Published in final edited form as:

Top Lang Disord. 2014 ; 34(1): 74–89. doi:10.1097/TLD.0000000000000004.

AGREEMENT AND COVERAGE OF INDICATORS OF RESPONSE TO INTERVENTION: A MULTI-METHOD COMPARISON AND SIMULATION

Jack M. Fletcher, Ph.D.,

Texas Institute of Measurement, Evaluation, and Statistics and Department of Psychology,
University of Houston

Karla K. Stuebing, Ph.D.,

Texas Institute of Measurement, Evaluation, and Statistics and Department of Psychology,
University of Houston

Amy E. Barth, Ph.D.,

Texas Institute of Measurement, Evaluation, and Statistics and Department of Psychology,
University of Houston

Jeremy Miciak, Ph.D.,

Texas Institute of Measurement, Evaluation, and Statistics and Department of Psychology,
University of Houston

David J. Francis, Ph.D., and

Texas Institute of Measurement, Evaluation, and Statistics and Department of Psychology,
University of Houston

Carolyn A. Denton, Ph.D.

Children's Learning Institute, University of Texas Health Science Center at Houston

Abstract

Purpose—Agreement across methods for identifying students as inadequate responders or as learning disabled is often poor. We report (1) an empirical examination of final status (post-intervention benchmarks) and dual-discrepancy growth methods based on growth during the intervention and final status for assessing response to intervention; and (2) a statistical simulation of psychometric issues that may explain low agreement.

Methods—After a Tier 2 intervention, final status benchmark criteria were used to identify 104 inadequate and 85 adequate responders to intervention, with comparisons of agreement and coverage for these methods and a dual-discrepancy method. Factors affecting agreement were investigated using computer simulation to manipulate reliability, the intercorrelation between measures, cut points, normative samples, and sample size.

Results—Identification of inadequate responders based on individual measures showed that single measures tended not to identify many members of the pool of 104 inadequate responders. Poor to fair levels of agreement for identifying inadequate responders were apparent between pairs of measures. In the simulation, comparisons across two simulated measures generated indices of agreement (kappa) that were generally low because of multiple psychometric issues inherent in any test.

Conclusions—Expecting excellent agreement between two correlated tests with even small amounts of unreliability may not be realistic. Assessing outcomes based on multiple measures, such as level of CBM performance and short norm-referenced assessments of fluency may improve the reliability of diagnostic decisions.

Keywords

decision reliability; diagnostic agreement; kappa; learning disabilities; response to intervention; simulation

INTRODUCTION

Identifying students who may be eligible for special education services because they have failed to respond adequately to educational interventions lies at the heart of all response to intervention (RTI) service delivery models. Many methods have been proposed for making these decisions. However, different methods often lead to different decisions for individual children, partly because different measures are used to operationalize inadequate response. The objective of this paper is to examine the problem of agreement across multiple methods for measuring response to intervention. We report on two studies, one an empirical evaluation of response to a Tier 2 reading intervention and the other a computer-based simulation of psychometric issues that affect agreement in decision making and that may explain why relatively low agreement has been reported in studies of RTI. Although the issue of diagnostic agreement has been addressed in research studies on assessment, many issues arise because multiple measures, methods, and criteria are used to determine the status of a student, and these issues are not adequately understood in practice. This problem is especially important for RTI models because (1) RTI decision making is complex and involves multiple measures, methods, and criteria, and (2) the lack of agreement across different measures, methods, and criteria for assessing intervention response is a common criticism of RTI (Hale et al., 2010; Reynolds & Shaywitz, 2009). In fact, the issue of agreement in decision making is not unique to RTI and is important in any clinical domain in which a student must be identified with a learning disability (LD), speech and language impairment, intellectual disability, or other developmental disorders where tests provide important diagnostic information.

To illustrate, a number of states have adopted identification methods for LD based on a RTI service delivery framework aligned with consensus statements about best practice in identification (Bradley, Danielson, & Hallahan, 2002). These methods often have three essential components. The first is evidence of inadequate achievement defined as performance on a norm referenced achievement test that is well below average (e.g., one standard deviation; 16th percentile) after a series of intensive interventions. The second

criterion is based on insufficient progress in response to intensive evidence-based interventions, usually based on progress monitoring data obtained during interventions. These assessments target rates of growth across the intervention period, with a student considered an inadequate responder if their growth rate is below a defined threshold (e.g., one standard deviation below typically achieving students). The third component involves statutory requirements for consideration of other disabilities and conditions that may explain low achievement and exclude the student from the LD category.

States generally do not stipulate how these criteria should be assessed, i.e., what tests should be used. Different districts (and different professionals) often vary in which tests are used to measure low achievement and intervention response. For example, let's consider two professionals charged with evaluating students for LD. To assess low achievement, the first selects a set of norm-referenced tests measuring achievement domains from the Woodcock Johnson III Tests of Achievement (WJ; Woodcock, McGrew, & Mather, 2001). For progress monitoring, this professional chooses a timed oral reading fluency curriculum-based measure (CBM), the Continuous Monitoring of Emergent Literacy Skills (CMERS; Mathes & Torgesen, 2008). The second decides to use achievement tests that include the same ones from the WJ, but also includes the Test of Word Reading Efficiency (TOWRE; Torgesen, Wagner, & Rashotte, 1999). For progress monitoring, fluency measures from AimsWeb MAZE (Shinn & Shinn, 2002) were selected. Both professionals implemented progress monitoring during intervention and computed slopes to determine if a student's growth trajectory will eventually result in average achievement and use the cut points required at the state level to determine adequate response to intervention. One question that emerges is whether the method used by professional 1 will identify the same students as the method used by professional 2 even though the tests are similar. A second question is whether the test of low achievement and the CBM will agree within each method. Finally, if the methods employed by professional 1 and 2 do not identify the same students, does lack of agreement reflect attributes of the students or of the tests?

These issues are fundamental questions for research, practice, and public policy involving children with disabilities. They are not new questions specific to the RTI service delivery framework. For LD, they were raised for IQ-achievement discrepancy methods when they were first implemented in 1977 because of differences in how aptitude and achievement discrepancies were assessed across districts and states and in the scientific literature on LD. Because of these differences in implementation, LD identification decisions often reflected the methods utilized, rather than attributes that occurred reliably across students. With the adoption of RTI models, the same issues emerge (Fuchs & Deshler, 2007).

RTI METHODS FOR IDENTIFICATION

Methods for identifying inadequate responders largely involve comparisons of assessments of final status (post intervention) to benchmarks and methods that incorporate rate of growth. Final status methods compare post intervention test scores to a criterion that may be norm-referenced or criterion-referenced based on a CBM. Because CBMs used for progress monitoring are repeated several times during the course of the intervention, the amount of

change over time can be computed as a slope (e.g., growth in words read correctly per minute) to quantify growth in the target reading skill that occurred during intervention.

The most widely accepted identification method using a measure of growth is a dual-discrepancy approach that compares the rate of growth (slope) and level of performance (final status) on a series of CBMs to same-aged peers or to local or national norms, and requiring deficits on both indices (Fuchs & Deshler, 2007). In comparing these final status and growth methods, Fuchs and Deshler concluded that the application of different methods resulted “in varying prevalence rates, severity, and stability. All of this suggests an urgent need for a data-based consensus about what RTI methods of disability identification (in combination with which measures, testing frequencies, and cut points) will be most useful. With such consensus, greater consistency in “reading disability” and “learning disability” designations across schools, districts, and states might be achieved” (p. 134).

RELIABILITY AND VALIDITY OF RTI METHODS

Several studies have evaluated the validity of methods based on RTI, generally showing that different methods disagree about which students are inadequate responders (Burns & Senesac, 2005; Vellutino, Scanlon, Small, & Fanuele, 2006). This suggests reliability problems when efforts are made to identify individual students as inadequate responders or as LD. Because the tests themselves are often highly reliable (> 0.90), however, the reliability problem might be addressed by comparing agreement across different measures and by asking how well single measures or sets of measures identify students who are considered inadequate responders.

This type of study is difficult because there is no “Gold Standard” for identifying inadequate responders or LD, so comparison of students identified across methods is a common approach. In studies of the agreement of different diagnostic approaches, it is common to use a statistical index of agreement such as kappa (κ ; Cohen, 1960), which corrects for chance agreement. By convention, $\kappa > .75$ is considered excellent; $.60-.74$ good; $.40-.59$ fair; and $< .40$ poor (Cicchetti & Sparrow, 1981). Across different areas of research, κ usually ranges between $.4$ and $.8$ (Kraemer, 1979), with strong consensus that $\kappa < .40$ is undesirable.

Only a few studies report the actual levels of agreement and coverage across methods derived from RTI (Barth et al., 2008; Brown Waesche, Schatschneider, Maner, Ahmed, & Wagner, 2011; Burns, Scholin, Kosciulek, & Livingston, 2010; Speece & Case, 2001). These studies show levels of agreement in the poor to fair range that are generally below desirable levels. In addition, they demonstrate the influence of different measures and cut points in the identification of inadequate responders.

Barth et al. (2008) examined differences in cut-points, tests, and methods used to identify inadequate responders after a Grade 1 Tier 2 reading intervention involving 399 students (Mathes et al., 2005). The methods included two assessments incorporating growth that estimated both a slopes-only and dual discrepancy method. In addition, four final status measures assessing decoding, fluency, and comprehension were used. Cut points were set at

0.5., 1.0., and 1.5 standard deviations below the normative group. Across over 800 comparisons, about two-thirds yielded $\kappa < .40$ (i.e., poor).

Burns et al. (2010) used progress monitoring data on 30 students receiving intervention to compare decisions about the adequacy of response using a dual discrepancy method and a method based on progress towards a benchmark. Dividing each set of data into two sets resulted in a poor level of agreement ($\kappa = .29$) for the growth to a benchmark method and a fair level of agreement for the dual discrepancy method ($\kappa = .58$). The authors concluded that the two methods would have resulted in different decisions for 40% of the students and that neither was adequate for high stakes decision making.

Brown Waesche et al. (2011) compared agreement for passage-based and nonsense word fluency CBM assessments in 288,114 students in Grades 1–3. Four definitions of reading disability were used, including absolute low achievement and three RTI definitions: final end point, growth, and a dual discrepancy method. Each definition was evaluated at 6 cut points ranging from the 3rd to the 25th percentile. Rates of agreement were generally poor to fair, with lowest agreement between absolute low achievement and the definitions based on RTI. In addition, agreement was lower for inadequate than adequate responders, especially at lower cut points.

Speece and Case (2001) reported that dual discrepancy models had better coverage for identifying inadequate responders than methods based on IQ-achievement discrepancy or end of year norm-referenced assessments. Only 25% of students with dual discrepancies met IQ-achievement criteria and 62% met low achievement criteria. Speece et al. (2003) found that in Grade 1 and 2 samples, dual discrepancy models identified children as inadequate responders who were not identified by simple low achievement measures. However, this contrast was based on the use of decoding measures for low achievement and IQ-discrepancy methods and fluency assessments for the dual discrepancy methods.

BASIS FOR LOW AGREEMENT

These studies demonstrate that the tests used to assess low achievement and instructional response and the placement of cut points on normally distributed skills affect identification of inadequate responders. There are no established standards for the placement of cut points that indicate low achievement or inadequate response, so variations in practice may reflect the number of children who can be served (Fuchs & Deshler, 2007). Cut points influence agreement and the kinds of misidentifications that tests make, both important, related issues in diagnostic decision making. All other things being equal, lower cut points (i.e., more severity) produce lower false positive rates (i.e., reduced identification of students as LD who are not really LD) and higher false negative rates (i.e., increased rates of missing students who should be identified as LD), which are always a tradeoff in any decision-making context. If the cut point is placed closer to the mean of the distribution, false positive rates rise. This kind of trade-off and the nature of these errors are universal across diagnostic decision making contexts, such as a psychiatric diagnosis, diagnosis of cardiovascular disease, an electroencephalogram for indicating a seizure disorder, or language assessments for identifying a language disorder. Evaluating decision errors is an important component of reliable identification in the development of systems for clinical decision making.

Measurement error and other issues affecting the correlations among tests also affect agreement in test-based decision-making. These problems are magnified if the attributes of interest (in this situation, level of achievement, growth on a CBM) represent unbroken continua (i.e., dimensions) with no qualitative thresholds indicating discrete groups (Markon, Chmielewski, & Miller, 2011). This would imply that LD is not categorical, as is true for many developmental disorders. When traits are continuous, cut-points are arbitrary. Moreover, measurement error makes it difficult to precisely locate a student's true status in relation to any chosen cut point, which reduces agreement across tests. Individuals' observed scores will fluctuate above and below the cut point because of the measurement error in an assessment (Francis et al., 2005). The proximity of the cut-score to the individual's true ability increases the likelihood that observed test scores for the individual will fluctuate across the boundary defined by the cut-score.

There is evidence that the attributes of LD or inadequate response exist on a continuum. Little empirical evidence shows qualitative breaks in the distributions of the observed indicators of these attributes (e.g., achievement, instructional response, cognitive discrepancies) that would indicate discrete categories (Ellis, 1984; Snowling & Hulme, 2011; Vellutino et al., 2006). The precision with which people can be placed in different parts of a continuous distribution is affected by the reliability of individual test scores. When individuals are being collocated on multiple dimensions at the same time (e.g., decoding and fluency, or IQ and achievement), precision is also affected by the intercorrelations of the underlying constructs measured by the tests and the distance of the cut point(s) from the mean of the distribution(s). Lower levels of agreement are expected when cut points are set low on the ability continuum because ability is measured with less precision at either end of the continuum, typically because there are fewer items to discriminate students with the lowest or highest abilities. While it is possible to design adaptive tests to give better precision at the extremes of ability, few tests used for disability identification are designed in such a way. Although not discussed to this point, when different tests are used, sampling differences may also affect agreement. In particular, differences in the normative samples can influence agreement, along with the sample size in which determinations of instructional response are made.

RATIONALE FOR THE PRESENT STUDIES

Agreement is a pivotal issue for any approach to the identification of a student as an inadequate responder and/or as LD, but there is not much information on how well different methods agree or the basis of low agreement. We evaluated agreement and coverage based on two final status norm-referenced tests of decoding and fluency administered after Tier 2 intervention. In addition, we utilized a passage fluency assessment repeatedly administered over the course of a Tier 2 intervention as a progress monitoring measure, which yielded assessments of final status based on CBM fluency rate and dual-discrepancy (slope and final status). In addition to evaluating agreement across these four methods, subgroups identified by different measures were compared on achievement levels to understand how the methods may be performing relative to an external standard. Based on previous research, we expected low agreement across single indicators, but expanded on previous research by evaluating subgroups against external criteria. To address the basis for low agreement, we then

simulated agreement across two reliable final status measures used in the empirical study to address effects of variations in reliability, test intercorrelations, cut points, normative differences, and sample size on agreement and coverage.

STUDY 1: OBSERVED AGREEMENT AND COVERAGE ACROSS INDICES OF RESPONSE TO INTERVENTION

METHODS

PARTICIPANTS AND DEFINITIONS OF INADEQUATE RESPONSE—In a study of inadequate responders to a Tier 2 Grade 1 intervention (Denton et al., 2011), Fletcher et al. (2011) identified 104 inadequate responders based on the separate application of three final status measures: norm-referenced assessments of untimed and timed decoding skills and fluency rates from a passage fluency CBM. Untimed decoding was assessed with the Woodcock-Johnson III Basic Reading Skills composite (WJ; Woodcock et al., 2001), representing a measure used to identify inadequate responders in several studies (Speece et al., 2003; Torgesen, 2000; Vellutino et al., 2006). Timed word reading fluency was assessed with the composite score from the Test of Word Reading Efficiency (TOWRE; Torgesen et al., 1999), which has not been used in previous studies. For both assessments, inadequate response was defined as a standard score below 91 based on previous studies (Speece et al., 2003; Torgesen, 2000; Vellutino et al., 2006). The passage fluency measure, representing a type of CBM commonly used to assess intervention response (e.g., Burns et al., 2010), was the Continuous Monitoring of Early Reading Skills (CMERS; Mathes & Torgesen, 2008), with an oral reading fluency rate of 20 words correct per minute (wcpm) or less indicating inadequate response to intervention. To implement a dual discrepancy method, we computed growth curves for the CBM measure over the 11 waves of progress monitoring over the course of the school year: four prior to intervention at two week intervals and then seven during the intervention at two week intervals. Based on previous research (Fuchs & Fuchs, 1998; Speece et al., 2003), we calculated the mean and standard deviation of the slopes for the typically achieving group and identified students as inadequate responders if both their slopes and final status measure were one standard deviation below the mean of the typical achieving group. The slope cut point was 2.42 words per minute per wave and the final status measure was about 20 words per minute. Note that these assessments are based on local norms.

Examining just the final status measures (WJ, TOWRE, CMERS benchmark), application of the cut points (see Table 1) yielded two subgroups with impairments in decoding and fluency. The first subgroup fell below criteria on both fluency assessments ($n = 19$); the second subgroup fell below criteria on the TOWRE, but not CMERS₂₀ ($n = 10$). Three subgroups had problems in fluency, but not decoding. The first fluency group ($n = 41$) fell below both TOWRE and CMERS₂₀ criteria. Fluency group 2 ($n = 14$) fell below only the CMERS₂₀ criterion, whereas Fluency group 3 ($n = 20$) fell below only the TOWRE criterion. In addition to these students judged to be inadequate responders, there were 85 who received intervention and were above criteria on all three measures and thus judged to be adequate responders. The study also included 69 typically developing students randomly

drawn from the original sample of students not at-risk and therefore not randomized to intervention.

Procedures

Single measures: To understand how the different measures detected students as inadequate responders, we treated the 104 students identified by one or more of the four methods as a pool of students who, at a latent level, could be true inadequate responders. However, at the observed level, measurement error and problems associated with the use of single indicators and different constructs were responsible for the lack of agreement across measures and criteria. In addition, we included the 85 responders who did not meet criteria on any of the instructional response measures. For the 104 inadequate responders, we first assessed the number of students in the pool identified by a single measure (coverage) and the number of students missed by single indicators, which we defined as “false negatives” relative to the total pool of inadequate responders. False positive rates cannot be meaningfully estimated in this situation for the original criteria because, by definition, any individual meeting any one of the original criteria was deemed to be an inadequate responder. Thus, one could say that the false positive rate has been artificially set at zero for the original criteria by virtue of our designation of all individuals meeting at least one criterion as inadequate responders at the latent level and our designation of all individuals meeting none of the criteria as adequate responders. We then evaluated the effect of adjusting the CMERS cut point upwards and also, a dual-discrepancy method on the final status decisions. As the criteria for identification are changed, the impact on both false positive and false negative rates can be judged.

External comparisons: Because there is no “gold standard” for defining an inadequate responder at the observed level, we then evaluated subgroups of inadequate responders against external achievement measures not used to identify inadequate responders. Because of the overlap in the groups and the previous demonstration that subgroups of adequate and inadequate responders could be differentiated on external cognitive variables, we focused on external achievement variables that might differentiate students identified only by a single criterion. If a subgroup validly represents inadequate responders, we expected lower scores on external measures compared to expectations for typically developing children. We included assessments of other academic skills not used to form the groups: norm referenced tests of spelling, reading comprehension, and math computations from the WJ (Woodcock et al., 2001), and a CBM assessment of reading fluency in which the child fills in passages with missing words over a three minute time period (AimsWeb MAZE; Shinn & Shinn, 2002). These measures are well-known with strong reliability and validity as described in Fletcher et al. (2011).

Agreement: Finally, we evaluated agreement across pairs of measures. This assessment was based on κ , which is a chance-corrected index of agreement. The results were interpreted using conventional rubrics for interpreting agreement identified above (> 0.75 excellent; 0.60 – 0.74 good; 0.40 – 0.59 fair; and < 0.40 poor (Cicchetti & Sparrow, 1981). Because κ can be low when agreement seems high (often termed a paradox), we also computed the average positive agreement (P_{pos} , which is the conditional probability of a positive

assignment on one measure if it was found on the other) and the average negative agreement (P_{neg} , which is the conditional probability of the classification made with one measure being negative if it was negative with the other measure) (Cicchetti & Feinstein, 1990). These two indices are not corrected for chance, but give the same value as κ if corrected for chance.

RESULTS

Table 1 presents subgroups from the entire sample ($n = 258$) resulting from application of the four methods. Table 1 provides raw counts, including subgroup and cumulative frequencies that lead directly to computation of coverage and misses, and levels of agreement.

Coverage and misses

Final status indicators: Table 2 presents coverage (percent agreement) and misses for final status single indicators. For the three final status measures, application of the WJ decoding criterion (by definition) missed all of the students in the three fluency-only subgroups, identifying only 29/104 (28%) of the pool of inadequate responders. However, the TOWRE criterion identified the entire decoding /fluency group and 61/74 of the group identified by CMERS₂₀, for total coverage of 90/104 (87%). The CMERS₂₀ indicator identified 19/29 in the decoding/fluency group, and 55/75 of the Fluency group, for a total coverage of 74/104 (71%). For each of these single indicators, the false negative rate is the inverse of percent agreement, and will prove useful when we evaluate the effects of adjusting the CMERS cut point and the dual discrepancy method.

Adjusting the CMERS criterion: The 20 wcpm cut point for CMERS may not be at the 25th percentile used for the WJ and TOWRE because it lacks a norm-referenced standardization (see Fletcher et al., 2011). If we adjust the CMERS criterion to an approximation of the 25th percentile using the Hasbrouck and Tindal (2006) norms (< 28 wcpm), false negatives are reduced from .29 to .11 (16/104). However, 24 students move from the adequate to inadequate responder group, with no typicals changing groups. Although these students would be designated as false-positives based on the definitions that we initially established to create known groups of adequate and inadequate responders, whether these students are truly false positives (as designated by the current criterion) or false negatives (based on the original criteria) is impossible to know, because children's true status as responders is unknown and there is no way to determine if the CMERS₂₀ or the CMERS₂₈ criterion is the correct criterion. Had we begun with the CMERS₂₈ criterion, these students would have begun with designations as inadequate responders, by definition. Because intervention response, as indexed by measures of status or growth, exists on a continuum, changing the criterion changes the designation for some individuals. In this case, such a change would significantly increase the size of the fluency group impaired only on CMERS. As Table 3 shows, this latter subgroup generally has standard scores on other achievement measures in the average range.

Dual-discrepancy comparisons: Table 2 shows that the dual-discrepancy method by itself identified 91% of the 104 inadequate responders (i.e., false negative rate = .09). However,

the identified pool ($n = 134$) is larger, with 29 final status adequate responders and one typically developing reader identified as inadequate responders. Table 3 shows that as a group these children are adequate achievers on achievement measures not used to define the groups. Of the final status sample of 104 inadequate responders, the TOWRE did not identify 13% of the dual-discrepancy group, but the dual-discrepancy criterion missed 8% of the students identified by the TOWRE. The WJ missed 65% of the dual-discrepancy group, but the dual-discrepancy method missed two (2%) of the severely impaired (Fletcher et al., 2011) decoding/fluency group and one (1%) in the CMERS₂₀-only group. If the cut point for CMERS is moved to 28 wcpm, the identified pool rises to 139, again suggesting identification of less impaired students as inadequate responders.

External measures of achievement—To help address possible false positive identifications, Table 3 provides means and standard deviations on external achievement measures for the subgroups identified only by each of the single final status indicators (CMERS₂₀, CMERS₂₈, and TOWRE). In addition, we included the 30 students identified by the DD criterion as inadequate responders and by the other measures as adequate responders. Some students identified by the DD criterion were also identified by the other criteria, but the focus for this table was those inadequate responders identified just by the final status indicators from the pool of 104 inadequate responders and the 30 identified just by the DD criterion. We did not analyze the data for statistical significance because some of the comparisons may be underpowered. Rather, we focused on differences in magnitude by considering descriptive estimates of effect size.

The three CMERS groups (CMERS₂₀-only, CMERS₂₈-only, and DD-only) are very similar to one another over all measures except for CMERS which was used to create groups. Standard score differences between the highest and lowest group means expressed as a fraction of the population SD of 15, range from 0.03 (for TOWRE) to 0.21 for WJ Passage Comprehension. AimsWeb maximum difference relative to a sample based SD of 0.29 was 0.17. The three CMERS-based groups differed from the group defined by CMERS₂₀-TOWRE-WJ by as little as a standardized difference of 0.2 for AimsWeb MAZE to as much as 1.24 (by design) for the WJ Basic Reading. The same three groups differed from the group defined by TOWRE only by as little as a standardized difference of .10 for the AimsWeb MAZE to as much as .79 (by design) on the TOWRE. The scores for the TOWRE-only group tended to be a little lower than the three methods involving the CMERS, especially for reading comprehension, but are still much higher than for students identified by two measures in Fletcher et al. (2011).

Agreement across pairs of measures—Table 4 presents κ , uncorrected agreement, Ppos, and Pneg for each pair of final status and dual discrepancy indicators, now also including agreement involving adequate response. These indices were computed based on Cicchetti and Feinstein (1990), which explicitly addresses the paradox of low κ , higher agreement when there is imbalance in sensitivity and specificity, suggesting a need to report all four indices. How single measures performed is a key to understanding agreement. Because the WJ missed all the fluency-impaired inadequate responders, κ is .19 for WJ and CMERS₂₀, .33 for WJ and TOWRE, and .14 for WJ and DD, all poor., The WJ/TOWRE

and WJ/CMERS₂₀ agree on about two-thirds of the students. This is because positive agreement (Ppos) is low ($< .50$) for each of these three pairs, with negative agreement much higher and imbalanced relative to Ppos for WJ and either TOWRE and CMERS₂₀ (.76 for both). The dual-discrepancy approach introduces discrepancies in both positive and negative agreements, so Ppos and Pneg are both lower, but still imbalanced. The dual discrepancy approach yields $\kappa = .48$ with either TOWRE or CMERS₂₀, with agreement on over 70% of the students and Ppos and Pneg, both $> .70$. Note that although the cut points were comparable for WJ and TOWRE, and both are highly reliable, low agreement is also due to measurement of different constructs. For CMERS₂₀ and TOWRE, which are different indicators of fluency, agreement is higher ($\kappa = .53$), reflecting agreement in over 75% of the students, and Ppos and Pneg $> .75$ and balanced, representing the best performance of pairs of measures. This combination agreed on 70/104 of the inadequate responders (71%), and together identified 90/104 inadequate responders (86%), with the children excluded representing children identified only by CMERS₂₀ (or CMERS₂₈ or DD). Table 3 shows that these 14 children include many with adequate achievement.

STUDY 2: SIMULATION

METHODS

Because Study 1 showed poor to fair levels of agreement, we simulated agreement for the WJ and TOWRE to better understand why agreement was not stronger. The simulation allowed us to manipulate the effect of differences in test reliability and intercorrelations, as well as cut points, normative differences, and sample size.

All data were generated using the SAS 9.2 (SAS Institute, 2008) RANNOR function. For each condition except for the small sample size conditions, 100,000 simulated observations were generated to ensure adequate precision in κ estimates. General true scores were generated for each observation with a mean of 0 and standard deviation of 1. Correlations among construct true scores were attained by allowing each construct score to load on the general true score in proportion to the square root of the desired correlation between constructs. For example, for conditions where only one construct was considered, this loading would be 1. For constructs that correlated .81, the loading was set to .9.

Construct level correlations were chosen to closely approximate the original study conditions by disattenuating the observed correlations for unreliability (Kenny, 1979). The disattenuated correlation is calculated by dividing by the observed correlation by the product of the square roots of the reliabilities of the two measures. Construct level variables assumed perfect reliability in the simulation that assessed κ with perfect measurement. Observed scores with less than perfect reliability were created for each measure by making a weighted composite of the construct true score and random error. Observed scores were then rescaled to have a mean of 100 and standard deviation of 15 and rounded to whole numbers. Cut points were applied to dichotomize the distributions into observations that met and did not meet each criterion for computing κ at standard scores of 80, 90, and 100.

RESULTS

If the WJ and TOWRE were perfectly reliable and correlated 1.0, κ would be 1.0, representing perfect agreement. The cut point would not matter. In our data, the observed κ for WJ and TOWRE was .38. The correlation of WJ and TOWRE was 0.88. The published reliabilities are .98 for WJ and .90 for TOWRE. Consequently, agreement will likely be lower than 1.0 because while highly correlated and reliable, the measures are not perfect.

Imperfect correlation—The disattenuated construct level correlation (correcting for unreliability in the observed measures) between these two assessments was 0.94. Simulating two normally distributed, perfectly measured variables with a correlation of 0.94, and a cut point at the 25th percentile, yielded $\kappa = 0.76$, which is the lower end of the excellent interval. The only factor manipulated was the correlation between the two constructs, so the correlation of 0.94 dropped κ from 1.0 to 0.76. If the cut point was placed lower ($SS = 80$), $\kappa = 0.72$; at 100 $\kappa = 0.77$.

Imperfect reliability—To further address the poor κ observed in our sample, we manipulated the reliability in the measurement while allowing the correlation between the two constructs to be 1.0. When both assessments measured exactly the same construct, but one had a reliability of 0.98 (WJ) and the other had a reliability of 0.90 (TOWRE) with cutoffs at the 25th percentile, κ was again 0.76. At 80, $\kappa = 0.72$; at 100, $\kappa = 0.78$.

Imperfect reliability and correlation—If the construct level correlation is 0.94 and we use the published reliabilities of 0.98 and 0.90, the simulation yields $\kappa = 0.67$ at the 25th percentile ($SS = 90$). At 80, $\kappa = 0.62$; at 100, $\kappa = 0.68$. These figures are higher than the 0.38 for WJ and TOWRE observed in Study 1, but comparable to the best levels of agreement in Table 4.

Normative sample differences—To address possible normative sample differences, we note that 11% of the WJ observations and 35% of the TOWRE observations in our sample fell below the nominal 25th percentile cutoff. Thus, our sample performed lower relative to the norming sample on TOWRE, and higher than the norming sample on WJ. Adjusting the cutoffs in the simulation to replicate the proportions in our sample (standard scores below 83 for the WJ and below 95 for the TOWRE) reduced κ to 0.39. It is possible that the differences between our sample and norming populations were due to selection bias in our sample.

Sample size—Finally, these values of κ are based on 100,000 observations, so sampling error is almost non-existent. This is not the case when sample sizes are smaller. We ran 100 replications of the simulation where the correlation between constructs was 0.94, the reliabilities were .98 and .90, the cutoffs were at the 35th and 11th percentiles, and the sample size (from Study 1) was 258. Across the 100 simulations, mean κ was .39, $SD = .05$ (range 0.27 to 0.51.).

DISCUSSION

Agreement

As the simulation showed, expecting excellent agreement across measures for identifying inadequate response to intervention or evidence of LD is not realistic when using commonly adopted standards and methods for estimating agreement. Agreement rarely reached levels that would be considered “excellent” for κ even when the two measures were highly reliable and the constructs highly correlated, and the samples were large and presumably with the same normative basis. Simulated agreement was much lower when the measures were less reliable or assessed constructs that were different or were less highly correlated, when decisions were based on more extreme cut points, or when study samples were not simple random samples of the population, and the sample size was smaller. Previous studies of identification methods based on response to intervention service delivery frameworks have yielded low rates of agreement largely because of these psychometric and sample size issues.

These conclusions were supported by the empirical study, which showed levels of agreement generally well below the excellent range. Kappas for measures of different constructs (decoding vs. fluency) were poor. For measures assessing the same construct, κ was higher and would be considered fair (Cicchetti & Sparrow, 1981), although values of κ between 0.4 and 0.8 are often considered acceptable (Kraemer, 1979). However, when coverage was examined, all pairs of measures of the same construct disagree on at least 30% of the inadequate responders.

Universality of agreement problems

These problems are not specific to response to intervention applications. The WJ Basic Reading composite and the TOWRE composite are highly reliable norm-referenced assessments. At the nominal 25th percentile, the observed κ was 0.38, indicating poor agreement between these measures, with simulated κ no higher than 0.76 under optimal conditions. Poor agreement would have emerged even if we had used different types of norm-referenced assessments for determining cognitive discrepancies (e.g., IQ-achievement discrepancies, cognitive profiles). Previous simulations of achievement cut points and in evaluations of observed cut points based on IQ-achievement discrepancies as well as simple low achievement (Francis et al., 1996; 2005; Macmann, Barnett, Lombard, Belton-Kocher, & Sharpe, 1989) have shown substantial fluctuations on both sides of different cut points on repeated assessments or across methods and measures. As these simulations of IQ-achievement discrepancies showed, the same problems would emerge if cognitive discrepancies were the basis for identification. The measures are correlated and are not perfectly reliable, so agreement and coverage will vary with any method. Such problems are inherent to attempts to create groups based on cut points imposed on normally distributed variables that are not perfectly reliable, are intercorrelated, sometimes measure different constructs, and have different normative samples.

Multiple indices may improve identification

When multiple achievement criteria were used, a more consistent level of agreement could be reached. Although the pool seems large, this is partly the consequence of applying multiple criteria and a relatively weak intervention designed to mimic Tier 2 interventions commonly employed by schools (see Denton et al., 2011). A single measure of untimed word reading (WJ) detected a group of severely impaired struggling readers, but missed a large number of students with fluency difficulties, many of whom also struggle with comprehension. Comparisons of CBM fluency and untimed decoding measures as in Speece et al. (2003) may be misleading because they measure different constructs. It also seems reasonable that decoding accuracy is a pre-requisite for decoding fluency because decoding accuracy is a necessary, but not a sufficient condition for decoding fluency. Indeed, in this study, all of the children who scored below the 25th percentile on WJ Basic Reading met at least one of the criteria for fluency problems. Agreement was much higher for end of year final status CBM measures and a timed, norm-referenced assessment of fluency (TOWRE). Moreover, these two measures together agreed on 70/104 students, excluding the 14 identified only by the CMERS₂₀ who were largely in the average range on other academic tasks (Table 3), and the 20 identified by the TOWRE, but not the CMERS. This level of agreement using two final status methods, one representing a norm referenced measure and the other derived from a progress monitoring measure, was also observed in Barth et al. (2008). In our previous study of this sample (Fletcher et al., 2011), subgroups identified by two indicators had much lower levels of achievement than is apparent in Table 3.

Sole use of a CBM level of performance measure following intervention may not be adequate for identification of inadequate response because many of the children identified just by the CMERS₂₀ show adequate levels of achievement on other variables. If the CMERS₂₀ criterion was adjusted to an approximation of the 25th percentile (< 28 wcpm), many less impaired students shifted into the inadequate responder group. Our application of the dual-discrepancy method raised issues similar to sole application of the CMERS₂₀ benchmark, missing some students in the final status pool of 104 inadequate responders, but identifying many questionably impaired students as inadequate responders. Speece et al. (2003) argued that dual discrepancy methods had better coverage than assessments using end of the year norm-referenced achievement tests (only decoding) because of the assessment of growth. This finding was not apparent in this study, possibly because we used a norm-referenced fluency measure. The dual discrepancy measure also yielded potential false positives, possibly because of the use of local norms. The findings should be replicated using a norm referenced progress monitoring measure and additional manipulations of cut points.

Implications for practice

The analysis of empirical agreement data suggests that decisions about intervention response should not be made with a single measure. However, these results are limited to the sample and measures employed. Other studies should attempt replication of the use of multiple measure approaches for identifying inadequate responders. Using multiple measures may ensure broader coverage, partly because this approach is less subject to the psychometric issues related to the capacity of a single test or criterion to precisely assess a student's status

relative to a set cut point. An alternative would be to use confidence intervals for single measures, but simply assessing outcomes based on level of CBM performance and then administering short norm-referenced assessments of fluency may provide better agreement and coverage.

Little in these studies directly addresses the issue of the level of cut points. Agreement was generally higher closer to the mean of the distribution and lower if the cut point was further from the distribution mean. As the comparison of the two CMERS cut points shows, there will inevitably be tradeoffs of false negative and false positive errors as the cut point is changed, all other things being equal. Ultimately such decisions are dependent on the base rate and resources.

An approach that takes advantage of the CBM data collected routinely by schools with the addition of short norm-referenced assessments may be easier for many schools to implement given the need to compute slopes based on CBM assessments if growth data are utilized and the difficulties in estimating and interpreting confidence intervals. The additional assessment would require less than 15 minutes and could identify a clearly reading-impaired pool of students. Such an approach could represent a multiple gating approach where highly sensitive measures (e.g., CBM progressing monitoring benchmarks) are used to trigger a need for short norm referenced tests, thus ensuring stronger reliability and a form of incremental validity. In addition, other data that might explain inconsistent performance across the criteria could be examined, such as measures of reading comprehension and other academic areas such as math. Further, attention, behavior, and other aspects of adaptive functioning may need to be evaluated for some students in order to develop a comprehensive intervention plan.

In returning to the identification methodology adopted by several states and outlined in the introduction, the use of criteria based on both low achievement and inadequate instructional response was supported by the empirical study. The problems that may emerge involve the use of a rigid cut point, especially given the proximity of many state-adopted thresholds to the floor of the tests, which may be a particular disadvantage for younger students. At the very least, the standard error of measurement should be considered in decision making. There will still be disparities across districts because different tests will be used, so some effort to standardize assessments, as advocated by Fuchs and Deshler (2007), may be useful

As mandated by IDEA (2004) and clearly embodied in the rules adopted by most states after IDEA 2004, identification of inadequate responders or children eligible for special education as LD requires multiple sources of information and is ultimately a decision made by an interdisciplinary team. The examples of methods in this paper are only part of the information needed for identification of LD. However, how interdisciplinary teams make decisions and basic issues like agreement and coverage are not well understood and should be studied. Despite the provisions for team based decision making, this question has not been adequately studied for issues related to assessments of instructional response and LD.

Alternatives to a purely test-based formula-based approach for identification of inadequate response or LD must be sought, which is clearly illustrated in the present set of studies. In

this regard, understanding that the attributes of inadequate responders and/or LD are dimensional and not representative of discrete subgroups suggests that approaches to identification should be more intervention oriented and less oriented to finding the right student. There are medical disorders where the defining and measured attributes are dimensional, such as obesity (weight) and hypertension (blood pressure). Identification of these disorders is a matter of degree and not of identifying discrete subgroups that are different in kind. Intervention is related to multiple indicators tied to assessments of functional outcomes (Ellis, 1984; Snowling & Hulme, 2011). We do not have these types of data in the LD area. In their simulation of decision making based on WISC-III profiles, Macmann and Barnett (1997) suggested as an alternative to psychometric decision making, which was not reliable in their study: “hypotheses about needed skills (what to teach) and needed environments to support the development of those skills (how to teach) can be guided by research on effective instruction and systematic analyses of classroom environments in relation to those principles” (p. 230–231).

One advantage of approaches to identification emanating from RTI frameworks, especially if they can be moved away from rigid cut points and psychometric criteria, is their focus on instruction and modifications of instruction as an ongoing process. In this respect, the critical question may well be need for services triggered by multiple dimensions related to treatment as an ongoing process and not as a static assessment with a single time point. Historically, IDEA has focused schools and practitioners on the idea that diagnosis is needed in order to provide treatment. In fact, this approach implies that diagnosis reflects a discrete, static category that may not exist. Many of the eligibility decisions for special education (and other domains) are predicated on this idea, as well as the idea that intervention follows (and requires) diagnosis. For LD, intervention may be needed for diagnosis because there is no way to identify someone as LD in the absence of adequate opportunity to learn. In a RTI model, intervention precedes diagnosis, and thus makes assessment of both the disorder and evidence that the disorder is a disability measureable, thus conforming to the two-pronged criteria inherent in IDEA (i.e., to receive special education status, there must be evidence of a disorder as well as evidence of educational need related to the disorder).

Acknowledgments

This research was supported by grants P50 HD052117, Texas Center for Learning Disabilities, and K08 HD068545-01A1, Language, Cognitive, and Neuropsychological Processes in Reading Comprehension, from the Eunice Kennedy Shriver National Institute of Child Health and Human Development. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health and Human Development or the National Institutes of Health.

REFERENCES

- Barth AE, Stuebing KK, Anthony JL, Denton CA, Mathes PG, Fletcher JM, Francis DJ. Agreement among response to intervention for identifying responder status. *Learning and Individual Differences*. 2008; 18:296–307. [PubMed: 19081758]
- Bradley, R.; Danielson, L.; Hallahan, D., editors. *Identification of learning disabilities: Research to practice*. Mahwah NJ: Erlbaum; 2002. www.air.org/ldsummit
- Brown Waesche JS, Schatschneider C, Maner JK, Ahmed Y, Wagner RK. Examining agreement and longitudinal stability of reading disability using the affected-status agreement index. *Journal of Learning Disabilities*. 2011; 44:296–307. [PubMed: 21252372]

- Burns MK, Senesac SV. Comparison of dual discrepancy criteria to assess response to intervention. *Journal of School Psychology*. 2005; 43(5):393–406.
- Burns MK, Scholin SE, Kosciulek S, Livingston J. Reliability of decision-making frameworks for response to intervention for reading. *Journal of Psychoeducational Assessment*. 2010; 28:102–114.
- Cicchetti DV, Feinstein AR. High agreement but low Kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*. 1990; 43:551–558. [PubMed: 2189948]
- Cicchetti DV, Sparrow SS. Developing criteria for assessing inter-rater reliability of specific items: Application to assessments of adaptive behavior. *American Journal of Mental Deficiency*. 1981; 86:127–137. [PubMed: 7315877]
- Cohen J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960; 10:37–46.
- Denton CA, Cirino PT, Barth AE, Romain M, Vaughn S, Wexler J, Fletcher JM. An experimental study of scheduling and duration of “Tier 2” first grade reading intervention. *Journal of Research on Educational Effectiveness*. 2011; 4(3):208–230. [PubMed: 21796271]
- Ellis AW. The cognitive neuropsychology of developmental (and acquired) dyslexia: A critical survey. *Cognitive Neuropsychology*. 1984; 2:169–205.
- Fletcher JM, Stuebing KK, Barth AE, Denton CA, Cirino PT, Francis DJ, Vaughn S. Cognitive correlates of inadequate response to intervention. *School Psychology Review*. 2011; 40:2–22.
- Francis DJ, Fletcher JM, Stuebing KK, Lyon GR, Shaywitz BA, Shaywitz SE. Psychometric approaches to the identification of learning disabilities: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities*. 2005; 38:98–108. [PubMed: 15813593]
- Fuchs D, Deshler DK. 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. Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice*. 1998; 13:204–219.
- Hale J, Alfonso V, Berninger V, Bracken B, Christo C, Clark E, Yalof J. Critical issues in response-to-intervention, comprehensive evaluation, and specific learning disabilities identification and intervention: An expert white paper consensus. *Learning Disability Quarterly*. 2010; 33:223–236.
- Hasbrouck J, Tindal GA. Oral reading fluency norms: A valuable assessment tool for teachers. *The Reading Teacher*. 2006; 59:636–644.
- Individuals with Disabilities Improvement Act of 2004, Pub. L. 108-466. Federal register. Vol. 70(No. 118):35802–35803.
- Kenny, DA. Correlation and causality. New York: Wiley; 1979.
- Kraemer HC. Ramifications of a population model for k as a coefficient of reliability. *Psychometrika*. 1979; 44:461–472.
- Macmann GM, Barnett DW. Myth of the master detective: Reliability of interpretations for Kaufman's “intelligent testing” approach to the WISC-III. *School Psychology Quarterly*. 1997; 12:197–234.
- Macmann GM, Barnett DW, Lombard TJ, Belton-Kocher E, Sharpe MN. On the actuarial classification of children: Fundamental studies of classification agreement. *The Journal of Special Education*. 1989; 23(2):127–149.
- Markon KE, Chmielewski M, Miller CJ. The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*. 2011; 137(5):856–879. [PubMed: 21574681]
- Mathes PG, Denton CA, Fletcher JM, Anthony JL, Francis DJ, Schatschneider C. An evaluation of two reading interventions derived from diverse models. *Reading Research Quarterly*. 2005; 40:148–183.
- Mathes, PG.; Torgesen, JK. Continuous monitoring of early reading skills. Dallas, TX: Istation; 2008.
- Reynolds CR, Shaywitz SE. Response to intervention: Ready or not? Or watch-them-fail. *School Psychology Quarterly*. 2009; 24:130–145. [PubMed: 20169006]
- SAS Institute, Inc. SAS Release 9.2. Cary, NC: SAS Institute Inc; 2008.
- Shinn, MR.; Shinn, MM. AIMSweb training workbook: Administration and scoring of reading maze for use in general outcome measurement. Eden Prairie, MN: Edformation; 2002.

- Speece DL, Case LP. Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology*. 2001; 93:735–749.
- Speece DL, Case LP, Molloy DE. Responsiveness to general education instruction as the first gate to learning disabilities identification. *Learning Disabilities Research & Practice*. 2003; 18:147–156.
- Snowling MJ, Hulme C. Annual research review: The nature and classification of reading disorders: a commentary on proposals for DSM-5. *Journal of Child Psychiatry and Psychology*. 2012; 53:593–607.
- Torgesen JK. Individual responses in response to early interventions in reading: The lingering problem of treatment resisters. *Learning Disabilities Research and Practice*. 2000; 15:55–64.
- Torgesen, J.; Wagner, R.; Rashotte, C. *Test of Word Reading Efficiency*. Austin, TX: Pro-Ed; 1999.
- Vellutino FR, Scanlon DM, Small S, Fanuele DP. Response to intervention as a vehicle for distinguishing between children with and without reading disabilities: evidence for the role of kindergarten and first-grade interventions. *Journal of Learning Disabilities*. 2006; 39(2):157–169. [PubMed: 16583795]
- Woodcock, RW.; McGrew, KS.; Mather, N. *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside Publishing; 2001.

Table 1

Frequency of Students Identified as Inadequate Responders by Four Methods

Group	CMERS ₂₀	WJ	TOWRE	Dual-Discrepancy	Frequency
Decoding/Fluency Problem	Yes	Yes	Yes	Yes	19
Decoding/Fluency Problem	No	Yes	Yes	Yes	8
Decoding/Fluency Problem	No	Yes	Yes	No	2
Fluency Problem	Yes	No	Yes	Yes	41
Fluency Problem	Yes	No	No	Yes	13
Fluency Problem	Yes	No	No	No	1
Fluency Problem	No	No	Yes	Yes	14
Fluency Problem	No	No	Yes	No	6
Responder	No	No	No	Yes	29
Responder	No	No	No	No	56
Typical	No	No	No	Yes	1
Typical	No	No	No	No	68

Yes = meets criterion for inadequate response; Continuous Monitoring of Early Reading Skill 20 words correct per minute; Woodcock-Johnson III Basic Reading Skills Cluster 25th percentile; Test of Word Reading Efficiency composite 25th percentile; Dual Discrepancy 1 SD below the mean of the typical students on both intercept and slope on CMERS.

Table 2

Coverage (% Identified) and Misses Based on Final Status and Dual Discrepancy Assessments of 104 Inadequate Responders

Measure	% identified	False Negatives
WJ Basic Reading	28	0.72
TOWRE	87	0.13
CMERS ₂₀	71	0.29
Dual Discrepancy ^a	91	0.09

Notes. WJ = Woodcock Johnson-III Basic Reading 25th percentile; TOWRE = Test of Word Reading Efficiency composite 25th percentiles; CMERS₂₀ = Computerized Monitoring of Early Reading Skills (20 words correct per minute)

^a Identified 29 "responders" on final status measures and one typical as inadequate responder.

Table 3

Means and Standard Deviations for Single Indicator Subgroups

Indicator	CMERS ₂₀ -only (n = 14)		CMERS ₂₈ -only (n = 19)		DD-only (n = 30)		TOWRE-only (n = 20)	
	M	SD	M	SD	M	SD	M	SD
CMERS ₂₀	15.5	3.7	23.7	2.2	26.9	6.7	27.8	6.5
TOWRE	97.4	4.7	97.7	5.8	97.3	5.3	85.9	3.3
WJ Basic Reading	108.7	4.3	109.7	7.8	109.0	8.8	99.0	4.5
WJ Passage Comprehension	94.1	5.1	97.3	6.0	97.0	6.5	88.2	6.5
WJ Spelling	102.7	8.2	101.0	8.7	100.9	9.3	96.5	5.7
WJ Calculations	105.5	14.5	107.1	10.4	106.0	12.2	100.6	11.7
AimsWeb Maze ^a	.30	.21	0.30	.19	.35	.20	.27	.20

Notes. TOWRE = Test of Word Reading Efficiency composite; CMERS = Computerized Monitoring of Early Reading Skills with cut points below 20 or 28 words correct per minute; DD = Dual Discrepancy. WJ = Woodcock Johnson-III Basic Reading.

^a AimsWeb Maze score = number correct/ no. attempted. The italics indicate criterion scores for this index, which are always lower by definition.

Table 4
Agreement (kappa) Among Pairs of Indicators Based on Final Status Assessments of 104 Inadequate Responders and 85 Responders

Indicators	K	% Agreement	Ppos	Pneg	Imbalance
WJ - TOWRE	.33	68	.49	.76	.32
WJ - CMERS ₂₀	.19	66	.37	.76	.39
TOWRE - CMERS ₂₀	.53	77	.73	.79	.82
WJ - DD	.14	48	.35	.56	.23
TOWRE - DD	.48	74	.77	.70	.73
CMERS ₂₀ - DD	.48	73	.74	.71	.60

Notes. WJ = Woodcock Johnson-III Basic Reading 25th percentile; TOWRE = Test of Word Reading Efficiency composite 25th percentile; CMERS₂₀ = Computerized Monitoring of Early Reading Skills 20 words correct per minute; DD = Dual Discrepancy.