students

Table 1. Performance and Demographics of Sample (n = 229).

Variable	*M*	*SD*	%
Performance			
WMTB Backward Digit Recall (raw score)	13.81	3.87	
WASI 2—Matrix Reasoning (raw score)	12.89	3.69	
WASI 2—Vocabulary (raw score)	24.78	4.59	
TOWRE SWE (SS)	95.24	7.53	
Near-Transfer Reading Comprehension (raw score)	15.08	3.67	
Mid-Transfer Reading Comprehension (raw score)	9.21	3.14	
WIAT III Reading Comprehension (SS)	93.49	7.40	
GMRT Reading Comprehension (NCE)	36.98	10.44	
Demographics			
Fourth grade			0.53
Male			0.47
Black or African American			0.44
Hispanic			0.30
Caucasian			0.19
Other race			0.07
Free/reduced lunch			0.56
IEP			0.04
Retained			0.01

Note. Proportions are based on the number of students with reported demographic data. WMTB = *Working Memory Test* Battery; WASI 2 = *Wechsler Abbreviated Scale of Intelligence*, Second Edition; TOWRE SWE = *Test of Word Reading Efficiency*, Sight Word Efficiency subtest; SS = standard score; WIAT III = *Wechsler Individual Achievement Test*, Third Edition; GMRT = *Gates-MacGinitie Reading Test*; NCE = normal curve equivalent; IEP = Individual Education Plan.

with complete pre- and posttreatment data. They came from 17 elementary schools and 11 middle schools. Since the children were tutored in pairs, we used TOWRE Sight Word Efficiency scores to match them as closely as possible.

Staff. Research assistants (RAs) were 22 master's and doctoral students who were tutors and testers. Two full-time staff members also assisted with tutoring and testing. RAs participated in extensive training and multiple fidelity checks in their roles as tutors and testers before they were permitted to work with children (see "Procedures" section).

Measures

IQ. Two subtests from the WASI (Wechsler & Hsiao-pin, 2011) were administered at pretreatment testing to obtain an estimate of IQ. Vocabulary evaluates expressive vocabulary and verbal knowledge. For each item, students see a picture or hear a word read aloud by the tester and identify the picture or provide a definition of it. Matrix Reasoning assesses nonverbal reasoning. Tasks require pattern completion, classification, analogy, and serial reasoning. For each item, students select one of five options that best completes a visual pattern. Sample-based Cronbach's alpha for the Vocabulary subtest in Years 4 and 5 were .70 and .66, respectively. For Matrix Reasoning, .57 and .52.

Working Memory. Working memory was assessed only at pretreatment with the Backward Digit Recall subtest of the *Working Memory Test Battery* for Children (WMTB; Pickering & Gathercole, 2001). Students recall in backward order a set of numbers read aloud by the tester. The test is divided into spans of six items that increase in difficulty, ranging from 2 to 7 digits. We modified the test's standard administration by stopping it when a student incorrectly answered four items instead of three items within a span. Because of this modification, we used only raw scores from the test in our analyses. Sample-based Cronbach's alpha in Years 4 and 5 were .72 and .76, respectively.

Attention. Attention was measured using the *Strengths and Weaknesses of ADHD–Symptoms and Normal-Behavior Rating Scale* (SWAN; Swanson et al., 2001). Teachers completed only the first nine of 18 items, which directed them to rate (on a 7-point Likert-type scale) their students' attention to detail and whether they listen, sustain attention, remember information, follow through, and stay organized compared with students of similar age. We obtained these ratings in fall and spring of the academic year, but used only the fall ratings in our analyses. Sample-based Cronbach's alpha in Years 4 and 5 were .97 and .98, respectively.

Word reading. Word reading, as mentioned, was assessed with the TOWRE Sight Word Efficiency subtest (Torgesen

et al., 2012), which requires students to read as many sight words as possible in 45 s from a word list that increases in difficulty. The examiner's manual reports test–retest reliability of .90 for students between 8 and 12 years, and alternative form reliabilities of .89 and .83 for 9- and 10-year-old children, respectively. Because of the timed nature of the test, we did not calculate a sample-based Cronbach's alpha.

Knowledge acquisition. This experimenter-created measure assessed whether students learned the science and social studies content of passages read during the tutoring program. It consists of 20 multiple choice items, each of which has one correct answer option and four distractors. Items assess student recall of vocabulary words and their meanings, cause-and-effect relationships, and important facts about the content in the passages. The items and questions are read aloud to students to lessen the impact of reading problems on performance.

Commercially developed tests of comprehension. Two commercially developed normative tests of reading comprehension were administered: The Reading Comprehension subtest of the *Wechsler Individual Achievement Test*, Third edition (WIAT-III; Wechsler, 2009) and the GMRT (MacGinitie et al., 2006). On the WIAT-III, students read a selection of (typically three) texts and answer factual and inferential questions about them. The questions are read aloud by the tester and students may view the texts as they answer. In Year 4, sample-based Cronbach's alphas for fourth graders at pre- and posttreatment were .70 and .68. For fifth graders, .63 and .69. In Year 5, the sample-based Cronbach's alphas at pre- and posttreatment for fourth graders were .61 and .48, and for fifth graders, .56 and .59.

On the GMRT, students have 35 min to read 11 short passages and answer multiple-choice questions about them. Sample-based Cronbach's alphas in Year 4 at fourth grade were .73 and .77 at pre- and posttreatment testing; at fifth grade, .75 and .84. In Year 5, comparable coefficients at Grade 4 were .55 and .80; at Grade 5, .72 and .83.

Experimenter-made tests of comprehension. We developed comprehension measures to align more and less closely with the tutoring program. That is, we explored student performance on NT and MT measures. Our NT test requires students to read four informational passages, each of which is between 100 and 160 words long. After reading each passage, they answer six multiple choice questions that assess whether they can identify paragraph-level and passage-level main ideas and answer factual and inferential questions. The questions are similar to those they were asked during tutoring. Each multiple-choice question has one correct answer and three distractors. Whereas the NT passages are different from the tutoring passages, which is to say students had not seen them previously, they draw from the same social studies or science topics discussed in tutoring. In Year 4, sample-based Cronbach's alphas for the NT test at pre- and posttreatment were .72 and .71, respectively. In Year 5, comparable alphas were .69 and .73.

Our MT test consists of two information passages on topics *not* addressed in tutoring. However, their format and design are similar to the tutoring program's instructional passages. Each passage has between 190 and 220 words. There are eight test questions per passage. Most are multiple choice questions with one correct answer and three distractors. (There are also "complete-the-blank" questions and "circle-the-correct-answer" questions.) The questions require students to identify paragraph-level and passage-level main ideas and answer factual and inferential questions similar to the questions asked of them in tutoring. In Year 4, sample-based Cronbach's alphas for the MT measure at pre- and posttreatment were .66 and .70; in Year 5, they were .65 and .66. The NT and MT passages and questions were written and rewritten by the research team over the course of the larger multiyear study.

Tutoring

Students were tutored three times per week, 45 min per session, for 14 to 15 weeks. Tutoring lessons were provided to student pairs in the same school who, as previously indicated, were matched on word-reading performance. They worked collaboratively on activities as Coach and Reader. Lessons were scripted to promote fidelity of implementation and to provide correction procedures for incorrect responses. As explained, in Years 4 and 5 of the study, two variations of the comprehension program were compared with each other and with controls. In Year 4, the program variants were comprehension instruction (COMP) and COMP with WM training. In Year 5, COMP was contrasted with COMP plus transfer training.

We obtained small effect sizes (Cohen's *d*) when comparing the two treatment variants against each other in Years 4 and 5 on the experimenter-made and commercially developed measures. In Year 4, effect sizes ranged from $\leq$0.10 to 0.32. In Year 5, effect sizes were $\leq$0.15. Because of these relatively small group differences, we combined tutored students across Years 4 and 5 and across treatment variations to create a single COMP group. COMP instruction reflected the combining of strategies and activities shown by previous research to promote understanding of information texts written for the intermediate grades. Our COMP instruction was organized by "before-reading," "during-reading," and "after-reading" activities.

For *before reading* in both study years, students were expected to learn vocabulary words by reading and discussing definitions in a glossary. Prior to reading each passage, they were encouraged to identify text features (titles, headings, maps, pictures, captions) and text structures (descriptive,

sequential, compare-contrast, problem-solution). They checked what they already knew about the topic of a passage and then watched videos meant to build knowledge about it. Last before reading, they made a prediction about the most important idea that they would learn from the passage. *During reading*, the children were encouraged to think while reading and to stop when confused. They were taught five clarification methods, including rereading, using background knowledge, and asking for help. They were encouraged to make connections between ideas in a passage and their own experiences, including what they may have read previously. *After reading*, students used a three-step strategy based on paragraph shrinking (D. Fuchs et al., 2000) to create a main idea for each paragraph. The same three-step strategy was used to create a big idea, or the most important idea of the entire passage. At the end of each lesson, students were expected to use an "In or Out" strategy to determine whether a question was factual (the answer could be found in the passage) or inferential (the answer would require a connection to background knowledge).

Procedures

Prior to pretreatment testing, the RAs were taught to administer and score tests in a standard manner. They received 11 hr of training across 5 weeks and were required to demonstrate to project staff at least 90% adherence to administration and scoring rules. RAs failing to meet this criterion had to repeat this fidelity check until they met the criterion. Before posttreatment testing, the RAs received an additional 2 hr of training and had to pass another round of fidelity checks for each measure in our test battery. The same 90% criterion was applied.

The RAs were trained in two 8-hr sessions across consecutive days to deliver tutoring lessons in standard fashion. They were then required to practice with a partner and to earn a minimum 90% score on a fidelity check before tutoring began. In Years 4 and 5, project staff collected tutoring fidelity data on every RA during two in-school observations and in an audio check of a third session. Across (a) Years 4 and 5, (b) the three fidelity checks of each RA, and (c) the treatment groups, program adherence ranged from 92.9% and 98.5%.

Analytic Approach

We varied response methods and reading measures when classifying students as responsive/nonresponsive. Reading measures were the commercially developed comprehension subtests of the WIAT-III and GMRT and the experimenter-created NT and MT comprehension tests. For each of these measures, we combined final status and growth methods of response. We used binary logistic regression analyses, which produce effect sizes as odds ratios, and we computed Cohen's kappa, which quantifies the chance-corrected agreement between the methods used to classify students as responsive or not.

Each logistic regression tested the predictive value of seven student-level variables (i.e., grade, pretreatment word reading, pretreatment score on the outcome measure, expressive vocabulary, nonverbal IQ, WM, and teacher ratings of attention). Researchers have found statistically significant, or marginally significant, effects for one or more of these variables with the exception of word reading (cf. Al Otaiba & Fuchs, 2002; Cho et al., 2015; Frijters et al., 2013; Ritchey et al., 2012; Wanzek et al., 2016). The seven student-level variables were entered into the model simultaneously, rather than in blocks. No stepwise regression methods were used.

Final status method. For the commercially developed norm-referenced measures, age-normed standard scores were calculated as described in the test manuals. A student with a posttreatment score of 100 (50th percentile) or greater was classified as responsive. This is a higher criterion than the typical criterion for normalization, which is a standard score of 90 (25th percentile; e.g., Torgesen et al., 2001). However, students in the Torgesen et al. study had more severe reading deficits than our students who had a mean pretreatment normal curve equivalent score of 36.88 on the GMRT, corresponding to a standard score of 90. Using the conventional 25th percentile, we could have classified 70% of our sample as responsive at pretreatment based on their WIAT-III scores. For this reason, we used what we considered was a more appropriate and meaningful criterion—the 50th percentile. Using this criterion, only 23% of the sample (53 of 229 children) was classified as responsive based on WIAT-III pretreatment scores. On the GMRT, none of the students met this criterion at pretreatment.

We modified the conventional normalization method in additional ways to use with our NT and MT measures. Insufficient resources prevented us from administering these measures to a representative sample, which is to say that we assessed only students identified as weak in reading comprehension during sample selection. So, there was no normative distribution with which to compare students' posttreatment performance. The NT and MT measures by design are aligned (more and less so) with content and strategies taught during tutoring. Thus, strong posttreatment performance would suggest that students learned the strategies presented in tutoring and that they applied them to the NT and MT passages and questions, both of which reflected similar, but not identical, content and format to what they experienced in tutoring.

As with most criterion-referenced measures, a cut-off score was required to determine whether students had performed adequately. Recognizing the arbitrariness of these scores, we explored the utility of several of them and we tried to think about them as a classroom teacher might.

Scores of 75% and 87.5% correct were chosen as final status criteria because we believe they represent meaningful levels of achievement: 87.5% of items correct on the NT and MT measures corresponds to 21 of 24 and 14 of 16 items correct, respectively. A teacher or clinician could reasonably infer that a student performing at these levels comprehended the material adequately. That said, the arbitrariness of these indices, as well as our lack of access to a normative group, should be seen as study limitations.

Growth method. Response was also classified by amount of growth demonstrated from pre- to posttreatment. Reliable change index scores were calculated for each student on the commercially developed measures. To accomplish this, we used the Jacobson–Truax formula (Jacobson et al., 1984) with Maassen's (2004) modification (see Note 1). Students were classified as responsive if the difference between their pre- and posttreatment scores was statistically significantly greater than expected after accounting for the measure's reliability, unequal pre- and posttreatment variance, and practice effects (Maassen, 2004).

In principle, the reliable change index criterion can be used with commercially developed normative measures and experimenter-made measures without normative data. However, the 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. Whenever possible, we located the necessary values from the commercially developed tests' technical manuals and entered them into the formula. The GMRT technical manual did not provide test–retest reliability data. Our estimate of this value was the correlation between fall and spring administrations of the measure, which we entered into the reliable change index calculation. This is likely a more conservative estimate of the true test–retest reliability of the GMRT. Nevertheless, the absence of a test–retest index is another study limitation.

We had less of a basis for applying the reliable change index method to our NT and MT measures. Instead, we used a "limited norm criterion" (L. S. Fuchs, 2003), which is based on only tutored students. It compares each student's growth with that of other tutored students in the sample. Average change scores were calculated on NT and MT measures from pre- to posttreatment for all tutored students. Those meeting or exceeding the average change score were classified "responsive."

Results

Predictor variables were converted to *z*-scores. Logistic regression analyses were then conducted to explore predictors of response to the tutoring, using 10 combinations of methods and measures of response. In each analysis, the overall model was statistically significant and, in most analyses, the pseudo R^2 ranged from .25 to .15. Two exceptions were the GMRT and WIAT growth models. Each had a relatively poor fit, indicated by a pseudo-R^2 value of .09 and .08. This may have been due to relatively few students identified as responders by these measures and methods. Complete results for each logistic regression analysis may be found in Supplemental Tables 4 to 13. Readers may derive probabilities from them.

Results are displayed in Table 2. The data should be understood as an increase (odds ratios greater than 1) or decrease (odds ratios less than 1) in the likelihood of classifying a child as a responder given a 1 *SD* increase on a given variable (see column headings in Table 2) relative to the sample mean with all other variables held at their respective means. Consider, for example, a student whose pretreatment performance is 1 *SD* greater than the sample mean on WASI Vocabulary and equal to the sample mean on all other predictors. This student is 1.52 times more likely than a student with average scores on all predictors to be identified as a responder at posttreatment when response is defined by the reliable change index on the WIAT. All odds ratios are presented in the table regardless of statistical significance. They should be understood as heuristic because they may have been influenced by our small sample size and relatively large number of variables and hypotheses.

Predicting Response: The Influence of Methods

Table 2 indicates that the child characteristics best predicting response depended on the methods and measures used to define response. For *final status*, the odds ratios for students' pretreatment performance on the comprehension measures were greater than 1. This indicated that children with stronger pretreatment scores were more likely classified as responders at posttreatment. For example, a student performing 1 *SD* above the sample mean on pretreatment WIAT-III, and performing equally to the sample mean on all other predictors, was 2.77 times more likely to be a responder at posttreatment than a student with all scores at the sample mean when response was defined by the final status method of normalization.

Conversely, for each *growth* method, the odds ratios for pretreatment performance on the comprehension measures were less than 1. This indicated that children with lower comprehension scores at pretreatment were more likely to be classified as responders at posttreatment when response was defined as growth, regardless of measure. So, a student with a WIAT-III score 1 *SD* above the sample mean at pretreatment, and a score equal to the sample mean on all other predictors, was 0.52 times *less* likely than a student at the mean on every predictor to be identified as a responder at posttreatment with response defined by reliable change index scores on the WIAT-III.

Table 2. Odds Ratios for Predictors for Responder Status as Influenced by Methods and Measures.

Method/measure	Grade	TOWRE SWE	WASI Vocab	WASI Matrix Reasoning	WMTB Backward Digit Recall	SWAN	Pretreatment Near-Transfer	Pretreatment Mid-Transfer	Pretreatment WIAT	Pretreatment Gates
75% correct Near-Transfer	0.90	1.20	1.83**	1.14	1.49*	1.42	2.55***	—	—	—
87.5% correct Near-Transfer	0.97	0.97	1.64**	1.27	1.37	1.14	2.05***	—	—	—
Limited norm Near-Transfer	0.99	1.16	1.36	1.17	1.22	1.50*	0.19***	—	—	—
75% correct Mid-Transfer	1.63	0.83	1.05	1.54**	0.76	1.04	—	2.84***	—	—
87.5% correct Mid-Transfer	1.76	0.68*	1.09	1.12	1.02	1.45	—	3.04***	—	—
Limited norm Mid-Transfer	1.77	0.70*	1.12	1.42*	0.94	1.19	—	0.20***	—	—
Normalization WIAT	1.23	0.87	1.67**	1.10	0.99	1.74**	—	—	2.77***	—
RCI WIAT	0.58	0.73	1.52*	0.90	1.21	1.30	—	—	0.52**	—
Normalization Gates	1.42	0.81	1.25	1.42*	1.10	1.30	—	—	—	2.84***
RCI Gates	1.70	0.81	1.37	1.18	1.22	1.31	—	—	—	0.56**

Note. TOWRE SWE = *Test of Word Reading Efficiency*, Sight Word Efficiency subtest; WASI = *Wechsler Abbreviated Scale of Intelligence*; WMTB = *Working Memory Test Battery*; SWAN = *Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Rating Scale*; WIAT = *Wechsler Individual Achievement Test*, Reading Comprehension subtest; Gates = Gates-MacGinitie Reading Comprehension subtest; RCI (WIAT and Gates) = "reliable change index"; — = the variable was not included as a predictor in the model.
*$p \leq .05$. **$p \leq .01$. ***$p \leq .001$.

In sum, students with higher pretreatment scores on the reading comprehension measures were more likely identified as responders with *final status* methods. Students with lower pretreatment scores on the same measures were more likely to be identified as responders with *growth* methods. Across the methods, children were more likely identified as responders when their pretreatment scores were greater on expressive vocabulary, nonverbal IQ, and teacher ratings of attention.

Predicting Response: The Influence of Methods and Measures

Several child-level variables proved strong predictors of response, irrespective of how response was defined. These included the Vocabulary and Matrix Reasoning subtests of the WASI and the SWAN teacher attention ratings. The odds ratios for these predictors were greater than 1. So, students with higher pretreatment scores were more likely identified as responders. Complicating this finding, however, was that WASI Vocabulary and SWAN ratings were significant predictors only when reading comprehension was defined by performance on NT and WIAT-III measures, not by the MT and GMRT measures. Conversely, performance on WASI Matrix Reasoning was a significant predictor of response on MT and GMRT tests, but not on NT or WIAT-III measures.

In a similar vein, performance on the Backward Digit Recall subtest of the WMTB was a significant predictor of response for just the NT measure and only when response was defined as 75% items correct. The TOWRE Sight Word Efficiency subtest was a significant predictor of performance on the MT measure and when response was defined as 87.5% items correct and by the "limited norm criterion." The TOWRE subtest was the lone predictor to show that students with *lower* pretreatment scores were more likely identified as responders at posttreatment, an effect that was statistically significant for both growth ("limited norm criterion") and final status methods (87.5% item-correct criterion) on the MT measure.

Figure 1 displays these findings in five graphs. Each shows the predicted probability of "responsiveness" (*y*-axis) across the distribution of a predictor variable (*x*-axis). Recall that the definition of response varied because of 10 pairings of methods and measures (e.g., growth on the MT measure or normalization on the GMRT). The dashed lines and solid lines in each graph represent two (see Figure 1a) to five (see Figure 1b) definitions or operationalizations of response for which the variable was a statistically significant, or marginally significant, predictor.

The lines show the relationship between students' pretreatment scores on a predictor variable and the likelihood of their classification as responders by that specific combination of method and measure. For example, TOWRE Sight Word Efficiency scores were identified as a significant (or marginally significant) predictor when response was defined as final status (87.5% item-correct criterion) or as growth on the MT test. Therefore, these operationalizations of response are represented by dashed and solid lines in Figure 1a. The dashed line represents the final status method (with an 87.5% item-correct criterion); the solid line, a growth method. The two lines in the figure have similar

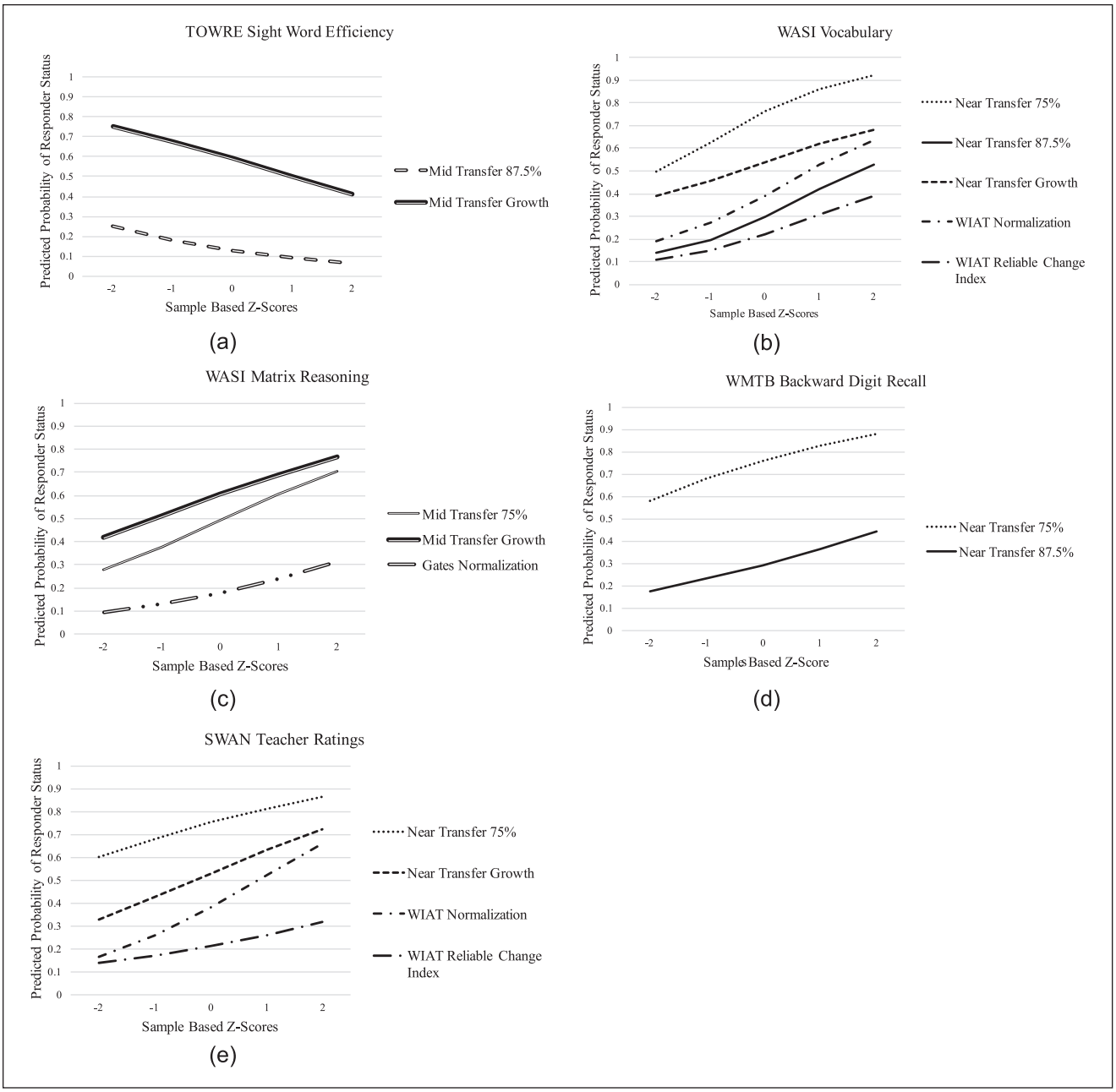

Figure 1. Predicted probabilities of student "responsiveness" status on (a) TOWRE sight words, (b) WASI Vocabulary, (c) WASI Matrix Reasoning, (d) WMTB Backward Digit Recall, and (e) SWAN teacher ratings for various methods of defining response. These probabilities were estimated using the "margins" command in Stata. Values of the predictor variable were fixed (i.e., at the mean, at 1 and 2 *SD* above and below the mean), whereas all other variables were held at their means.

slopes, but the solid line is higher. This indicates that predicted probabilities across the distribution of TOWRE scores are higher for the growth/MT definition than for the final status/MT definition. That is, students were more likely identified as responsive when response was defined by growth than by final status on MT. The likelihood of a response classification was greater when students had lower pretreatment TOWRE scores for both operationalizations.

To further explain this, consider two students and, again, Figure 1a. The first student has a TOWRE score 2 *SD*s below the sample mean (see the *x*-axis). The second student's TOWRE performance is 2 *SD*s above the sample mean. The student 2 *SDs* below the mean has a 75% chance of being viewed as a responder with a growth method on MT (see left end of solid line) when all other variables are held at the sample mean. The second student has a 40%

Table 3. Proportion of Sample Identified as Responders as Influenced by Method and Measure.

	Method						
	Final status			Growth		Final status or growth	
Measure	Norm.	75%	87.5%	RCI	Limited Norm.	75%	87.5%
Near-transfer	—	0.70	0.33	—	0.50	0.80	0.61
Mid-transfer	—	0.50	0.20	—	0.54	0.72	0.60
WIAT	0.42	—	—	0.24	—	0.49	
GMRT	0.23	—	—	0.19	—	0.30	

Note. This table should be read as follows using the first, "Near-Transfer" (NT) row as an example. It shows the proportion of the sample identified as responders using the NT comprehension measure in combination with the various final status and growth methods of response. The first column under the "Final Status" header (in the first row) has a dash, indicating that the normalization method was not used in combination with the NT measure. The next two columns show that 70% and 33% of the sample were identified as responders when the final status criterion was 75% and 87.5% items correct, respectively, on the NT measure. The first column under the "Growth" header also has a dash, indicating the RCI method was not used with the NT measure. The next column under "Growth" indicates 50% of the sample was identified as responders when the response method was the "limited norm," and so on. Norm. = normalization method; RCI = reliable change index scores method; WIAT = *Wechsler Individual Achievement Test*, Reading Comprehension subtest; GMRT = *Gates-MacGinitie Reading Test*, Reading Comprehension subtest; — = the method and measure combination was not explored.

chance (see right end of solid line). With all other variables held at the sample mean, the first student has a 25% chance of being classified as a responder with a final-status method on MT; the second student has about a 9% chance.

The remaining graphs in Figure 1 illustrate a strikingly different relationship between (a) pretreatment scores on the WASI Vocabulary and Matrix Reasoning, WMTB Backward Digit Recall, and SWAN teacher ratings and (b) the probability of a response classification. For these predictors, students scoring at the higher end of the distribution were more likely to be classified as responsive to tutoring.

Proportions of Responsive Students Identified by Combinations of Methods and Measures

Between 19% and 80% of students were identified as responders at posttreatment by the various definitions of response (see Table 3). The method-measure combinations that identified the highest proportion (80%) of students as responders was the "final status or growth" method (see Table 3) and the NT measure, followed by the "final status or growth" method and the MT measure. The method-measure combinations yielding lowest proportions of responders were (a) the reliable change index and GMRT (19%), (b) normalization and GMRT (23%), and (c) the 87.5% items-correct criterion and MT (20%).

Across all combinations of methods and measures, the average proportion of the sample identified as responsive was 38%. Averaged across methods of response, the GMRT identified the smallest proportion (30%) of responders; the NT measure identified the largest proportion (61% or 80% for the 75% or 87.5% items-correct criterion, respectively). When response was defined by the combination of "final status or growth" method and the WIAT-III, 49% of the tutored children was responsive. With the same response method and MT, 72% or 60% of the sample was responsive with the 75% and 87.5% items-correct criteria, respectively (see Table 3).

Different Method–Measure Combinations Identify Different Responsive Students

We used Cohen's Kappa to determine whether the various definitions (method–measure combinations) of response identified similar or different groups of children. Overall, the chance-corrected agreement between the various definitions of response ranged from negative or chance agreement ($k = -0.05$, *ns*) to moderate agreement ($k < 0.47$). Moderate agreement was only obtained for GMRT final status and GMRT growth methods ($k = 0.47$). Rates of agreement among the comprehension measures, when combined with the final status method, ranged from poor to fair. The NT test (87.5% correct-item criterion), MT test (75% correct-item criterion), and WIAT-III showed highest rates of agreement with other final status methods. However, the magnitude of kappa statistics for these definitions indicated only fair agreement (Landis & Koch, 1977). Agreement was poor among the comprehension measures when combined with growth methods. Each combination of method and measure registered in the negative or chance range. In sum, growth methods identified nearly completely different groups of responsive students.

Discussion

Our first objective was to determine the child-level variables that predicted response to a multicomponent reading comprehension intervention for at-risk fourth- and

fifth-grade students. A second objective was to explore how various definitions of response influenced the predictive value of the child-level variables. Below we discuss findings concerning these objectives and then implications for RTI as a means of disability identification.

Predictors of Response

Students with higher pretreatment scores on expressive vocabulary, nonverbal IQ, teacher ratings of attention, and reading comprehension were more likely classified responsive across methods and measures used to define response. Frijters et al. (2013) also found that students with higher pretreatment expressive vocabulary and nonverbal IQ were more likely identified responsive on a measure of reading comprehension, irrespective of methods of response. Results from these studies suggest, however tentatively, that it may eventually be possible to "fast-track" students to more appropriate (intensive) instructional programs on the basis of performance on a select set of cognitive and linguistic tests (cf. Compton et al., 2012).

In our study, word reading skill was *not* among the child characteristics predictive of a poor treatment response. Children with poorer pretreatment word reading still made relatively successful use of comprehension strategies on the MT measure *if* they also had relatively high teacher ratings of attention and higher scores on vocabulary and nonverbal IQ measures. This result is generally consistent with findings of Frijters et al. (2013) who found students with weaker pretreatment word reading and rapid letter naming were identified as responders when response was defined by a growth method. Findings from the two studies suggest that, without our proximal measures and comparison of final status versus growth methods, we may have mistakenly recommended the fast-tracking of students with weak pretreatment word reading—but also with higher scores on the other predictors—to a more intensive instructional program when they were likely to benefit from the tutoring. Such a possibility calls attention to the importance of the methods and measures used to define response in RTI frameworks. (We should also mention that some variables in our analyses may have exerted different effects if we had explored them independently rather than in combination with other predictor variables. This possibility may be an interesting avenue of future research.)

Methods and Measures of Response

Methods. Children with stronger pretreatment scores on the cognitive, linguistic, and reading comprehension measures (i.e., lower-risk students) were more likely identified as responsive by final status methods than by growth methods. This result is consistent with Frijters et al.'s (2013) and Wanzek et al.'s (2016) respective classification and moderation analyses, which indicated students with lower pretreatment comprehension scores were more likely viewed as responders by growth than by final status methods. Despite the apparent agreement across research teams on the influence of response methods, our findings may be partly explained by the psychometric characteristics of the measures that were paired with them.

To explain, remember that we used a "limited norm criterion" (L. S. Fuchs, 2003) to operationalize growth on our NT and MT measures. Students were deemed responders if their raw-score change from pre- to posttreatment exceeded the average change (3.5 points) of all tutored children in the sample. A student answering 22 of 24 comprehension questions correctly on the NT measure at pretreatment would not be classified as a responder, even if he achieved a perfect posttreatment score. Such ceiling effects complicate interpretations of the stand-alone importance of response methods.

Another consideration potentially confounding comparisons between final status and growth methods of response is that we paired the different growth methods with criterion-referenced measures or with norm-referenced measures but not with both. Nevertheless, across such pairings, our findings were similar: Higher risk students (with lower pretreatment scores) were more likely identified as responders with growth methods. This suggests that results may not have been entirely an artifact of the psychometric limitations or idiosyncrasies of NT and MT measures and the "limited norm criterion" response method.

Measures. Whereas, generally speaking, practitioners and researchers may not be aware of the possibility that different methods of response can identify different children as responsive, many do recognize that measures of reading comprehension are markedly different from each other and, as a result, the same child may perform adequately on one comprehension test and inadequately on another. Similarly, a growing number of researchers are recognizing that one comprehension measure may indicate an instructional program is beneficial, whereas another comprehension measure may indicate it is of little value (cf. Catts & Kamhi, 2017).

Contributing to the variation among comprehension tests is the complexity of the construct. Reading with understanding depends on an interaction of cognitive processes like attention, WM, reasoning, and inferential thinking; on sensitivity to the structure of language; on background knowledge and vocabulary development; on motivation; on the use of strategies like self-monitoring; and, of course, on word reading (e.g., Gough & Tunmer, 1986; Nation, 2009; Perfetti, 1985). Furthermore, these processes interact with text features like genre, structure, and complexity, and task demands to influence how much a reader understands and learns from text ([RAND] Reading Study Group, 2002).

The complex and covert nature of reading comprehension represents significant challenges to those attempting to measure it. Test developers must choose which dimension(s) of comprehension to measure and which ones to ignore. In the absence of a consensual definition of comprehension, test developers choose a smaller set of components around which to build their tests (just as program developers must decide which components should be targeted by their intervention programs). Indeed, test developers have created reading comprehension tests that often address uniquely different sets of skills and strategies. Not surprisingly, studies of the psychometric properties of such tests reveal that they do not correlate as strongly with each other as might be expected (e.g., Clemens & Fuchs, 2019; Cutting & Scarborough, 2006; Francis et al., 2006; Keenan et al., 2008). All this has implications for identifying students who are responsive and not, as well as reading programs that are effective and ineffective.

Consider a student who participated in a reading comprehension program and failed to reach a conventionally accepted normalization criterion on a commercially developed comprehension test. This putatively unresponsive child may nevertheless have mastered comprehension skills and strategies addressed by the program but ignored by the measure used to determine her response. A relatively proximal measure—one deliberately aligned with the instructional program—may have revealed the child's growth. Proximal measures, like our NT and MT tests, are not necessarily substitutes for the more distal, commercially developed, norm-referenced tests. Rather, they can supplement the distal tests, reveal student learning missed by them, and, arguably, lead to more valid judgments about children's response to instruction, especially with regard to reading comprehension instruction.

Study Limitations

We have already described several study limitations such as the arbitrariness of our final status criteria of 75% and 87.5% items correct. We wish to discuss several more. First, we lacked multiple data points on our study participants beyond pre- and posttreatment measurement. This precluded exploration of student response when indexed by CBM slope and dual discrepancy methods. Dual discrepancy, in particular, could have strengthened the importance of our efforts, conceptually and practically, because it combines growth and final status response methods (cf. L. S. Fuchs & Fuchs, 1998). Second, we were unable to compare tutored students' performance on our experimenter-created measures with a normative group. A normative group would have permitted a more meaningful growth criterion for the experimenter-created measures. Instead, growth on the NT and MT measures was determined by the "limited norm" method, which requires use of the sample's average growth as the criterion. This resulted in classifying about half of the tutored students as responsive on each measure. Third, we conducted a relatively large number of analyses and did not control for family-wise error. As mentioned previously, we did so in part because we viewed the study as exploratory, and much of what we did (and didn't do) was for heuristic purposes. Nevertheless, reported results may reflect Type I error.

Implications for RTI Frameworks

IQ-achievement discrepancy. The 2004 reauthorization of IDEA endorsed RTI as an important adjunct to procedures for identifying children and youth with LD. In a sense, this endorsement represented the culmination of efforts to diminish the importance of IQ-achievement discrepancy—if not to eliminate it in its entirety (e.g., Lyon et al., 2001)—as a principal component in the identification process. Such efforts began shortly after passage of PL 94-142 (aka Education of All Handicapped Children's Act of 1975), which formally established LD as a disability category. From 1975 onward, scholars, policymakers, and practitioners have criticized the discrepancy method as biased (because of the necessary use of IQ tests), unfair (because a child's academic achievement must be sufficiently below classmates' achievement to qualify for special services), and harmful (because, it has been alleged, it has contributed to the incorrect classification of children as disabled and their assignment to purportedly stigmatizing special education programs).

Perhaps the strongest and most persuasive criticism of IQ-achievement discrepancy is that it leads to arbitrary decision making. There have been at least two forms of this argument. The first is illustrated by Ysseldyke et al. (1982) who claimed their research failed to show meaningful differences between poor readers with and without a discrepancy on many tests addressing a broad range of functioning. Reschly and Hosp (2004) expressed a second argument when writing that variations in percentages of students with LD across states suggest that a child in Iowa, let's say, might be given an LD label but not in neighboring Wisconsin or Missouri.

RTI. Advocates of RTI describe it as an objective means of identifying disability because it is based on (replicable) observations of student performance on clear, meaningful academic tasks. Many proponents see it as a necessary supplement to traditional tests and procedures; for others, it is a necessary and sufficient means of identifying LD (e.g., North Carolina Department of Public Instruction Exceptional Children Division, 2015).

Such support of RTI notwithstanding, findings from this study, and several related studies (e.g., Barth et al., 2008; D. Fuchs et al., 2004, 2008), raise questions about its current

use as a means of disability identification. When educators combine methods and measures to operationalize "response," they are knowingly or otherwise creating pairings that define it differently. For a given child or group of children, one method–measure combination may signal "response"; another combination may indicate "inadequate response." Because we focused on reading comprehension, it is possible similar results would be found for other skill areas in reading and in other academic domains. Without an evidence-based consensus about which methods and measures should be used in concert to define response, RTI seems as arbitrary an approach to disability identification as IQ-achievement discrepancy with all of the attendant practice and policy-related issues and problems.

Acknowledgments

We gratefully acknowledge the contributions of the following members of our research team: Meagan Walsh, Jenny Gilbert, Wen Zhang Tracy, Sam Patton, Nicole Davis-Perkins, Wooliya Kim, Amy Elleman, Peng Peng, and Lynn Fuchs. We also thank the teachers, principals, and administrators of the Metro-Nashville Public Schools for their interest and cooperation.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper was supported in part by a grant (R324D130003) from the National Center for Special Education Research in the Institute of Education Sciences (U.S. Department of Education) and Core Grant HD15052 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The paper does not necessarily express positions or policies of the funding agencies and no official endorsement by them should be inferred.

Supplemental Material

Supplemental material for this article is available online.

Note

1. $\mathrm{RCI} = (x_2 - x_1) / \mathrm{SEM}_{\mathrm{diff}}$, where x_2 and x_1 are the student's post- and pretreatment scores, respectively. The standard error of measurement of the difference score ($\mathrm{SEM}_{\mathrm{diff}}$) 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 posttreatment scores, respectively, and r_{xy} is the test–retest reliability of the measure.

References

Al Otaiba, S., & Fuchs, D. (2002). Characteristics of children who are unresponsive to early literacy intervention: A review of the literature. *Remedial and Special Education*, *23*(5), 300–316.

Barth, A. E., Stuebing, K. K., Anthony, J. L., Denton, C. A., Mathes, P. G., Fletcher, J. M., & Francis, D. J. (2008). Agreement among response to intervention criteria for identifying responder status. *Learning and Individual Differences*, *18*, 296–307.

Cain, K., Oakhill, J., & Bryant, P. (2004). Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, *96*(1), 31–42.

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.

Clemens, N., & Fuchs, D. (2019). *Commercially developed tests of reading comprehension: Gold standard or fool's gold?* [Unpublished manuscript].

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. (2006). How should "unresponsiveness" to secondary intervention be operationalized? It is all about the nudge. *Journal of Learning Disabilities*, *39*, 170–173.

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.

Consortium for Evidence-Based Early Intervention Practices. (2010). *A response to the Learning Disabilities Association of America (LDA) white paper on specific learning disabilities (SLD) identification*. http://www.rtinetwork.org/images/content/articles/learn_about_RTI/LDAResponsefinal.pdf

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*, 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.

Francis, D. J., Snow, C. E., August, D., Carlson, C. D., Miller, J., & Iglesias, A. (2006). Measures of reading comprehension: A latent variable analysis of the diagnostic assessment of reading comprehension. *Scientific Studies of Reading*, *10*(3), 301–322.

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., Compton, D. L., Fuchs, L. S., Bryant, J., & Davis, G. N. (2008). Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. *Reading and Writing*, *21*(4), 413–436.

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., & Stecker, P. M. (2010). The "blurring" of special education in a new continuum of general education placements and services. *Exceptional Children*, *76*, 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, L. S. (2003). Assessing intervention responsiveness: Conceptual and technical issues. *Learning Disabilities Research & Practice*, *18*(3), 172–186.

Fuchs, L. S., & Fuchs, D. (1998). A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice*, *13*, 204–219.

Gough, P., & Tunmer, W. (1986). Decoding, reading and reading disability. *Remedial and Special Education*, *7*, 6–10.

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.

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), 12–19.

Keenan, J. M., Betjemann, R. S., & Olson, R. K. (2008). Reading comprehension tests vary in the skills they assess: Differential dependence on decoding and oral comprehension. *Scientific Studies of Reading*, *12*(3), 281–300.

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

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394.

Landis, J. R., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Learning Disabilities Association. (2010). *The Learning Disabilities Association of America's white paper on evaluation, identification, and eligibility criteria for students with specific learning disabilities*.

Lyon, R. G., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., . . . Olsen, R. (2001, May). Rethinking learning disabilities. In C. E. Finn, A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259–288). Thomas B. Fordham Foundation and Progressive Policy Institute.

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.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., Dreyer, L. G., & Hughes, K. E. (2006). *Gates-MacGinitie reading tests* (4th ed.). Riverside.

McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children*, *71*, 445–463.

Nation, K. (2009). Reading comprehension and vocabulary: What's the connection? In R. K. Wagner, C. Schatschneider, & C. Phythian-Sence (Eds.), *Beyond decoding: The behavioral and biological foundations of reading* (pp. 176–194). Guilford Press.

North Carolina Department of Public Instruction Exceptional Children Division. (2015, April). *Proposed policy revisions: Specific learning disabilities white paper*. https://ec.ncpublicschools.gov/gcs04-taskforce-report.pdf

Perfetti, C. A. (1985). *Reading ability*. Oxford University Press.

Pickering, S., & Gathercole, S. E. (2001). *Working Memory Test Battery for Children (WMTB-C)*. The Psychological Corporation.

Rand. (2002). Reading for understanding: Toward an R&D program in reading comprehension. RAND.

Reschly, D. J., & Hosp, J. L. (2004). State SLD identification policies and practices. *Learning Disability Quarterly*, *27*(4), 197–213.

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.

Swanson, J., Deutsch, C., Cantwell, D., Posner, M., Kennedy, J., & Spence, A. (2001). Genes and attention-deficit hyperactivity disorder. *Clinical Neuroscience Research*, *1*, 207–216.

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*. Pro-Ed.

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.

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*(suppl. 1), 45–66.

Wechsler, D. (2009). *Wechsler Individual Achievement Test* (3rd ed.). Pearson.

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

Ysseldyke, J. E., Algozzine, B., Shinn, M. R., & McGue, M. M. (1982). Similarities and differences between low achievers and students classified learning disabled. *Journal of Special Education*, *16*(1), 73–85.