Published in final edited form as:
Learn Individ Differ. 2008 September ; 18(3): 296–307. doi:10.1016/j.lindif.2008.04.004.

Agreement among response to intervention criteria for identifying responder status

Amy E. Barth[a,*], **Karla K. Stuebing**[a], **Jason L. Anthony**[b], **Carolyn A. Denton**[b], **Patricia G. Mathes**[c], **Jack M. Fletcher**[a], and **David J. Francis**[a]

[a] *University of Houston, United States*

[b] *University of Texas Health Science Center Houston, United States*

[c] *Southern Methodist University, United States*

Abstract

In order to better understand the extent to which operationalizations of response to intervention (RTI) overlap and agree in identifying adequate and inadequate responders, an existing database of 399 first grade students was evaluated in relation to cut-points, measures, and methods frequently cited for the identification of inadequate responders to instruction. A series of 543 2×2 measures of association (808 total comparisons) were computed to address the agreement of different operationalizations of RTI. The results indicate that agreement is generally poor and that different methods tend to identify different students as inadequate responders, although agreement for identifying adequate responders is higher. Approaches to the assessment of responder status must use multiple criteria and avoid formulaic decision making.

Keywords

Response to intervention

1. Introduction

The reauthorization of the Individuals with Disabilities Act (IDEA, 2004), which provides states with guidelines for operationalizing the federal definition of learning disabilities (LD), allows school districts to use a process based on students' response to quality, research-based instruction as one part of the identification process for the category of specific LD. This process, generally known as a Response to Intervention (RTI) approach, may be used as an alternative to traditional psychometric discrepancy approaches. Response to Intervention is an approach to prevention and remedial instruction that generates data that not only informs instructional decisions but may help identify students with LD (Fuchs & Fuchs, 1998; National Joint Committee on Learning Disabilities [NJCLD], 2005; Vaughn & Fuchs, 2003). Three key components of approaches to identification that incorporate RTI are (1) use of scientific, research-based instructional methods that are monitored for integrity, (2) measurement of students' response to these methods, and (3) changing instruction based on these data (Fuchs & Fuchs, 1998; Fuchs, Mock, Morgan, & Young, 2003; NJCLD, 2005; Vaughn & Fuchs, 2003).

[*] Corresponding author. *E-mail address*: aebarth@uh.edu (A.E. Barth).

These key components are typically operationalized within the framework of a multi-tiered instructional model (Bradley, Danielson, & Hallahan, 2002; Donovan & Cross, 2002; Fuchs, Fuchs, & Speece, 2002; NJCLD, 2005; Vaughn & Fuchs, 2003). In a three-tier model, Tier 1 (primary) intervention provides differentiated scientifically-based instruction to all students in general education classrooms. Data from universal screening and repeated progress monitoring over time are used to inform instructional decision making and to guide the differentiation of students and instruction. Students whose level of academic performance or rate of learning is significantly below that of their same grade peers (based on classroom, school, district, state, or national norms) are identified as at-risk. If at-risk students do not make adequate progress in Tier 1, they advance to Tier 2 (secondary or supplemental) intervention. Tier 2 provides students with more specialized instruction that permits increased intensity and more differentiation, usually through the use of small groups and additional instructional time. Progress monitoring data continues to determine intervention effectiveness and guide instructional decisions. Students who do not respond adequately to Tier 1 and Tier 2 interventions are advanced to Tier 3 (tertiary or intensive) intervention. Tier 3 represents an even more intense and differentiated intervention and may also be a point for initiating a comprehensive evaluation by a multi-disciplinary team to determine eligibility for special education. Intervention response to either Tier 2 or Tier 3 intervention (or both) has been proposed as a partial basis for disability determination.

1.1. Operationalizing RTI

From this brief description of RTI, the importance of the measurement of response should be apparent. In principal, it is meant to the primary means by which teachers determine which students enter secondary and tertiary intervention. It could even be argued that the success of RTI hinges on the establishment of criteria that delineates response-nonresponse to instruction (Fuchs, Compton, Fuchs, Bryant, & Davis, 2008). However, the establishment of a criterion has proven challenging because intervention response-nonresponse (which exists on a continuum) (Fletcher, Lyon, Fuchs, & Barnes, 2007) must be considered a binary outcome. Intervention response-nonresponse must be considered a binary outcome because students nonresponsive to instruction will be advanced through the multiple tiers of intervention whereas responsive students will not. Thus, the field needs a criterion that dichotomizes students into two groups: (1) students not responding to instruction who may be later identified as having LD and (2) students responding to instruction that will not be later identified as having LD, so that the most vulnerable students are advanced to interventions of increasing intensity and frequency.

Prior research has begun to evaluate the extent to which screening procedures that incorporate various RTI dimensions (i.e., methods of establishing adequate response, response groups, measures for assessing learning, and cut-points) successfully identify the risk pool that should enter secondary and tertiary interventions. These studies have typically reported sensitivity, specificity, and weighted kappa statistics to quantify the extent to which approaches maximize classification accuracy and minimize classification errors. Sensitivity represents the probability that students at-risk are identified by the screening procedure whereas specificity is the probability that students not at-risk are not identified by the screening procedure. Cohen’s kappa, also a measure of agreement, represents a more robust index of inter-rater reliability. However, the results of these studies are inconclusive (Fuchs et al., 2008) and do not definitively explain the extent to which different RTI criteria identify the same group of inadequate and adequate responders. For this reason, the degree of overlap among different RTI criteria (e.g., methods for establishing response, reference groups, measures for assessing learning, and cut-points) serves as the focus of this paper.

1.2. Methods of establishing inadequate response

According to Fuchs and Deshler (2007), recent implementations of RTI approaches in classroom settings have primarily measured RTI using three methods: (1) final status, represented by both “normalization and “final benchmark” methods, (2) slope-discrepancy methods, and (3) dual-discrepancy methods. Final status methods compare students’ post intervention test scores to a criterion that may represent a norm referenced score or a criterion-referenced benchmark. Slope/discrepancy models compare students’ learning rates (i.e., slopes) to the average rate of learning for a reference group (such as same grade peers from a class, district, state, or then nation) (Marsten, 1989). Students’ with slower rates of learning than the reference group (Fuchs, Fuchs, & Compton, 2004) or students whose performance is in the bottom half of the distribution (i.e., median split) are designated as inadequate responders (Vellutino et al., 1996). The dual discrepancy method (Fuchs & Fuchs, 1998; Speece & Case, 2001) compares both students’ rate of growth (i.e., slope) and level of achievement (i.e., final status) to the referent group (Fuchs, 2003; Fuchs & Fuchs, 1998; Fuchs et al., 2002; Speece & Case, 2001). Only students with achievement levels relative to a benchmark or intercept and learning rates below the reference group are considered inadequate responders (Fuchs, 2003; Fuchs & Fuchs, 1998).

1.3. Reference group

The second dimension that must be considered when determining whether a student is responding to instruction is the group to which the student is referenced. Fuchs (2003) suggests that schools frequently use three different types of reference groups: (a) a normative sample, (b) a limited norm sample, and (c) a benchmark. The normative reference group represents the full range of student abilities (e.g., how students perform relative to the 30th percentile on a norm-referenced test). A limited norm sample represents the range of student abilities for those who participated in the same intervention (e.g., how students perform relative to other students who participated in Tier 2). A benchmark approach represents a target to be attained as a function of participating in the intervention (e.g., reading 40 words correctly per minute on an oral reading fluency measure following Tier 2 instruction) and may be relative to peers in the same classroom, or school, or to some type of national benchmark.

1.4. Measures for assessing learning

The third dimension is the selection of measures used to screen for students at-risk of later academic failure, monitor student progress, and to inform classroom instruction. Four different types of measures have been typically used: (a) growth measures, (b) curriculum-based measures, (c) norm-reference tests, and (d) criterion-referenced tests. Growth measures refer to assessments that can be repeatedly administered over time and are used to measure rate of learning. Slope parameters are frequently generated from growth measures, but the final assessment can also be used as a benchmark. Curriculum-based measures, otherwise known as general outcomes measures, assess a student’s performance on either basic skill such as math, reading and spelling or content area knowledge. Norm-referenced tests are a type of test in which the score of the tested individual is compared to a sample of peers (i.e., normative sample). The translated score indicates whether the student did better or worse than the normative sample. Norm-referenced test scores allow one to measure progress against a fixed goal. Finally, criterion-referenced tests are a type of test in which student scores are compared to a criterion. Many criterion-referenced tests involve a cut-score, where the student passes if their score exceeds the cut-score and fails if it is below (i.e., 40 words read correct per minute). The cut-score often represents the degree or level of mastery students should attain to not be considered at-risk for academic failure.

1.5. Cut-points

The final dimension is the cut-point used to differentiate students into adequate and inadequate responder groups. It is not always obvious where the cut-point should be placed in order to achieve optimal decision making (Swets, 1992) because the location of the cut-point will significantly impact the types of instructional services that individual student will receive and the incidence of non-response in the sample. Thus, the location of the cut-point or decision threshold is open to debate. To date, cut-points of 0.5, 1.0, and 1.5 standard deviations below the mean (e.g., class, district, state, nation, norm-referenced sample) (Fuchs, 2003), have been employed to determine response to instruction, along with methods based on criterion-referenced benchmarks and median splits.

1.6. Previous research

Previous research has begun to manipulate these RTI dimensions (e.g., method, reference group, measure, and cut-point) to determine which combination consistently identifies the same risk pool who should enter secondary and tertiary interventions. For example,Vellutino et al. (1996) evaluated students' word reading abilities several times over the course of a multi-year study. To differentiate students who responded adequately and inadequately to intervention, they rank-ordered the students' Woodcock Reading Mastery Test scores and performed a "median-split" on word reading slopes. Students with slopes in the bottom half were designated as inadequately responding to instruction. In a similar vein, Torgesen et al. (2001) tested students with the Woodcock Reading Mastery Test after completion of a 67.5 h Tier 2 reading intervention. Students attaining a word reading accuracy scale score of 90 or below were designated as inadequately responding to instruction, students performing above the criterion were designated as adequately responsive or "normalized".

Alternately, Case, Speece, and Molloy (2003) employed the dual discrepancy method and examined whether reading difficulties varied as a function of severity. Dual discrepancy was defined as below one standard deviation below class level and slope. Results revealed that students defined as frequently dually discrepant had more severe reading deficits and obtained poorer teacher ratings of behavior. Similarly, McMaster, Fuchs, Fuchs, and Compton (2005) explored the sensitivity of the dual discrepancy approach relative to performance-level only and growth-rate only approaches. The dual discrepancy approach was defined as performance below 0.50 standard deviations below the average performer's level and slope on non-word fluency and Dolch word probes. Performance-level approach was defined as performance below the 30th percentile on the WRMT-R Word Identification and Word Attack subtests or reading less than 40 words correct per minute. The growth only approach defined limited growth as less than 10 words gained on the WRMT-R Word Identification subtest and less than 5 words gained on the Word Attack subtest; no growth as defined as zero words gained. Results indicated that the performance level approach yielded fewer inadequate responders than the dual discrepancy approach; however, several students who attained adequate levels of performance continued to present slopes that were significantly below average performers. Similarly, the 40 words correct per minute benchmark yielded many more inadequate responders than the dual discrepancy approach, with many students presenting above average slopes. Growth approaches resulted in fewer inadequate responders than the dual discrepancy approach (McMaster et al., 2005). Further, Burns and Senesac (2005) compared four definitions of dual discrepancy (i.e., student growth below the 25th, 33rd, 50th, percentiles and 1 standard deviation below the mean). Results suggest that cut-point plays a critical role in differentiating response, resulting in varying estimates of the incidence of inadequate response.

Finally, to provide greater information about different operationalizations of RTI, Fuchs et al. (2004) contrasted three measures (Dolch Word List, Nonsense Word Fluency, and CBM Oral Reading Fluency) and four methods (i.e., Dolch slope median split, nonsense word fluency slope median split, normalized posttreatment status, and benchmark posttreatment status) to judge response to Tier II intervention. Although participants selected for Tier II interventions were at least 0.5 standard deviations below the reference group on slope and level, the dual discrepancy method and varying cut-points were not further examined. Findings indicated that (a) incidence varied as a function of method (i.e., 3.5% for median split, 1.4% for normalized posttreatment status, and 8.4% for final benchmark) (b) median split on word fluency slope differentiated adequately and inadequately responsive groups whereas the median split on nonsense fluency slope did not, (c) classification accuracy of the final normalized method was greater than the final benchmark method.

Collectively, these studies show how different definitions of adequate and inadequate response to instruction elicit different incidence rates of RD and identify different groups of students with varying degrees of reading difficulty. Final normalization methods resulted in acceptable incidence rates of reading disabilities but elicited mixed hit rates, sensitivity and specificity (Fuchs & Deshler, 2007). Benchmark and median split methods generally overidentified reading disabilities. Slope and dual discrepancy also tended to overidentify reading disabilities but elicited acceptable hit rates, sensitivity, and specificity statistics (Fuchs & Deshler, 2007). Needed is additional research that explores which RTI approaches appear viable. Additional research must examine the extent to which alternate RTI approaches differentiate response-nonresponse to instruction.

1.7. Research questions

This study systematically examined the extent to which different methods, cut-points, and measures overlap and agree in the identification of student responder status in order to help identify the strengths and weaknesses of different operationalizations of RTI. Although it is likely that responsiveness is dimensional and represents a continuous attribute (Denton, Fletcher, & Anthony, 2006; Fletcher et al., 2007), at some point a decision must be made that subdivides responsiveness into different hypothesized classes, such as adequate and inadequate responders. Based on this two-class model, we asked to what extent do different operationalizations of RTI overlap and identify the same students as adequate and inadequate responders, controlling for cut-point, method, and measure. For psychometric reasons, we hypothesized that cut-point would be a major determinant of agreement because of its impact on the observed base rate of responder subgroups.

2. Method

2.1. Participants

To further examiner the extent to which different operationalizations of RTI overlap, we retrospectively analyzed the data from a reading intervention study involving explicit, intense reading interventions (see Mathes et al., 2005). This dataset was selected because students at risk for later reading failure were provided intensive intervention and response to intervention was measured before, during, and after intervention. Both of these features are key components of RTI approaches and make this dataset suitable for reanalysis.

2.1.1. Schools—This research was conducted in six schools in a large urban school district in Texas that participated in a multi-tiered Grade 1 intervention study. We selected these schools for the original intervention studies (Denton et al., 2006; Mathes et al., 2005) because they had been designated as adequately performing schools in reading by the state’s department of education. This designation suggests that classroom (Tier 1) reading was adequate. None of these schools was Title 1-eligible and all served diverse student populations in terms of ethnicity and socio-economic status.

2.1.2. Students—During each of two consecutive years, Mathes et al. (2005) identified within these schools a sample of first-graders who showed significant risk for reading difficulties. In order to determine which students were at-risk for reading difficulty, classroom teachers and the research team screened all students at the end of kindergarten and beginning of Grade 1 using the screening portions of the Texas Primary Reading Inventory (TPRI; Foorman, Fletcher, & Francis, 2004), the Woodcock-Johnson III (W-J III; Woodcock, McGrew, & Mather, 2001) Letter-Word Identification subtest, the Observation Survey of Early Literacy Achievement text reading subtest (Clay, 2002), and a 1-min oral reading fluency sample. Students identified as at-risk were designated as “not developed” on the TPRI or unable to read (a) five or more words correctly on the WJ-III, (b) texts designated as Level D or higher (Fountas & Pinnell, 1999) with at least 90% accuracy, or (c) five or fewer words correctly per minute on the 1-min oral reading fluency sample.

All students who received their reading instruction in regular education classes were eligible for the study, including students who qualified for special education based on the identification of a learning disability, speech or language impairment, or “other health impairment.” The researchers excluded students with limited English proficiency that were served in bilingual classrooms and students served primarily in self-contained special education classes, which represented two classrooms across the six schools.

Once identified, all students designated as at-risk *within* a school were randomly assigned to one of three conditions: Tier 1 only (Enhanced Classroom Instruction) or conditions involving both Tier 1 and Tier 2, the latter represented by two approaches to small group instruction labeled Proactive and Responsive (Mathes et al., 2005). These small groups received 40 min of daily instruction in groups of three students with a certified teacher for 30 weeks. The difference in the interventions is not relevant for this study; they essentially represent a comparison of a direct instruction approach (i.e., Proactive) and an explicit approach in which lessons were planned by teachers based on ongoing student assessments (i.e., Responsive) (Mathes et al., 2005). In addition, a sample of Typically Achieving readers was randomly selected from among all students in the same classrooms who evidenced no risk for reading problems. The purpose of this socio-demographically comparable Typically Achieving group was to provide a benchmark of reading development in these classrooms.

To increase sample size, the study was conducted over two successive school years with two cohorts of students. The initial sample (n = 399) included 92 students in the Proactive intervention group, 92 students in the Responsive intervention group, 114 students in the at-risk enhanced classroom condition, and 101 students who were typically achieving. After the effects of attrition, 78 Proactive Reading students, 83 Responsive Reading students, 91 at-risk students who received quality classroom instruction with no researcher-provided supplemental intervention (a small number of these students received some supplement intervention provided by their schools), and 94 typically achieving students were assessed at post-test (n = 346). Attrition was not selective. Comparisons of students who left the study and those who remained indicated that skill strengths and weaknesses were not significantly for the two groups. Also, for the analyses conducted, intervention groups (i.e., Proactive and Responsive) were combined in order to increase variance.

Table 1 summarizes the demographic information and educational status information for all students who began the intervention. No statistically significant differences among the at-risk groups were detected for any of the demographic or educational status variables.

2.2. Measures

2.2.1. Rationale—The measures included in the present study represented a subset of measures from a larger assessment battery administered in the Mathes et al. (2005) study.

Because proficiency in reading requires, at a minimum, that children are able to read words and text accurately and fluently, and understand the meaning of text, measures assessing each of these three domains were selected. Assessments of fluency were done throughout the year to assess the impact of the interventions on growth in word reading and fluency. End of year assessments of word reading, fluency, and comprehension were conducted using norm-referenced tests.

2.2.2. Growth assessments—Word reading fluency was assessed four times during the year at 2-month intervals beginning in October using the Sight Word Efficiency and Phonemic Decoding Efficiency subtests from the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). For these subtests, students read as many words or decoded as many pseudowords as they could in 45 s per list. Each list of words and non-words was arranged so that items increased in difficulty. We included both words and non-words to ensure that we measured both phonological decoding ability and sight recognition of familiar or partially familiar words. Internal consistency exceeds .95 for both subtests.

In additional to these bi-monthly measures, passage reading fluency was measured as words read correctly per minute (WCPM) on timed 1 min oral reading samples of end-of-first-grade level passages that had been developed for Continuous Monitoring of Early Reading Skills software (CMERS; Mathes, Torgesen, & Herron, in press). The passages used to evaluate oral reading fluency were subjected to substantial field-testing to determine equivalence of difficulty. These measures were given every three weeks by trained research assistants for a total of 11-13 assessments over each school year.

2.2.3. End-of-year assessments—Measures that were administered only at the end of the school year (i.e., post-test only) included the WJ-III Word Attack, Letter-Word Identification, and Passage Comprehension subtests. Reliability ranges from .87-.97. The Word Attack subtest is a measure of accurate decoding of non-words, whereas Letter-Word Identification is a measure of the ability to read sight words in lists. Passage Comprehension is measured through a cloze procedure, where students read a sentence or brief passage in which certain words have been taken out and students are required to produce the missing words or acceptable substitutions for them. TOWRE and CMERS measures were also administered at the end of the year to determine final status.

2.3. Analytic approach

2.3.1. Estimation of growth parameters—To utilize the TOWRE and CMERS for growth-related assessments of RTI, we used SAS PROC MIXED (Singer, 1998). Individual growth parameters were estimated for the TOWRE Composite and CMERS words correct per minute (WCPM). Growth trajectories in word reading efficiency and non-word reading efficiency were estimated from four occasions of measurement. In contrast, growth trajectories in CMERS were estimated from 11 occasions of measurement for Year 1 participants and 13 occasions for Year 2 participants.

A two-level model was specified for each reading-related skill. Level 1 modeled the repeated measures nature of the data (i.e., within-in student variability due to Time) and Level 2 modeled between-student variability in growth trajectories. Time was centered at the final test administration, which was near the end of first grade. Centering at the end of first grade allowed direct estimation of end-of-year performance levels following Tier 1 instruction or Tier1 instruction plus Tier 2 intervention. Additionally, the intercept terms could be directly compared to the standardized achievement scores also obtain at the end of first grade. Individual growth parameters were estimated using linear growth models with random intercepts and random slopes. Linear modeling was selected because it provided an adequate approximation of more complex growth processes (Raudenbush & Bryk, 2002) and because it closely aligns with the practices employed by schools. Random intercept and random slope terms were significant at the .05 alpha level for all measures analyzed.

2.3.2. Cut-points—The slope and intercept estimates for each student for each measure were saved. For each measure, the mean and standard deviation for the intercept and slope parameters of the Typically Achieving students was calculated to serve as the reference group for determining RTI. Three cut-points were applied to both growth parameters for all students in the study: 0.5, 1.0 and 1.5 standard deviations below the mean of the Typically Achieving students. Thus, this sample included students in the Typically Achieving group as the reference group, so all decisions are relative to that sample. To illustrate, students with slope parameters above the criterion of 0.5 standard deviations below the mean of the Typical Achievers would be classified as a Responder; otherwise, Inadequate Responders.

For the end of year tests, we used the national norm referenced samples and cut-points of at the 30th percentile for the norm referenced tests based on Torgesen (2000) and 40 WCPM for the CMERS, which represents the 35th percentile for WCPM based on the score distribution on the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good, Wallin, Simmons, Kame’enui, & Kaminski, 2002). These represent standard employed benchmarks for these types of measures. We did not make these decisions relative to the entire sample because the means for the Typical achievers varied above the average range and in some instances (e.g., word recognition) exceeded the 80th percentile on the norm referenced tests (Mathes et al., 2005). This decision will likely lead to higher identification rates at 0.5 SD for the growth measures, but is consistent with how these kinds of measures are used for decision making.

3. Results

Table 2 includes the different combinations of methods and cut-points that were evaluated for each of the growth and end of year measures and the proportion of students identified as inadequate responders for each operationalization of responder status. Altogether, we compared 808 different combinations based on 543 association tables. As Table 2 shows, these different approaches identify different proportions of students as inadequate responders. A major determinant of the incidence of inadequate responders is the cut-point, with less stringent cut-points (e.g., 0.5 SD) generating more inadequate responders than more stringent cut-points (e.g., 1.5 SD). However, across cut-points, intercept methods from growth assessments identify more inadequate responders than methods that incorporate slope (either slope alone or dual discrepancy) or end of year assessments. These estimates of inadequate responders likely reflect the high performance level of the Typically Achieving group. For end of year measures, fluency assessments tend to generate more inadequate responders, with a benchmark assessment generating the highest incidence.

3.1. 2×2 measures of association

All possible combinations of cut-point, method, and measure were computed, including separate assessments of cut-point (Tables 3-5), method (Table 6), and measure (Table 7). Each table includes the overall agreement between the two approaches for identifying adequate and inadequate responders, as well as the proportion of students identified as inadequate responders for each approach and the agreement on this decision. It is important to recognize that the denominator for inadequate and adequate responders is based on the total of all agreements and disagreements between methods, so does not sum to the overall agreement rate.

Each table includes the kappa statistic, which is an estimate of the overall concordance between approaches in identifying students as adequate or inadequate responders after controlling for chance occurrence. Kappa (Cohen, 1960) measures interjudge agreements and is often used to examine the reliability of ratings or agreement of the identification approaches (Kraemer, 1979). It is likely to be between 0.4 and 0.8, and will likely have a lower magnitude for populations with very high or very low proportions of any subgroup (Kozan, 1979). Nonetheless, RTI approaches that yield higher kappa values and concordance rates may be superior to approaches in which these statistics are lower depending on the observed base rates for different decisions. It is unrealistic to assume that an approach could yield kappa and concordance rates that equal 1.0. This would occur only if the RTI approaches perfectly agreed in identifying adequate and inadequate responders. Perfect classification is not attainable because of measurement error and the influence of factors such as the base rate of adequate responders in a population, school, etc.

Kappa was selected over the overall level of agreement because the latter combines the agreement of adequate and inadequate responders and usually inflates estimates of classification agreement since the base rate of adequate responders in the population is (hopefully) much larger than the number of inadequate responders. Thus, all 2×2 tables are organized and ranked by kappa value, with only those comparisons yielding kappas of at least .40 reported in the tables. Complete results can be found at www.texasldcenter.org/RTITABLES.

3.1.1. Cut-point—As a means of controlling the effect of cut-point, all possible combinations of methods and measures were examined within each of the three cut-points in order to identify which combination of method and measure elicited the greatest overlap within each cut-point. Table 3 presents results for each cut-point for the growth measures. Of a possible 186 combinations, only 8 comparisons yielded a kappa value of at least 0.40. Although the differences in kappa values and overall agreement are not large across the 8 comparisons, differences in agreement for adequate and inadequate responders vary considerably, reflecting likely overidentification of inadequate responders at 0.5 SD and the low proportions of inadequate responders at 1.0 and 1.5 SD. At 1.0 and 1.5 SD, the agreement for adequate responders is higher, but agreement for inadequate responders is generally poor. Thus, when the observed base rate of adequate responders increases, overall agreement increases, but kappa remains low because of the error rate for inadequate responders.

Next consider the overlap among the end of year measures (i.e., norm-referenced tests and benchmark assessments) for each of the three cut-points (0.5, 1.0, and 1.5 standard deviations below the mean relative to the norm referenced sample or 40 WCPM for the CMERS) (see Table 4). A total of 85 association tables were computed, with 16 yielding kappa values of .40 or higher. Kappa values remain low at 0.5 SD despite high overall concordance rates driven by agreement on the proportion of adequate responders. Even at -1.5 SD, agreement rates are very high (.96-.98), but these results reflect agreement rates for adequate responders (.96-.98). Although a small number of inadequate responders is identified (.02-.05 of the sample), the agreement between measures is low (.31-.45). Interestingly, Table 4 shows a clear tendency for different kinds of measures to yield better kappa values.

Finally, consider the overlap among the growth measures (i.e., TOWRE and CMERS) and end of year measures for the cut-points of 0.5, 1.0, and 1.5 standard deviations below the mean. A total of 196 associations were computed and 31 yielded kappas of 0.40 or higher.Table 5 shows higher kappa values and concordance rates for combinations of end of year and growth measures. However, with exceptions, this pattern is driven by methods derived from the same test. For example, the highest kappa (.88) was for the combination of dual discrepancy and end of year benchmark from the CMERS, where the two methods agreed on .86 of the inadequate responders and .92 of the adequate responders. Note that these methods identify about a third of the sample as inadequate responders. For cross-measure combinations, no association yields an agreement rate for inadequate responders greater than .43.

3.1.2. Method of RTI determination—To evaluate the second criterion, method of determining RTI, all possible combinations of measure and cut-point were examined for each of the three methods (i.e., dual discrepancy, intercept, and slope) in order to identify which combination of measure and cut-point elicited the greatest overlap for each method. Table 6 is based on a total of 69 combinations, with 10 yielding kappa values of at least 0.40 when method was controlled. In Table 6, it is surprising that the highest kappa values were not elicited when common methods (e.g., slope and slope, dual discrepancy and dual discrepancy, or intercept and intercept) were compared. Thus, overlap was not driven by method. Rather, overlap was driven by cut-point, with similar cut-points yielding the greatest kappa values. Even this agreement was poor. The overall concordance ranged from 0.69-0.88. Combinations that agreed upon inadequate responders at .50 and above reflected fluency assessments (TOWRE, CMERS) and achieved this agreement with a tendency towards poorer agreement for identification of adequate responders.

3.1.3. Measure—To evaluate the final criterion, measure, all possible combinations of methods and cut-points were examined in relation to each end of year measure in order to identify which measures yielded the highest overlap. For these assessments, the benchmarks were set at the 30th percentile for the norm referenced tests and at 40 WCPM for CMERS. A total of 272 combinations were estimated, with 71 yielding kappas of at least 0.40. As with Table 5, Table 7 shows that similar measures produce the highest degree of overlap. For example, when two different operationalizations using the CMERS passage fluency measure were compared, overlap was high. Also, operationalizations comparing measures within similar constructs result in high overlap. In other words, measures assessing a similar construct, such as the WJ-III Word Identification and WJ-III Word Attack tests, elicited high overlap. Interestingly, cut-point also played a significant role in the determination of overlap. The lowest kappas were produced by operationalizations comparing disparate cut-points (i.e., cut-point of -0.5 compared to a cut-point of -1.5). Further, the highest kappas were obtained when similar cut-points were used. Thus, although similar measures do overlap, it can also be argued that this is also driven in large part by the cut-point employed.

3.1.4. Influence of reference group—These assessments are relative to the reference group of typical achievers. The results imply that the different approaches tended to identify different students as inadequate responders. To assess this possibility, Table 8 displays the proportion of students meeting criteria for inadequate response across all end of year tests using norm referenced criteria in which inadequate response was defined as a standard score score <93 on the WJ III and TOWRE, and a benchmark of 40 WCPM on CMERS, representing commonly used benchmarks for assessing inadequate response. As Table 8 shows, the proportion of students meeting criteria for inadequate response on all five end of year tests was 0.043. In contrast, the proportion of students meeting criteria for adequate response for all five tests was only 0.50. Interestingly, a small proportion of students (0.05) appear to have begun to apply phonic and structural analysis skills to the pronunciation of unfamiliar printed words as demonstrated by adequate performance on the WJ-III Word Attack subtest. However, this subgroup appears to experience difficulties decoding real words in isolation and in connected text, as demonstrated by inadequate performance on WJ-III Letter-Word Identification, WJ-III Passage Comprehension, TOWRE, and CMERS Passage Fluency. In addition, a small proportion of students (0.04) have adequate decoding and fluency skills yet experience difficulties comprehending connected text. Further, a large proportion of students (0.14) appear to present reading fluency deficits; however, this deficit does not significantly impact their reading comprehension abilities. The latter proportion may reflect a benchmark that is set to low to determine inadequate responders, especially in comparison to a norm-referenced benchmark for the TOWRE (standard score score <93), which uniquely identified only 0.03 of the sample as inadequate responders.

4. Discussion

Given the sheer number of potential RTI operationalizations, we sought to evaluate the extent to which different operationalizations of RTI agree in dichotomizing students into adequate and inadequate responder groups. Across the 808 combinations computed for this study, most did not yield kappa values of a minimum level of agreement (0.40), with the 136 kappas>0.40 representing slightly over 15% of the combinations. The kappa values that reached this threshold rarely showed significant agreement for identifying inadequate responders, and high overall concordance was driven largely by agreement on adequate responders. Agreement for inadequate responders was seen only when the same measures were used to generate the criteria for identification, which is hardly surprising. The superiority of growth over end of year benchmarks, or of dual discrepancy approaches over other assessments of growth, was not apparent. For example, assessments of dual discrepancy, slope, or intercept that used two different fluency measures (CMERS, TOWRE) did not generate high levels of agreement for identifying inadequate responders. This low concordance may reflect the use of 4 (TOWRE) vs. 11-13 (CMERS) time points, but is still surprisingly low given the correlation of the TOWRE and CMERS scores at the final time point and within measure correlations over time.

The results of this study clearly show that cut-point is the most significant determinant of responder status. Different cut-points derive different incidences of the proportion of adequate and inadequate responders. The variation in these errors is influenced by the observed base rate as well as errors in the diagonals of the 2×2 decision tables. These diagonals, which in traditional decision tables reflect the sensitivity and specificity of a decision when the true out come is known, influence the magnitude of kappa (Kraemer, 1979). In addition, cut-point is important because response status dichotomizes a continuous distribution of scores. It is well-known that placing cut-points on continuous dimensions will lead to instability in classification. Instability occurs because scores inevitably fluctuate around the established criterion due to the measurement error inherent to the test (Francis, Fletcher, & Stuebing, 2005; Shepard, 1980). Even if a test is appropriate, the cut-point is located at the score distribution’s maximal point of precision, and the assessment is administered repeatedly to improve the reliability of the estimate of ability, instability around the cut-point will still occur. In research on RTI assessment, cut-point must be controlled or differences in the agreement, sensitivity, and specificity of a particular approach will be masked by differences in the observed base rates for different criteria.

Because cut-points ultimately divide a continuous distribution in two, students performing below the cut-point enter secondary or tertiary interventions while students above the cut-point do not. A stringent and literal implementation of cut-points implies that there is a real difference in the instructional needs of children who score just above or below this arbitrary cut-point. But psychometric evaluations of cut-points demonstrate that students just above or below the arbitrary cut-point frequently present the same strengths and weaknesses (Francis et al., 2005) and thus present similar instructional needs. For example, students whose reading skills are slightly below expectation (but above a cut-point) would still benefit from more time in explicit instruction than a student at grade expectation. Therefore, placement into a continuum of RTI services could be guided by the use of cut-points and then confirmed by expert teacher judgment. This is supported by emerging research that suggests that instructional models that operate on a continuum and incorporate expert teacher judgment are more effective than instructional models that adhere strictly to an arbitrary cut-point (Connor, Morrison, Fishman, Schatschneider, & Underwood, 2007).

This naturally highlights the difficult decisions schools ultimately face in selecting and implementing RTI approaches for identification. A school’s answer to the question likely rests with the number of classification errors they can *afford* to make. There is a trade off in terms of error rates and resources. Errors in which a child who is struggling is misidentified as an adequate responder (false negative) are likely more serious than errors in which a true adequate responder is identified as needing intervention (false positive). Here a major issue would be resources available for providing intervention. Schools must weigh how many false positives they can afford to advance within the multiple intervention tiers in hopes of preventing later academic difficulties. However, in many instances, schools lack the resources to serve those students who might not really need intervention and can only focus on those students currently presenting significant reading difficulties. Different issues may pertain in terms of deciding whether RTI results indicate a disability, where the risk of not identifying a child with a disability (and affording them the due process rights inherent in special education identification) must be weighed against the risk of mislabeling a child. Thus, although an RTI approach can provide guidance as to who might benefit from additional instruction, resources and the consequences of labeling must be weighed in the decision-making process in RTI and special education identification.

These issues highlight the importance of using multiple criteria for determining responder status and major decisions like special education status. Although the use of confidence intervals would help deal with instability, test performance cannot be the sole determinant of special education status. It would be tragic if the determination of responder status became formulaic and was used in schools in the same way as approaches based on ability-achievement discrepancy. The psychometric issues are similar and the approaches used in this paper are all examples of alternative approaches to the estimation of a discrepancy, albeit relative to a benchmark or age/normative expectation.

The consensus summary from the Learning Disabilities Summit convened by the Office of Special Education Program (Bradley et al., 2002) recommended that the determination of LD be based on three criteria: response to intervention, assessments of achievement, and the application of traditional exclusionary criteria that should not be the primary cause of low achievement if the student’s difficulties reflect LD. In some respects, this model was incorporated into IDEA through requirements for assessing intervention integrity, a comprehensive evaluation, and traditional exclusionary criteria. It seems important to differentiate the determination of instructional response, low achievement, and the application of exclusions as separate parts of the determination of a learning disability, particularly for eligibility decisions. It is obvious that students may have inadequate RTI that could be due to factors other than LD, such as limited English proficiency. The low agreement identified in this paper based on growth assessments is a problem when identification is based solely on progress monitoring assessments. In the present study, low agreement may be partly a function of the success of the interventions in the study and the fact that relatively few children should be considered inadequate responders. Future studies should examine agreement using different studies and measures.

Table 8 is especially disturbing since 50% of the sample could be identified as inadequate responders if multiple assessments were conducted. To a certain extent, this is not surprising since students could manifest difficulties in one or more domains of reading. It is also because of the measurement error of the different tests, some of which measure similar reading constructs. Other factors involve the setting of benchmarks. Clearly the criterion benchmark for the end of year CMERS is set too low and likely represents differences in the difficulty level of the CMERS and DIBELS passages from which the benchmark was derived. Perhaps setting the norm-referenced benchmarks at 0.5 SD below the mean was also too liberal. Changing this benchmark to, for example, 1.0 SD (standard scores scores <86) would decrease false positives, but increase the false negative rate. Again, where to set the cut-point is driven by decisions about the tolerance for different kinds of errors and by resources.

Another approach not formally evaluated in this paper would be to identify inadequate responders based on a smaller number of dimensions and focusing on end of the year assessments. For example, if the benchmark assessment on the CMERS was not used and the identification focused on the major constructs of interests (word recognition accuracy, fluency, and comprehension), a norm referenced assessment could be completed with three measures: a composite of Word Identification and Word Attack, the TOWRE composite, and Passage Comprehension from the Woodcock or an alternative assessment of reading comprehension to control for higher levels of agreement due to method variance. Using the present sample, such a method would have identified 10% of the sample as inadequate responders, which may not be unreasonable since multiple categories are being used. Thus, in accordance with the LD Summit consensus (Bradley et al., 2002), it may be best to use growth measures to determine passage through successive tiers as one of the potential criteria and to rely upon the use of the highly reliable norm-referenced assessments in a multiple category approach. It is reasonable to think that children could demonstrate inadequate response at the end of a Grade 1 intervention because of problems in any of these three major domains of reading proficiency. Even here, cut-point is a critical issue. The cut-point that was established would be designed to minimize errors in identifying children as adequate responders and ensure that the children who needed assistance would receive it. A standard score that was set lower than the 30th percentile on these measures would likely yield higher levels of agreement, but potentially increase the risk of false negative errors.

Many of the issues raised by this paper could be better resolved by adopting some type of “gold standard” against which different operationalizations of RTI could be compared. Then alternative methods for predicting the gold standard, such as receiver operator curves (Burns & VanDerheyden, 2006; Compton, Fuchs, Fuchs, & Bryant, 2006) or logistic regression could be used to help determine optimal cut-points that are placed along the continuous distribution of “responsiveness.” In addition, latent class models that examine multiple indicators of responder status may be useful in establishing whether some type of gold standard is viable, which would facilitate decision-making about responder status.

Another issue that has never been evaluated is the reliability of decision-making by experts who have multiple sources of information. In many respects, the development of kappa statistics and concerns about agreement/overlap stem from classification work involving categorical psychiatric classifications (Cicchetti,1981). Given the apparent weakness of statistical decision-making for continuous attributes like LD and the trade-off in false negative versus false positive decisions, it is possible that decision-making may be more reliable if it is made using multiple criteria by a group of experts. However, this determination is also an empirical question and should be carefully evaluated using the sorts of analyses outlined in this study. Examining the reliability of expert decision-making does mimic the type of decision-making that might be done in a school in which an interdisciplinary team is convened to consider the issue of special education eligibility. The data in this paper do not address the reliability or level of agreement of these approaches to decision-making, but the same types of evaluations should be conducted to assess their reliability and validity.

Altogether, the results of this study indicate that choice of cut-point, method, and measure does influence who is classified as adequate and inadequate responders and should move through the multiple tiers of intervention. However, these findings are limited because the sample likely contained a small number of children who did not respond favorably to the interventions provided and the retrospective nature of the analyses that were conducted. Findings require corroboration with either multiple diagnostic efficacy studies or a longitudinal study that examines the nature of RTI in a large, heterogeneous, and representative sample of children. These studies are necessary because many of the indices used to quantify classification accuracy (i.e., sensitivity and specificity) are sensitive to base rate fluctuations. Because this sample may not generalize to the base rate of adequate or inadequate responder in other educational settings, the findings must be cautiously treated. Future research would also be strengthened by the use of statistical models that focus more broadly on latent class issues and better use of growth data (e.g., Compton et al., 2006).

In the end, the goal is reliable ascertainment of students who are consistently non-responsive and demonstrate intractability in their instructional response. Such students are important to isolate since they may epitomize the “unexpected underachievement” construct that is the heart of the concept of LD. However, there continues to be large gaps in our knowledge of how to isolate students who consistently do not respond to instruction and may be LD (Fletcher et al., 2007). Thus, one of the critical questions remaining to be answered is which single measure or combination of measures most accurately identifies students who will experience serious and chronic reading difficulties that will prevent reading for understanding and will limit their ability to function successfully as adults in today’s technologically advanced society (Fuchs et al., 2004).

Acknowledgments

Grants from the Interagency Educational Research Initiative (National Science Foundation, National Institute of Child Health and Human Development, and Institute of Educational Sciences), NSF REC 9979968 and the National Institute of Child Health and Human Development, P50 HD052117, Texas Center for Learning Disabilities, supported this paper.

References

Bradley, R.; Danielson, L.; Hallahan, D., editors. Identification of learning disabilities: Research to practice. Erlbaum; Mahwah NJ: 2002. www.air.org/ldsummit

Burns MK, Senesac BV. Comparison of dual discrepancy criteria to assess response to intervention. Journal of School Psychology 2005;43:393–406.

Burns MK, VanDerheyden AM. Using response to intervention to assess learning disabilities: Introduction to the special series. Assessment for Effective Intervention 2006;32:3–5.

Case LP, Speece DL, Molloy DE. The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual differences and contextual factors. School Psychology Review 2003;32:557–582.

Cicchetti D. Testing the normal approximation and minimal sample size requirements of weighted kappa when the number of categories is large. Applied Psychological Measurement 1981;5:101–104.

Clay, MM. An observation survey of early literacy achievement. 2nd ed.. Heinemann; Portsmouth, NH: 2002.

Cohen J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement 1960;10:37–46.

Compton DL, Fuchs D, Fuchs LS, Bryant J. Selecting at-risk readers in first grade for early intervention: A two year longitudinal study of decision rules and procedures. Journal of Educational Psychology 2006;98:394–409.

Connor CM, Morrison FJ, Fishman BJ, Schatschneider C, Underwood P. The early years: Algorithm-guided individualized reading instruction. Science 2007;313:464–465. [PubMed: 17255498]

Denton C, Fletcher JM, Anthony J. An evaluation of intensive intervention for students with persistent reading difficulties. Journal of Learning Disabilities 2006;39:447–466. [PubMed: 17004676]

Donovan, MS.; Cross, CT. Minority students in special and gifted education. National Academy press; Washington, DC: 2002. http://www.nap.edu/catalog/10128.html

Fletcher, JM.; Lyon, R.; Fuchs, LS.; Barnes, MA. Learning disabilities: From identification to intervention. The Guilford Press; New York, NY: 2007.

Foorman, BF.; Fletcher, JM.; Francis, D. Texas Primary Reading Inventory. Texas Educational Agency and the University of Texas System; 2004.

Fountas, IC.; Pinnell, GS. Matching books to readers. Heinemann; Portsmouth, NH: 1999.

Francis DJ, Fletcher JM, Stuebing K. Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. Journal of Learning Disabilities 2005;38:98–108. [PubMed: 15813593]

Fuchs LS. Assessing intervention responsiveness: Conceptual and technical issues. Learning Disabilities Research & Practice 2003;18:172–186.

Fuchs D, Deshler DD. What we need to know about responsiveness to intervention (and shouldn't be afraid to ask). Learning Disabilities Research & Practice 2007;22:129–136.

Fuchs LS, Fuchs D. A unifying concept for reconceptualizing the identification of learning disabilities. Learning Disabilities Research & Practice 1998;13:204–219.

Fuchs D, Compton DL, Fuchs LS, Bryant J, Davis N. Making "secondary intervention" work in a three-tier responsiveness-to-intervention model: Finding from the first-grade longitudinal reading study of the National Research Center on Learning Disabilities. Reading and Writing: An Interdisciplinary Journal 2008;21:413–436.

Fuchs D, Fuchs LS, Compton DL. Identifying reading disabilities by responsiveness-to-instruction: specifying measures and criteria. Learning Disability Quarterly 2004;27:216–227.

Fuchs LS, Fuchs D, Speece DL. Treatment validity as a unifying construct for identifying learning disabilities. Learning Disability Quarterly 2002;25:33–45.

Fuchs D, Mock D, Morgan PL, Young CL. Responsiveness-to-intervention for the learning disabilities construct. Learning Disabilities Research & Practice 2003;18:157–171.

Good, RH.; Wallin, J.; Simmons, DC.; Kame'enui, EJ.; Kaminski, RA. System-wide Percentile Ranks for DIBLES Benchmark Assessment (Technical Report 9). University of Oregon; Eugene, OR: 2002.

Kraemer HC. Ramifications of a population model for k as a coefficient of reliability. Psychometrika 1979;44:461–472.

Kozan LM. The reliability of clinical methods, data, and judgments. New England Journal of Medicine 1979;293:642–646.

Marsten, DB. A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In: Shinn, MR., editor. Curriculum-based measurement: Assessing special children. The Guilford Press; New York: 1989. p. 18-77.

Mathes PG, Denton CA, Fletcher JM, Anthony JA, Francis DJ, Schatschneider C. The effects of theoretically different instruction and student characteristics on the skills of struggling readers. Reading Research Quarterly 2005;40:148–182.

Mathes, PG.; Torgesen, JK.; Herron, J. Continuous monitoring of early reading skills (CMERS) [Computer software]. ProEd; Austin, Texas: in press

McMaster KL, Fuchs D, Fuchs LS, Compton DL. Responding to nonresponders: An experimental field trial of identification and intervention methods. Exceptional Children 2005;71:445–463.

National Joint Committee on Learning Disabilities. Responsiveness to intervention and learning disabilities: A report prepared by the National Joint Committee on Learning Disabilities representing eleven national and international organizations. Author; Washington, DC: 2005.

Raudenbush, SW.; Bryk, AS. Hierarchical linear models: Applications and data analysis method. 2nd ed.. Sage Publications, Inc.; Thousand Oaks, CA: 2002.

Shepard L. An evaluation of the regression discrepancy method for identifying children with learning disabilities. Journal of Special Education 1980;14:79–91.

Singer JD. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. Journal of Educational and Behavioral Statistics 1998;323:23–355.

Speece DL, Case LP. Classification in context: an alternative approach to identifying early reading disability. Journal of Educational Psychology 2001;93:735–749.

Swets JA. The science of choosing the right decision threshold in high-stakes diagnostics. American Psychologist 1992;47:522–532. [PubMed: 1595983]

Torgesen JK. Individual differences in response to early interventions in reading: The lingering problem of treatment resisters. Learning Disabilities Research and Practice 2000;15:55–64.

Torgesen JK, Alexander A, Wagner R, Rashotte C, Voeller K, Conway T. Intensive, remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. Journal of Learning Disabilities 2001;34:33–58. [PubMed: 15497271]

Torgesen, JK.; Wagner, RK.; Rashotte, CA. The test of word reading efficiency. Pro-Ed; Austin, TX: 1999.

U. S. Department of Education. Individuals with Disabilities Improvement Act of 2004, Pub. L. 108-466. Federal register 2004;70(118):35802–35803.

Vaughn S, Fuchs LS. Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. Learning Disabilities Research & Practice 2003;18:137–146.

Vellutino FR, Scanlon DM, Sipay ER, Small SG, Pratt A, Chen R, et al. Cognitive profiles of difficult-to-remediate and readily remediate 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 1996;88:601–638.

Woodcock, R.; McGrew, K.; Mather, N. Woodcock Johnson-III tests of achievement. Riverside; Itasca, IL: 2001.

Table 1

Sample statistics by comparison group

	Normal controls	At-risk controls	Proactive intervention	Responsive intervention
N	101	114	92	92
Age in months (SD)	79(4.8)	78(4.8)	78(4.9)	78(4.2)
Sex				
Male	62%	60%	57%	58%
Female	38%	40%	43%	42%
Race				
Caucasian	31%	22%	26%	32%
African American	41%	46%	43%	45%
Hispanic American	24%	24%	25%	23%
Asian American/Other	5%	9%	5%	1%
Special education				
Yes	1%	3%	2%	3%
No	96%	80%	86%	88%
Unknown	3%	18%	12%	9%
Speech therapy				
Yes	2%	7%	7%	3%
No	95%	75%	82%	88%
Unknown	3%	18%	12%	9%
Bilingual/ESL services				
Yes	4%	5%	5%	0%
No	92%	77%	83%	91%
Unknown	3%	18%	12%	9%

Table 2

Proportion of inadequate responders identified for each operationalization of response to instruction

Measure	Method	Cut-point		
		-0.5 SD **Proportion of IR**	**-1.0 SD** **Proportion of IR**	**-1.5 SD** **Proportion of IR**
Growth measures				
TOWRE	Dual discrepancy	.25	.10	.05
	Slope	.29	.17	.10
	Intercept	.55	.25	.11
CMERS passage Fluency	Dual discrepancy	.54	.35	.18
	Slope	.54	.36	.20
	Intercept	.70	.44	.20
End of year measure norm referenced tests				
WJ-III passage comprehension	Final status	.27	.12	.05
WJ-III letter-word identification	Final status	.14	.06	.02
WJ-III word attack	Final status	.08	.04	.02
TOWRE composite	Final status	.29	.12	.04
End of year benchmark tests				
CMERS	Benchmark 40 wcpm	.38		

Note. TOWRE = Test of Word Reading Efficiency; WJ-III = Woodcock-Johnson-III; CMERS = Continuous Monitoring of Early Reading Skills; IR = Inadequate Responders.

Table 3

Comparison of the degree of overlap among different RTI operationalizations that employed a cut-point of -0.5, -1.0, and -1.5 standard deviations below the mean of the typically developing control group and allowed method and growth measures to vary

Measure 1	Measure 2	Cut-point	Kappa	Agreement	IR M1	IR M2	Agreement IR	Agreement AR
TOWREINT	CMERBOTH	-0.5 SD	0.43	0.72	0.54	.54	0.59	0.53
TOWREINT	CMERSLOPE	-0.5 SD	0.43	0.72	0.54	.54	0.59	0.53
TOWREINT	CMERINT	-0.5 SD	0.47	0.75	0.54	.70	0.66	0.50
TOWRESLOPE	CMERBOTH	-1.0 SD	0.42	0.76	0.20	.35	0.39	0.72
TOWRESLOPe	CMERSLOPE	-1.0 SD	0.43	0.76	0.20	.36	0.40	0.72
TOWREINT	CMERBOTH	-1.0 SD	0.45	0.76	0.26	.35	0.44	0.71
TOWREINT	CMERSLOPE	-1.0 SD	0.46	0.76	0.26	.36	0.45	0.71
TOWREINT	CMERBOTH	-1.5 SD	0.40	0.85	.11	.18	0.32	0.84

Note. Growth measures included the Test of Word Reading Efficiency (TOWRE) composite and Continuous Monitoring of Early Reading Skills (CMERS). Growth methods included Intercept (INT), Slope (SLOPE), and Dual Discrepancy (BOTH). Growth cut-points included 0.5, 1.0, and 1.5 standard deviations below the mean of typically developing control group. IR M1 represents the proportion of students identified as inadequate responders (IR) for Measure 1 (M1). IR M2 represents the proportion of students identified as inadequate responders (IR) for Measure 2 (M2). Agreement IR represents the proportion of students identified as inadequate responder on Measure 1 and Measure 2. Agreement AR represents the proportion of students identified as adequate responder on Measure 1 and Measure 2.

Table 4

Comparison of the degree of overlap among different RTI operationalizations that employed a cut of -.5, -1.0, and -1.5 standard deviations below the mean of the typically developing control group and allowing method and end of year measures to vary

Measure 1	Measure 2	Cut-point	Kappa	Agreement	IR M1	IR M2	Agreement IR	Agreement AR
W5PC05	W5CMER40	-0.5 SD	0.40	0.73	0.27	0.38	0.42	0.67
W4TOWRECOMP05	W5CMER40	-0.5 SD	0.46	0.75	0.29	0.38	0.47	0.69
W5WA05	W5LWI05	-0.5 SD	0.47	0.89	0.08	0.14	0.36	0.89
W5PC05	W4TOWRECOMP05	-0.5 SD	0.51	0.80	0.27	0.29	0.48	0.76
W5LWI05	W4TOWRECOMP05	-0.5 SD	0.51	0.83	0.14	0.29	0.43	0.81
W5LWI05	W5PC05	-0.5 SD	0.57	0.85	0.14	0.27	0.48	0.83
W4TOWRECOMP10	W5PC10	-1.0 SD	0.52	0.90	0.12	0.12	0.41	0.89
W5LWI10	W4TOWRECOMP10	-1.0 SD	0.56	0.93	0.06	0.12	0.43	0.92
W5LWI10	W5PC10	-1.0 SD	0.58	0.93	0.06	0.12	0.44	0.92
W5WA10	W5LWI10	-1.5 SD	0.59	0.96	0.04	0.06	0.43	0.96
W5WA15	W4TOWRECOMP15	-1.5 SD	0.46	0.97	0.03	0.04	0.31	0.96
W5WA15	W5PC15	-1.5 SD	0.48	0.96	0.02	0.05	0.33	0.96
W5LWI15	W5PC15	-1.5 SD	0.51	0.97	0.02	0.05	0.35	0.97
W5LWI15	W5WA15	-1.5 SD	0.52	0.98	0.02	0.02	0.36	0.98
W5LWI15	W4TOWRECOMP15	-1.5 SD	0.59	0.97	0.02	0.04	0.43	0.97
W4TOWRECOMP15	W5PC15	-1.5 SD	0.60	0.97	0.04	0.05	0.45	0.96

Note. End of year measures included: WJ-III Passage Comprehension (W5PC), WJ-III Word Attack (W5WA), WJ-III Letter-Word Identification (W5LWI), the Test of Sight Word Efficiency (TOWRE) Composite in the follow-up year, and the Continuous Monitoring of Early Reading Skills (CMERS) in the follow-up year. The end of year method used was the final status score. The end of year cut-points included .5, 1.0, and 1.5 standard deviations below the normative mean or performance below the benchmark. IR M1 represents the proportion of students identified as inadequate responders (IR) for Measure 1 (M1). IR M2 represents the proportion of students identified as inadequate responders (IR) for Measure 2 (M2). Agreement IR represents the proportion of students identified as inadequate responder on Measure 1 and Measure 2. Agreement AR represents the proportion of students identified as adequate responder on Measure 1 and Measure 2.

Table 5

Comparisons of the degree of overlap among end of year measures and growth measures when employing a cut of -.5, -1.0, and -1.5 standard deviations below the mean of the typically developing control group, and allowing method to vary

Measure 1	Measure 2	Cut-point	Kappa	Agreement	IR M1	IR M2	Agreement IR	Agreement AR
CMERBOTH05	W4TOWRECOMP05	-0.5 SD	0.41	0.69	0.54	0.29	0.46	0.58
CMERSLOPE05	W4TOWRECOMP05	-0.5 SD	0.41	0.69	0.54	0.29	0.46	0.58
TOWRESLOP05	W5CMER40	-0.5 SD	0.41	0.73	0.33	0.38	0.45	0.66
W5CMER40	CMERINT05	-0.5 SD	0.41	0.68	0.38	0.70	0.54	0.48
TOWREBOTH05	W5CMER40	-0.5 SD	0.41	0.74	0.25	0.38	0.42	0.68
W4TOWRECOMP05	TOWRESLOPE05	-0.5 SD	0.44	0.77	0.28	0.32	0.44	0.71
W4TOWRECOMP05	TOWREINT05	-0.5 SD	0.51	0.75	0.28	0.53	0.52	0.65
TOWREBOTH05	W4TOWRE_COMP05	-0.5 SD	0.59	0.84	0.25	0.28	0.53	0.81
CMERBOTH05	W5CMER40	-0.5 SD	0.68	0.84	0.54	0.38	0.70	0.74
CMERSLOPE05	W5CMER40	-0.5 SD	0.68	0.84	0.54	0.38	0.70	0.74
TOWREINT10	W5CMER40	-1.0 SD	0.43	0.75	0.26	0.38	0.44	0.69
TOWREBOTH10	W5PC10	-1.0 SD	0.44	0.89	0.10	0.12	0.33	0.88
W4TOWRECOMP10	TOWRESLOPE10	-1.0 SD	0.45	0.86	0.12	0.19	0.36	0.84
W5PC10	TOWREINT10	-1.0 SD	0.46	0.83	0.12	0.26	0.38	0.81
W4TOWRECOMP10	TOWREINT10	-1.0 SD	0.56	0.86	0.12	0.25	0.46	0.85
W5LWI10	TOWREBOTH10	-1.0 SD	0.57	0.93	0.06	0.10	0.43	0.93
W5CMER40	WRINT10	-1.0 SD	0.60	0.80	0.38	0.48	0.62	0.71
TOWREBOTH10	W4TOWRECOMP10	-1.0 SD	0.73	0.95	0.10	0.12	0.61	0.94
W5CMER40	CMERINT10	-1.0 SD	0.84	0.92	0.38	0.44	0.83	0.88
CMERSLOPE10	W5CMER40	-1.0 SD	0.87	0.94	0.36	0.38	0.85	0.91
CMERBOTH10	W5CMER40	-1.0 SD	0.88	0.94	0.35	0.38	0.86	0.92
TOWREINT15	W5PC15	-1.5 SD	0.42	0.91	0.11	0.05	0.30	0.91
W5PC15	TOWREINT15	-1.5 SD	0.42	0.91	0.05	0.11	0.30	0.91
W5LWI15	TOWREBOTH15	-1.5 SD	0.44	0.96	0.02	0.05	0.29	0.96
TOWREBOTH15	W5PC15	-1.5 SD	0.49	0.95	0.05	0.05	0.35	0.95
W5PC15	TOWREBOTH15	-1.5 SD	0.49	0.95	0.05	0.05	0.35	0.95
W4TOWRE_COMP15	TOWREINT15	-1.5 SD	0.52	0.93	0.04	0.11	0.38	0.93
CMERBOTH15	W5CMER40	-1.5 SD	0.54	0.80	0.18	0.38	0.48	0.76
CMERINT15	W5CMER40	-1.5 SD	0.57	0.82	0.20	0.38	0.52	0.77
CMERSLOPE15	W5CMER40	-1.5 SD	0.59	0.82	0.20	0.38	0.53	0.78
W4TOWRECOMP15	TOWREBOTH15	-1.5 SD	0.72	0.98	0.04	0.05	0.58	0.98

Note. End of year measures included: WJ-III Passage Comprehension (W5PC), WJ-III Word Attack (W5WA), WJ-III Letter-Word Identification (W5LWI), Test of Sight Word Efficiency (TOWRE) Composite in the follow-up year, and Continuous Monitoring of Early Reading Skills (CMERS) in the follow-up year. The method used was the final status score. The cut-points included .5, 1.0, and 1.5 standard deviations below the mean the normative sample for TOWRE, WJ-III Letter-Word Identification, Word Attack, and Passage Comprehension and a benchmark of 40 WCPM on the CMERS. Growth measures included the Test of Word Reading Efficiency (TOWRE) composite and Continuous Monitoring of Early Reading Skills (CMERS). Growth methods included Intercept (INT), Slope (SLOPE), and Dual Discrepancy (BOTH). Growth cut-points included .5, 1.0, and 1.5 standard deviations below the mean of typically developing control group. IR M1 represents the proportion of students identified as inadequate responders (IR) for Measure 1 (M1). IR M2 represents the proportion of students identified as inadequate responders (IR) for Measure 2 (M2). Agreement IR represents the proportion of students identified as inadequate responder on Measure 1 and Measure 2. Agreement AR represents the proportion of students identified as adequate responder on Measure 1 and Measure 2.

Table 6

Comparing RTI operationalizations employing the different response to instruction methods allowing measure and cut-point to vary

Measure 1	**Measure 2**	**Kappa**	**Agreement**	**IR M1**	**IR M2**	**Agreement IR**	**Agreement AR**
Dual discrepancy method							
TOWREBOTH05	CMERBOTH10	0.46	0.77	.25	.35	0.44	0.71
TOWREBOTH10	CMERBOTH15	0.50	0.88	.10	.18	0.39	0.87
CMERBOTH15	TOWREBOTH05	0.53	0.84	.18	.25	0.46	0.81
Intercept method							
CMERINT15	TOWREINT10	0.41	0.79	.20	.25	0.37	0.76
TOWREINT05	CMERINT05	0.47	0.75	.55	.70	0.66	0.50
Slope method							
CMERSLOPE05	TOWRESLOP05	0.40	0.69	.54	0.33	0.47	0.57
TOWRESLOPE10	CMERSLOP10	0.43	0.76	.20	0.36	0.40	0.72
CMERSLOPE15	TOWRESLOP05	0.47	0.79	.20	0.33	0.43	0.75
TOWRESLOPE05	CMERSLOP10	0.49	0.77	.33	0.36	0.50	0.70
TOWRESLOPE10	CMERSLOP15	0.51	0.85	.20	0.20	0.44	0.82

Note. Growth measures included the Test of Word Reading Efficiency (TOWRE) composite and Continuous Monitoring of Early Reading Skills (CMERS). Growth methods included Intercept (INT), Slope (SLOPE), and Dual Discrepancy (BOTH). Growth cut-points included .5 (05), 1.0 (10), and 1.5 (15) standard deviations below the mean of typically developing control group. IR M1 represents the proportion of students identified as inadequate responders (IR) for Measure 1 (M1). IR M2 represents the proportion of students identified as inadequate responders (IR) for Measure 2 (M2). Agreement IR represents the proportion of students identified as inadequate responder on Measure 1 and Measure 2. Agreement AR represents the proportion of students identified as adequate responder on Measure 1 and Measure 2.

Table 7

Comparing degree of overlap between growth measures and gold standard end of year tests

Measure 1	Measure 2	Kappa	Agreement	IR M1	IR M2	Agreement IR	Agreement AR
End of year CMERS: benchmark of 40 wcpm							
TOWRESLOPE05	W5CMER40	0.41	0.73	0.33	0.38	0.45	0.66
W5CMER40	CMERINT05	0.41	0.68	0.38	0.70	0.54	0.48
TOWREBOTH05	W5CMER40	0.41	0.74	0.25	0.38	0.42	0.68
TOWREINT10	W5CMER40	0.43	0.75	0.26	0.38	0.44	0.69
CMERBOTH15	W5CMER40	0.54	0.80	0.18	0.38	0.48	0.76
CMERINT15	W5CMER40	0.57	0.82	0.20	0.38	0.52	0.77
CMERSLOPE15	W5CMER40	0.59	0.82	0.20	0.38	0.53	0.78
W5CMER40	CMERBOTH05	0.68	0.84	0.38	0.54	0.70	0.74
W5CMER40	CMERSLOPE05	0.68	0.84	0.38	0.54	0.70	0.74
W5CMER40	CMERINT10	0.84	0.92	0.38	0.44	0.83	0.88
CMERSLOPE10	W5CMER40	0.87	0.94	0.36	0.38	0.85	0.91
CMERBOTH10	W5CMER40	0.88	0.94	0.35	0.38	0.86	0.92
End of year WJ-III Letter-word identification							
W5LWI05	CMERSLOPE15	0.43	0.84	0.14	0.20	0.36	0.82
W5LWI15	TOWREBOTH15	0.44	0.96	0.02	0.05	0.29	0.96
W5LWI05	CMERINT15	0.45	0.85	0.14	0.20	0.37	0.83
W5LWI05	CMERBOTH15	0.45	0.85	0.14	0.18	0.37	0.84
TOWREBOTH10	W5LWI05	0.51	0.89	0.10	0.14	0.39	0.89
W5LWI10	TOWREINT15	0.53	0.93	0.06	0.11	0.40	0.92
W5LWI05	TOWREINT10	0.55	0.85	0.14	0.26	0.46	0.83
W5LWI10	TOWREBOTH10	0.57	0.93	0.06	0.10	0.43	0.93
W4TOWRECOMP15	W5LWI10	0.61	0.96	0.04	0.06	0.45	0.96
TOWREBOTH15	W5LWI10	0.61	0.96	0.05	0.06	0.46	0.96
TOWREINT15	W5LWI05	0.70	0.93	0.11	0.14	0.59	0.93
End of year WJ-III passage comprehension							
TOWREBOTH15	W5PC10	0.40	0.91	0.05	0.12	0.29	0.90
W5PC10	CMERSLOPE15	0.41	0.84	0.12	0.20	0.33	0.82
CMERSLOPE15	W5PC05	0.41	0.79	0.20	0.27	0.38	0.75
TOWREINT15	W5PC15	0.42	0.91	0.11	0.05	0.30	0.91
W5PC15	TOWREINT15	0.42	0.91	0.05	0.11	0.30	0.91
W5PC10	CMERINT15	0.42	0.85	0.12	0.20	0.34	0.83
W5PC10	CMERBOTH15	0.43	0.85	0.12	0.18	0.34	0.84
TOWREBOTH10	W5PC10	0.44	0.89	0.10	0.12	0.33	0.88
W5PC10	TOWREINT10	0.46	0.83	0.12	0.26	0.38	0.81
TOWREINT15	W5PC05	0.46	0.83	0.11	0.27	0.38	0.81
TOWREBOTH15	W5PC15	0.49	0.95	0.05	0.05	0.35	0.95
W5PC15	TOWREBOTH15	0.49	0.95	0.05	0.05	0.35	0.95
TOWREINT10	W5PC05	0.55	0.82	0.26	0.27	0.50	0.78
TOWREINT15	W5PC10	0.59	0.92	0.11	0.12	0.47	0.91
End of year WJ-III word attack							
W5WA10	TOWREINT15	0.41	0.92	0.04	0.11	0.29	0.91
W5WA15	W5WA05	0.44	0.94	0.02	0.08	0.30	0.94
W4TOWRECOMP15	W5WA05	0.48	0.94	0.04	0.08	0.34	0.94
W5WA05	TOWREINT15	0.49	0.91	0.08	0.11	0.37	0.91
W5WA10	TOWREBOTH15	0.55	0.96	0.04	0.05	0.40	0.96
End of year TOWRE composite							
W4TOWRECOMP05	CMERBOTH05	0.41	0.69	0.29	0.54	0.46	0.58
W4TOWRECOMP05	CMERSLOP05	0.41	0.69	0.29	0.54	0.46	0.58
TOWRESLOP15	W4TOWRECOMP10	0.43	0.88	0.12	0.12	0.38	0.76
TOWRESLOP10	W4TOWRECOMP05	0.43	0.79	0.19	0.28	0.33	0.88
TOWREBOTH10	W4TOWRECOMP05	0.44	0.82	0.10	0.28	0.35	0.80
W4TOWRECOMP05	TOWRESLOP05	0.44	0.77	0.28	0.32	0.44	0.71
W4TOWRECOMP10	TOWRESLOP10	0.45	0.86	0.12	0.19	0.36	0.84

Measure 1	Measure 2	Kappa	Agreement	IR M1	IR M2	Agreement IR	Agreement AR
W4TOWRECOMP05	W5CMER40	0.46	0.75	0.29	0.38	0.47	0.69
W4TOWRECOMP05	CMERINT10	0.46	0.74	0.29	0.44	0.48	0.66
W4TOWRECOMP10	TOWREBOTH05	0.46	0.84	0.12	0.25	0.37	0.82
TOWREINT15	W4TOWRECOMP05	0.48	0.83	0.11	0.28	0.39	0.81
W4TOWRECOMP10	CMERINT15	0.48	0.86	0.12	0.19	0.39	0.85
W4TOWRECOMP10	CMERSLOP15	0.49	0.86	0.12	0.20	0.39	0.84
CMERINT15	W4TOWRECOMP05	0.49	0.81	0.19	0.29	0.44	0.77
CMERBOTH15	W4TOWRECOMP05	0.49	0.81	0.19	0.29	0.43	0.78
W4TOWRECOMP05	CMERBOTH10	0.50	0.78	0.29	0.36	0.49	0.72
CMERSLOP15	W4TOWRECOMP05	0.50	0.81	0.20	0.29	0.45	0.78
W4TOWRECOMP10	CMERBOTH15	0.51	0.87	0.12	0.19	0.41	0.86
W4TOWRECOMP05	W5CMER35	0.51	0.80	0.29	0.30	0.48	0.75
W4TOWRECOMP05	CMERSLOP10	0.51	0.78	0.29	0.36	0.50	0.72
W4TOWRECOMP05	TOWREINT05	0.51	0.75	0.28	0.53	0.52	0.65
W4TOWRECOMP15	TOWREBOTH10	0.51	0.94	0.04	0.10	0.37	0.93
W4TOWRECOMP15	TOWREINT15	0.52	0.93	0.04	0.11	0.38	0.93
End of year TOWRE composite							
TOWREBOTH15	W4TOWRECOMP10	0.54	0.93	0.05	0.12	0.40	0.93
W4TOWRECOMP10	TOWREINT10	0.56	0.86	0.12	0.25	0.46	0.85
TOWREBOTH05	W4TOWRECOMP05	0.59	0.84	0.25	0.28	0.53	0.81
W4TOWRECOMP15	TOWREBOTH15	0.72	0.98	0.04	0.05	0.58	0.98
TOWREBOTH10	W4TOWRECOMP10	0.73	0.95	0.10	0.12	0.61	0.94
TOWREINT10	W4TOWRECOMP05	0.81	0.93	0.25	0.28	0.76	0.91
TOWREINT15	W4TOWRECOMP10	0.84	0.97	0.11	0.12	0.75	0.96

Note. End of Year measures included: WJ-III Passage Comprehension (W5PC), WJ-III Letter-Word Identification (W5LWI), WJ-III Word Attack (W5WA), TOWRE Composite (W4TOWRECOMP), and Continuous Monitoring of Early Reading Skills (W5CMER). For each end of year measure the method used was the final status score. The cut-points included .5, 1.0, and 1.5 standard deviations below the mean the normative sample for TOWRE, WJ-III Letter-Word Identification, Word Attack, and Passage Comprehension and a benchmark of 40 WCPM on the CMERS . Growth measures included the Test of Word Reading Efficiency (TOWRE) composite and Continuous Monitoring of Early Reading Skills (CMERS). Growth methods included Intercept (INT), Slope (SLOPE), and Dual Discrepancy (BOTH). Growth cut-points included .5, 1.0, and 1.5 standard deviations below the mean of typically developing control group. IR M1 represents the proportion of students identified as inadequate responders (IR) for Measure 1 (M1). IR M2 represents the proportion of students identified as inadequate responders (IR) for Measure 2 (M2). Agreement IR represents the proportion of students identified as inadequate responder on Measure 1 and Measure 2. Agreement AR represents the proportion of students identified as adequate responder on Measure 1 and Measure 2.

Table 8

Proportion of students identified as inadequate responder or adequate responder on each end of year measure

TOWRE composite	WJ-III passage comprehension	WJ-III letter-word identification	WJ-III word attack	CMERS benchmark 40 wcpm	Proportion
Inadequate	Inadequate	Inadequate	Inadequate	Inadequate	0.04
Inadequate	Inadequate	Inadequate	Inadequate	Responder	0.01
Inadequate	Inadequate	Inadequate	Responder	Inadequate	0.06
Inadequate	Inadequate	Inadequate	Responder	Responder	0.02
Inadequate	Inadequate	Responder	Inadequate	Inadequate	0.01
Inadequate	Inadequate	Responder	Inadequate	Responder	0.01
Inadequate	Inadequate	Responder	Responder	Inadequate	0.04
Inadequate	Inadequate	Responder	Responder	Responder	0.01
Inadequate	Responder	Inadequate	Responder	Inadequate	0.01
Inadequate	Responder	Responder	Responder	Inadequate	0.04
Inadequate	Responder	Responder	Responder	Responder	0.03
Responder	Inadequate	Inadequate	Inadequate	Inadequate	0.01
Responder	Inadequate	Inadequate	Responder	Responder	0.003
Responder	Inadequate	Responder	Inadequate	Responder	0.003
Responder	Inadequate	Responder	Responder	Inadequate	0.04
Responder	Inadequate	Responder	Responder	Responder	0.04
Responder	Responder	Responder	Inadequate	Responder	0.01
Responder	Responder	Responder	Responder	Inadequate	0.14
Responder	Responder	Responder	Responder	Responder	0.50

Note. *N*=323 students. Inadequate response was defined as performing below 0.5 standard deviations below the national norm for the TOWRE, WJ-III Passage Comprehension, Letter-Word Identification, and Word Attack subtests and the benchmark of 40 WCPM for the CMERS Passage Fluency subtest.