Published in final edited form as:

Learn Individ Differ. 2008 ; 18(3): 308–315. doi:10.1016/j.lindif.2008.04.005.

The importance of measuring growth in response to intervention models: Testing a core assumption[☆]

Christopher Schatschneider^{a,c,*}, Richard K. Wagner^{a,c,*}, and Elizabeth C. Crawford^{b,c,*}

^aDepartment of Psychology, Florida State University, 1107 W. Call Street, Tallahassee, Florida 32306-4301, USA

^bDepartment of Communication Disorders, Florida State University, Tallahassee, Florida 32306, USA

^cFlorida Center for Reading Research, Tallahassee, Florida 32301, USA

Abstract

A core assumption of response to instruction or intervention (RTI) models is the importance of measuring growth in achievement over time in response to effective instruction or intervention. Many RTI models actively monitor growth for identifying individuals who need different levels of intervention. A large-scale ($N=23,438$), two-year longitudinal study of first grade children was carried out to compare the predictive validity of measures of achievement status, growth in achievement, and their combination for predicting future reading achievement. The results indicate that under typical conditions, measures of growth do not make a contribution to prediction that is independent of measures of achievement status. These results question the validity of a core assumption of RTI models.

Keywords

RTI; Identification; Reading; Dyslexia

“The greatest enemy of the truth is not the lie-deliberate, contrived, and dishonest, but the myth-persistent, pervasive, and unrealistic.” John F. Kennedy

1. Introduction

Reading disability or developmental dyslexia refers to unexplained poor performance in reading. When the concept of reading disabilities was formalized in the 1960's, the common assumption made when children were unable to learn to read was that they were impaired intellectually. Later, it was recognized that an inability to learn to read could exist despite the absence of general intellectual impairment (Kirk, 1962). As a result, the traditional operational definition of reading disability in the United States was rooted in a comparison of performance in reading and performance on a measure of cognitive ability. In addition to requiring that the observed poor performance in reading could not be explained as a result of a general intellectual impairment, other potential explanations that needed to be ruled out

[☆]Support for this research has been provided by Grant P50 HD052120 from NICHD.

© 2008 Published by Elsevier Inc.

*Corresponding authors. schatschneider@psy.fsu.edu (C. Schatschneider), rk.wagner@psy.fsu.edu (R.K. Wagner), ecrawford@fcrr.org (E.C. Crawford).

included lack of an opportunity to learn and impaired sensory capacities in areas required for reading, especially vision (Lyon, Shaywitz, & Shaywitz, 2003).

The need to distinguish among low-achievers in reading is rooted in the acknowledgement that there are multiple causes of poor reading skills, including low general cognitive ability, minimal opportunities to learn how to read at home (or perhaps also at school), sensory impairments, or possibly the existence of some specific neurological impairment. Presumably the identification of specific causes would then lead to specific treatments for poor reading. The traditional approach in looking for children with specific neurological impairments in reading has been to use adequate overall intellectual function as a proxy to rule out these other causes of low reading achievement.

However, the traditional approach to identification has been challenged over the years on a number of grounds. The most recent challenge has come from proponents of an approach to identification of individuals with reading disability on the basis of their failure to respond to provision of effective instruction and intervention. In the present article, we (a) review concerns that have been raised about the traditional approach to identification, (b) review approaches to identification based on response to instruction and intervention, (c) present results from a large-scale, longitudinal study comparing the predictive validity of measures of achievement status, growth in achievement, and their combination for predicting future reading achievement, and (d) consider implications of the results for evaluating the potential of response to intervention models for addressing limitations associated with traditional models.

2. The discrepancy approach to identification

The most common approach for determining whether a child is eligible for special education services in the United States due to the presence of a specific learning disability in reading has required evidence of a discrepancy between IQ and reading performance. This IQ-achievement discrepancy approach has come under attack for three major reasons (Wagner, in press). These include viewing the approach as a “wait to fail” model, having concerns about the reliability of IQ-achievement discrepancy scores, and questioning the validity and educational relevance of the distinction between poor readers who are IQ-discrepant and poor readers who are not (Fletcher, Lyon, Fuchs, & Barnes, 2007; Fletcher, Francis, Rourke, Shaywitz, & Shaywitz, 1992; Fletcher, Morris, & Lyon, 2003; Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Lyon et al., 2003; President's Commission on Excellence in Special Education, 2002; Siegel, 1992; Spear-Swerling & Sternberg, 1996; Stanovich, 1991; Stanovich & Siegel, 1994; Stuebing et al., 2002.). We briefly review each of these concerns about the IQ-achievement discrepancy model.

2.1. A “wait to fail” model?

Under the traditional approach to identification, it indeed has been the case that most children with reading disability have not been identified and provided additional services until second grade. This is problematic to the extent that reading problems become more intractable the longer they exist. There is support for the idea that reading problems become increasingly intractable from the results of longitudinal correlational studies of the development of reading. Individual differences in reading skills become remarkably stable by second grade (Francis et al., 1996; Wagner et al., 1997). For example, Wagner et al. (1997) reported year-to-year correlations for word-level decoding of .84, .96, and .96, for the time periods 1st to 2nd grade, 2nd to 3rd grade, and 3rd to 4th grade respectively, in their five-year longitudinal study of over 200 children. A related concern of a “wait to fail” approach is that a child who struggles with early reading may experience unfortunate concomitants such as a negative academic self-concept or an aversion to schooling.

There are at least three reasons for the fact that most children with reading disability are not identified and provided additional services until second grade. First, reading instruction in the United States begins in earnest in first grade when children typically are six years old. This marks the beginning of compulsory education. Although there is a trend of teaching more about reading in the previous kindergarten year and even in preschools, there is great variability in the educational experience of children prior to first grade. Kindergarten typically is optional, and may exist as either a half- or full-day program. Preschools vary tremendously in the extent to which early literacy activities are emphasized. If reading instruction does not begin in earnest until first grade, any approach that is based on an observed failure to acquire reading skills that are taught formally in school cannot hope to identify children until some time after such instruction has begun. The second, related reason is that common measures of reading achievement show floor effects through the beginning of first grade, as might be expected if they assess reading skills that are taught in first grade. These floor effects make it virtually impossible to observe a discrepancy between aptitude and achievement much before the end of first grade or early second grade. The third reason is that resources limit the number of students who can be referred for evaluation, which adds an additional time lag into the provision of services.

There is no disputing the fact that the traditional IQ-achievement discrepancy approach typically does not result in getting help to children with reading disability until second grade on average, and this is problematic. However, new developments in our understanding of early literacy potentially could eliminate or reduce the “wait-to-fail” nature of the traditional approach. Reading should not be thought of as a skill that appears out of whole cloth when children are first taught to read. Rather, reading is better conceptualized as a developmental phenomenon that builds on early print awareness, rudimentary phonological awareness, and vocabulary. These skills can be assessed reliably in children as young as 3 years of age, and are highly predictive of later decoding (Burgess & Lonigan, 1998; Lonigan, Wagner, Torgesen, & Rashotte, 2007). If measures of print awareness replace measures of reading achievement, it should be possible to identify children whose print awareness is discrepant from their IQ prior to their entry in kindergarten or first grade. Of course, it would be important to evaluate whether early IQ-print awareness achievement discrepancies are stable and become IQ-reading achievement discrepancies at a later time.

2.2. Are IQ-achievement discrepancy scores unreliable?

In practice, an operational definition of an IQ-achievement discrepancy is required if it is to be useful for identification of individuals with reading disability. For example, an operational definition might be a difference of 15 standard score points between accepted measures of cognitive ability and of reading. Under some circumstances, differences scores can be unreliable. The formula for calculating the reliability of a difference score is the following:

$$r_{\text{diff}} = \frac{[(r_{aa} + r_{bb}) / 2] - r_{ab}}{1 - r_{ab}}$$

In this formula, r_{diff} is the reliability of the difference score, r_{aa} and r_{bb} are the reliabilities of the two scores used to create the difference score, and r_{ab} is the correlation between these two scores. When the two scores that comprise the difference score are uncorrelated, the reliability of the difference score equals the average of the reliabilities of the two scores. This can be seen in the formula above by using zero as the value of r_{ab} . The numerator simplifies to $(r_{aa} + r_{bb}) / 2$ (i.e., the average of the reliabilities of the two scores) and the denominator simplifies to one. But now consider what happens as the correlation between the two scores begins to approach their reliabilities. The numerator approaches zero as r_{ab}

approaches $(r_{aa} + r_{bb}) / 2$. In the extreme case, when the correlation between the two scores is equal to their average reliability, the numerator becomes zero, and consequently, the reliability of the difference score becomes zero. For most common measures of IQ and achievement, their reliabilities are quite high but they are correlated substantially. Therefore, the reliability of IQ-achievement differences tends to be less than the reliability of typical IQ or achievement scores but certainly greater than zero.

Fletcher et al. (2007) identified two additional problems that reduce the reliability of IQ-achievement difference scores. First, when an IQ-achievement discrepancy score is used for identification, a cutoff score must be identified. Because the distributions of IQ and achievement scores are continuous, so is the distribution of IQ-achievement discrepancy scores. Consequently, an arbitrary cut-point is imposed on a continuous distribution and this can have a detrimental effect on reliability. Reliability is reduced because individuals who score close to the cut-point are likely to vary on which side of the cut-point they land upon repeated testing. In an empirical investigation of this phenomenon using both real data from the Connecticut Longitudinal Study and simulated data, Francis et al. (2005) reported that classification decisions showed instability longitudinally that could be explained in part to the imposition of arbitrary cut-points on a continuous distribution.

The second problem identified by Fletcher et al. (2007) that could affect the reliability of classification decisions based on IQ-achievement discrepancy scores is regression to the mean. Regression to the mean refers to the fact that when individuals are selected on the basis of an extreme score on a test, repeated testing with any test that is substantially correlated with the first test will result in scores that are closer to the mean of the tests than was the original extreme score. The reason regression to the mean occurs when extreme scores are selected is unreliability in the test. Any observed test score that contains measurement error may differ from an individual's "true" score. Sometimes the observed score will be higher than the true score, and other times the observed score will be lower than the true score. If extreme low scores are selected, they will contain a greater proportion of individuals whose observed scores were lower than their true score for the specific test used for selection. Conversely, if high scores are selected, they will contain a greater proportion of individuals whose observed scores were higher than their true scores for the specific test used for selection. When tested again, the scores will regress to the "true" scores, which means that low scorers will tend to score higher and high scorers will tend to score lower.

If lessened reliability of IQ-achievement discrepancy scores was the only concern associated with the traditional approach, the concern could be remedied by obtaining multiple measures of ability and achievement. The multiple measures could be used to create multiple discrepancy scores, and reliability could be enhanced by using the mean or median discrepancy score. However, other concerns remain, including concerns about the validity of IQ-achievement scores.

2.3. Are IQ-achievement discrepancy scores valid?

Do IQ-discrepant poor readers differ meaningfully from non-IQ-discrepant poor readers? For common measures of decoding and of phonological awareness, phonological memory, and rapid naming, differences tend to be small albeit statistically significant (Fletcher et al., 2007). As would be expected, differences tend to be higher for more IQ-related variables such as vocabulary and syntax. Does IQ-discrepancy status have implications for the kind of intervention that is required? The consensus with respect to interventions designed to improve word-level decoding skills is that IQ-discrepancy status is irrelevant (Fletcher et al., 1992, 2007, 2003; Francis et al., 1996; Lyon et al., 2001, 2003; Shaywitz, Fletcher, Holahan,

& Shaywitz, 1992; Shaywitz & Shaywitz, 2003; Siegel, 2003; Stanovich, 1991; Stanovich & Siegel, 1994).

However, what happens when we consider reading connected text for comprehension, the actual form of reading that matters in the everyday world? Most arguments that question the pedagogical validity of the distinction between IQ-discrepant and non-discrepant poor readers focus on word-level decoding. There may be reason to be skeptical, based on logical grounds, that IQ-discrepant and non-discrepant poor readers are equivalent with respect to their educational and intervention needs when we move from word-level decoding to reading connected text for comprehension, although there is little direct empirical evidence to support this position. Vocabulary, which is one of the best predictors of reading comprehension once children have mastered basic decoding, also provides the single best estimate of verbal aptitude. Although it is true that poor decoders show weaknesses in measures of phonological processing regardless of their IQ-discrepancy status, some poor decoders have weaknesses that are confined largely to phonological processing whereas others have weaknesses that extend to broader oral language areas including vocabulary and comprehension. Even if children with broader oral language deficiencies can acquire adequate decoding skills, their comprehension will be affected by their limited vocabulary and other language deficiencies (Torgesen, 2000).

3. Response to instruction or intervention (RTI) approaches to identification

According to the report of the President's Commission on Excellence in Special Education (2002), traditional approaches for identifying individuals with learning disabilities and providing intervention have not succeeded in helping individuals meet the ever increasing literacy-related demands of modern society. One solution proposed in the report was to base identification on limited response to effective instruction or intervention (RTI) (Case, Speece, & Molloy, 2003; Compton, Fuchs, Fuchs, & Bryant, 2006; Fuchs & Fuchs, 1998; Fuchs, Fuchs, McMaster, & Al Otaiba, 2003; Fuchs, Mock, Morgan, & Young, 2003; Lyon et al., 2003; McMaster, Fuchs, Fuchs, & Compton, 2005). An RTI approach seeks to avoid the possibility that the observed poor reading results from inadequate instruction by first ensuring that students receive effective reading instruction in their general education classrooms. Response to effective reading instruction is measured, and only students who don't respond are considered as possible candidates for more focused intervention, and if warranted, eventual identification as an individual with a learning disability.

When utilized as an instruction model, a Response to Instruction (RTI) model primarily uses a three-tier model with the core reading program or the main curriculum identified as Tier I instruction that all children receive in the general education classroom provided by the general education teacher. Tier II intervention involves instruction that supplements the Tier I curriculum providing extra strategies to support the students at risk for reading difficulties, oftentimes delivered to small groups of identified children within the general classroom instructional framework. Tier III, the most intense instruction in a three-tier model, consists of instruction for children who continue to struggle learning to read even with the support of the general core curriculum and supplemental support of Tier II. Tier III instruction is more individualized for the specific student and is often delivered outside of the regular classroom instruction (Vaughn & Fuchs, 2003; Reschly, 2005; Vaughn & Chard, 2006).

When schools incorporate RTI into the process of identification, they typically screen all students at the beginning of the year and determine which students require intervention. The school then monitors all student progress periodically over the year while more frequently progress monitoring the students identified as being 'at risk' and who are receiving

intervention in the area being assessed. If children continue to demonstrate little or no growth as a result of this more intense instruction, these children could potentially be labeled as having a learning disability (Fletcher et al., 2007).

At a general level, response to intervention is closely related to other approaches to assessment. For example, Grigorenko (in press) carried out a theoretical analysis of similarities and differences between response to intervention and dynamic assessment. She concluded that they are two sides of the same coin. Both approaches to assessment are examples of a family of methodologies that blend assessment and intervention into a single, holistic activity. Both approaches are responses to addressing important practical problems, and both are embedded in a context that favors service provision over diagnosis. There are differences between the approaches as well. For example, the history and amount of existing research is considerably greater for dynamic assessment than for response to intervention. One of the insights provided by Grigorenko's analysis is the likely value of mining the research literature on dynamic assessment for implications on effective implementation of response to intervention models.

In fact, one can argue that the traditional IQ-achievement discrepancy form of identification is an example of a response to instruction model broadly defined. Children arrive at school and receive instruction in reading. Children who do not respond to instruction fall behind and eventually become eligible for identification and services by virtue of a discrepancy between their reading achievement and expectations based on their cognitive ability.

Although there are obvious parallels between RTI and other approaches to assessment, recently proposed RTI models can be distinguished from other forms of assessment in some of their more detailed assumptions. Perhaps the most important of these detailed assumptions is the belief that one must actively monitor progress in the context of the provision of effective instruction and intervention. Failure to respond appropriately under these conditions becomes the operational definition of a reading disability.

As yet, consensus has not emerged around a single method for identifying non-responders in an RTI model. Fuchs and Deshler (2007) described five alternative methods that have been used in research studies or in practice. Vellutino et al. (1996) used a *median-split of slope estimates* from a longitudinal study that used hierarchical linear modeling of growth. Students below the median were classified as non-responders.

Torgesen et al. (2001) used a *normalization criterion* that consisted of scoring at the 25th percentile on nationally standardized tests of reading. Individuals scoring below the normalization criterion were considered to be non-responders. Good, Simmons, and Kame'enui (2001) used a *final benchmark criterion* of falling below an empirically-derived benchmark on oral reading fluency. Fuchs and Fuchs (1998) used a *dual-discrepancy criterion* of scoring one or more standard deviations below classroom peers on both slope and final status on reading fluency. Finally, Fuchs, Fuchs, and Compton (2004) used a *slope discrepancy criterion* in which non-responders are identified using a cut-point on slope.

Results from a recent study by Fuchs, Compton, Fuchs, Bryant, and Davis et al. (in press) indicate that the alternative methods for identifying non-responders identify different students as nonresponders. They carried out a longitudinal study of 253 first grade children who were selected on the basis of being at risk for reading disability. End of second grade reading disability status was the criterion to be predicted. Both traditional and RTI approaches were compared in terms of their prediction of end of second grade reading disability status. The traditional approaches evaluated by the study were initial low achievement and IQ-achievement discrepancy at the end of first grade. The RTI approaches that were examined were normalization, final benchmark, median-split of slopes, slope

discrepancy, and dual-discrepancy. The approaches were compared in terms of prevalence estimates, sensitivity, and specificity. In the present context, sensitivity was defined as the proportion of children with a reading disability who were correctly identified as such by the identification procedure. Specificity was defined as the proportion of children who were adequate readers who were correctly identified as such by the identification procedure. Acceptable sensitivity and specificity rates were set at .80 or higher.

Beginning with the traditional identification procedures, initial low achievement in word identification yielded reasonable prevalence estimates, acceptable specificity, and poor sensitivity. The IQ-achievement discrepancy yielded low prevalence estimates, strong specificity, and poor sensitivity. Of the RTI methods, final normalization yielded reasonable prevalence estimates, marginal specificity, and marginal sensitivity. Both slope discrepancy and dual-discrepancy methods yielded high prevalence estimates, adequate sensitivity, and adequate specificity. Benchmark and median-split methods yielded high prevalence estimates and marginal sensitivity and specificity (both below .80).

Although these various methods differed in terms of their overall prevalence and predictive utility, many of the RTI methods have as a core assumption that measuring growth on progress monitoring measures yields information beyond what can be obtained with a single point assessment. Three of these methods (the median-split on slopes, slope discrepancy, and dual-discrepancy) rely on the idea that measuring growth in these skills is critical to the identification process. Inherent in the notion of RTI is the idea that frequent assessments would identify children as being responsive or nonresponsive to increasingly intensive instruction based upon how quickly or slowly they grow on these assessments. The present study attempts to investigate whether growth does add additional information to predicting future reading performance.

4. The present study

One key difference among the approaches is whether non-responders are identified on the basis of low achievement status (i.e., Good et al., 2001a,b,c; Torgesen et al., 2001; Vaughn, Linan-Thompson, Hickman, 2005), low growth (Fuchs et al., 2004; Speece & Case, 2001; Vellutino et al., 1996), or both (Fuchs et al., 2004). The purpose of the present study was to compare the predictive validity of measures of achievement status, growth, and both achievement status and growth, in a two-year, large-scale longitudinal study beginning in first grade. The participants were students who are potentially at a higher-than-average level of risk for reading difficulties by virtue of their attending Reading First schools. Reading instruction was provided using an approved reading program with support from reading coaches, and could be considered representative of a tier one level of service in a three-tier RTI program.

4.1. Method

4.1.1. Participants—The participants were first grade children from districts throughout the state of Florida who were attending Reading First schools. Reading First is the largest federal initiative in the history of the United States to improve the reading performance of poor readers. The initial sample consisted of 32,016 first grade students from across the state who were in first grade in the fall of 2003 during the first year of Florida's Reading First Initiative. During the summer, K-3 teachers from the Reading First Schools attended a four day workshop that focused on core Tier 1 (whole classroom) instruction and how to use the results from the DIBELS assessments to group struggling readers into effective Tier 2 interventions. Additionally, reading coaches were assigned to each school to assist teachers in implementing the Reading First plan. From this initial sample of students, we retained only those students for whom scores were available on all four oral reading fluency

assessments during their first grade school year, and for whom end of second grade reading achievement scores were available. This reduced our final sample to 23,438 students. Forty-eight percent of the participants were female, 72.3% were eligible for free or reduced lunch, and the ethnic breakdown was as follows: 41.1% White, 31.8% Black, 21.5% Hispanic, 4.0% mixed race, 1.3% Asian, and .3 Native American.

4.1.2. Measures—The reading measures assessed both oral reading fluency and reading comprehension of passages read silently.

4.1.3. DIBELS oral reading fluency (ORF)—ORF (5th Edition; Good, Kaminski, Smith, Laimon, & Dill, 2001) assesses oral reading fluency for reading grade-level connected text. This test is individually administered and assesses the number of words accurately read in 1 min. The ORF was administered four times during the months of September, December, February, and April.

Performance was measured at each time point by having students read three passages aloud for one minute each, with the cue to “be sure to do your best reading.” Words omitted, substituted, and hesitations of more than three seconds were scored as errors. Words self-corrected within three seconds were scored as accurate. The assessor noted errors, and the number of correct words read per minute was used as the score. The median of the scores from the three test passages was used as the final data point for each assessment period. Median alternate-form reliability for oral reading of passages has been shown to be stable at .94 (Good, Kaminski, Smith, & Bratten, 2001).

4.1.4. Stanford Achievement Test (10th edition; SAT-10)—SAT-10 (Harcourt Brace, 2003) is a standardized and nationally-normed measure of reading comprehension. Classroom teachers administered this untimed test in a group format. Students answered multiple choice items that assess initial understanding, interpretation, critical analysis, and awareness and usage of reading strategies after reading literary, informational, and functional text passages. The reliability coefficient for SAT-10 reading comprehension was .91 for a nationally representative sample of students assessed at the end of second grade. Considerable evidence of content, criterion-related, and construct validity is available for the test (Harcourt Brace, 2003). Although we expect that the teachers administered the SAT-10 according to state-wide protocol, we have no direct observational data that corroborates this, which could be seen as a limitation of this assessment results.

4.1.5. Procedures—The data used in this study were obtained using Florida's Progress Monitoring and Reporting Network (PMRN) maintained by the Florida Center for Reading Research. The PMRN is a centralized data collection and reporting system through which Reading First schools in Florida both report reading data and receive reports of the data for instructional decision making. Trained assessors (not the classroom teachers) collected the progress monitoring data, including the DIBELS ORF assessments, four times a year and entered the data into the PMRN's web-based data entry facility. The SAT 10 assessments, which are administered by the schools, were obtained directly from Harcourt Brace and are match-merged into the PMRN.

4.2. Results

4.2.1. Overview of growth curve analyses—The data analyses were designed to compare the predictive validity of estimates of (a) student growth in oral reading fluency, (b) student status or level of oral reading fluency, and (c) combined measures of growth and status, for prediction of concurrent and future reading skills. To obtain estimates of student growth in oral reading fluency, we fit two individual growth curve models (Raudenbush &

Bryk, 2002) to the four oral reading fluency time points. The first model was a linear growth model that assumed that each student's growth over the first grade period could be adequately modeled as a straight line. Although this model may be unrealistic, it does have the useful property that estimates of linear growth often have higher reliabilities than growth estimates obtained from more complex models. The second model fit was a model that allowed each person's growth over the year to be modeled by a quadratic function. A quadratic model allows for each person's growth to accelerate or decelerate and may be a more realistic model for students in first grade for whom more rapid growth may occur in reading skill toward the end of the year. Both models centered time at the last assessment period.

From these models, ordinary least squares estimates of growth can be obtained. For the linear model, the slope estimate represents the amount of linear growth per assessment period (roughly two months) and remains constant through the year. For the quadratic model, rate of growth can change throughout the year. When the quadratic model is centered at the last assessment period, the linear slope estimate from this model represents the estimated rate of growth at the last assessment period, while the quadratic estimate represents instantaneous change in the slope throughout the year. Estimates of growth from both models were obtained and used in a multiple regression to predict reading comprehension scores. From the linear model, slopes that represent the average growth observed over the year were used, while from the quadratic model, the growth rate estimated at the final measurement point was used.

4.2.2. Descriptive statistics—Means and standard deviations are reported in Table 1. For the overall sample, performance on the ORF assessments indicate that growth in this skill is accelerating over the school year, with approximately 6 words per minute growth from September to December, and approximately 15 words per minute change from February to April. The scores on the SAT 10 Reading comprehension indicate that the sample was very close to the nationally-normed sample, with percentile scores close to 50 for the end of first grade and the end of second grade.

4.2.3. Individual growth curves—As described above, two growth curve models were fit to the oral reading fluency data. The first model of linear growth allowed both the intercept and the slopes to be random. The second model added a random quadratic term to the model. Estimates of the fixed and random effects from both models are presented in Table 2.

For the linear model, the intercept of 47.7 is the estimated end of first grade ORF performance and this estimate is close to the observed April mean of 49.6. The reliability of the intercept (.97) was quite high. The estimate of the slope of 6.1 words per minute represents the amount of improvement in oral reading fluency between each two month interval. The slope estimate was moderately reliable (.81). The significant random effects for intercept (993.8) and for slope (11.1) indicate that there were reliable individual differences in these parameters.

For the quadratic model, the presence of a significant fixed and random effect for the quadratic term indicated that the quadratic model provided a better fit to the data than the linear model, and that there were reliable individual differences in the quadratic term as well as in the intercept and slope. The slope estimate from the quadratic model (9.3) exceeded that of the linear model (6.1) because the slope estimate for the quadratic model was at the final measurement point and the slope increased from the beginning to the end of the year, whereas the slope from the linear model was an estimate of growth across the entire year. The positive quadratic term (0.54) indicates that the slope was accelerating at the end of first

grade. Although the reliability of the intercept remained comparable to that of the linear model (.97), the reliabilities of the linear (.57) and quadratic (.25) terms were substantially less than that of the growth parameter of the linear model (.81).

Because the growth estimate from the linear model was more reliable than the linear growth estimate from the quadratic model, but the quadratic model appeared to provide at least a modestly better fit to the data, we decided to investigate the unique contributions of growth using estimates from both the linear and quadratic models. Ordinary least squares estimates of growth from both models were obtained for every student and were used to predict concurrent and future reading performance.

4.2.4. Multiple regression analyses—The three dependent variables that served as reading criteria to be predicted were SAT 10 Reading Comprehension at the end of first grade, SAT 10 Reading Comprehension at the end of second grade, and ORF measured at the beginning of second grade. For each dependent variable, two hierarchical multiple regression analyses were carried out to determine relative contributions of slope and end of year status in ORF. The first hierarchical multiple regression had slope as the sole predictor in the first step and added end of year status in the second step. The second hierarchical multiple regression reversed the order of entry of the two predictors.

Correlations among the predictors and dependent variables used in the multiple regression analyses are presented in Table 3. For all three dependent variables, correlations were highest with end of year status in ORF, next highest for linear slope, and lowest for slope at the end of first grade assessment.

The results of multiple regression analyses for parameter estimates from the linear model are presented in Tables 4. The pattern of results was identical across the three dependent variables. End of year status in ORF made a large contribution to prediction that was independent of the contribution to prediction of slope, whereas slope made little or no independent contribution to prediction independent of status. Results from the quadratic model are presented in Table 5. The pattern of results was identical to that of the linear model, with a large independent contribution to prediction made by end of year status but not by slope.

5. Discussion

The purpose of the present study was to estimate the relative contributions of measures of growth and status of oral reading fluency for predicting concurrent and future reading skills. Several of the proposed reading disability classification models grounded within RTI use growth on curriculum based assessments as a primary component of their definition (Fuchs & Fuchs, 1998; Speece & Case, 2001; Vellutino et al, 1996). The results of this study seem to indicate that, for first grade students, growth is not adding unique information to prediction of future reading skills above and beyond what can be obtained using a single time point assessment at the end of the year. Furthermore, this result was obtained on a large and representative group of students who attend schools that have a higher-than-normal risk for having children with reading difficulties.

There are a number of reasons why growth may not add to our prediction of future reading performance. First, a sizable amount of information about growth is carried in the final assessment. For the vast majority of students, there is only one way to obtain a low ORF score at the end of the year — and that is to grow slowly, or not at all. Second, as is typically the case, the reliability of the end of year status estimates exceed that of the slope estimates

by a considerable margin, and this difference in reliability may have been responsible for the observed differences in correlations and regression results that were obtained.

More generally, there may be reason to be skeptical that RTI models will address the concerns raised about the traditional IQ-achievement discrepancy model until convincing data prove otherwise. Recall that traditional IQ-achievement models were called “wait to fail” models because students were not identified routinely earlier than second grade. This criticism would seem to apply equally to RTI models as they (a) are most likely not to be implemented before first grade, (b) take a substantial amount of time to measure a child's response first to tier one effective classroom instruction and then tier two supplemental intervention, and (c) require failure in the form of failing to respond to instruction and intervention before identification of a reading disability is achieved.

A second criticism of IQ-achievement discrepancy scores was that they were less reliable than well constructed measures of IQ or achievement. Given this criticism and the fact that RTI models have been proposed as a better alternative, it is disconcerting that there appear to be no published studies that assess the reliability of an RTI approach. Fletcher et al. (2007) suggest one reason why the reliability of an RTI approach might be better than that of identification based on a single assessment. The RTI approach uses multiple measures collected over an extensive time period, and performance across multiple measures is assumed to be more reliable than performance assessed once. However, although the reliability with which change is measured increases as a function of the number of measurements obtained, measures of change tend to be noticeably less reliable than are measures of status. This was the case for the present study.

A reason to be even more skeptical about the reliability of identification based on an RTI approach is uncontrolled variability associated with teacher quality and differential effectiveness of instruction and intervention. Any such variability will contribute to unreliability in identification.

Unreliability associated with applying a cut-point to a continuous distribution will apply to the RTI approach as much as it does the traditional IQ-achievement discrepancy approach, in that measures of response to intervention are likely to have continuous distributions and a cut-point will need to be used to determine which children advance through the tiers of an RTI model.

Finally, measures that are suitable for progress monitoring do not yet match traditional measures of aptitude and achievement in desirable psychometric characteristics. Development of progress monitoring measures represents an even greater challenge because creation of multiple, equivalent forms is required.

Although our results did not support the value of measures of growth relative to status in RTI models, several limitations of our study should be noted. First, our measures of growth all were obtained during first grade. It is possible that the results could have been different had we studied growth in other grades. Second, we studied growth in a single measure, namely, oral reading fluency. Whether the results generalize to other measures of reading remains to be seen. Third, our study was carried out in the context of a tier one level of a three-tier RTI model. Although effective classroom instruction was presented, we did not implement intensive tutoring for students who were lagging behind in growth and we predicted the performance of students who varied considerably in reading performance as opposed to focusing solely on children who had severe reading disability. Fourth and finally, our measures of growth were based on four assessments distributed throughout the year. It is possible to assess more frequently, and doing so might have improved the reliability of the slope parameter. The greater reliability of the status parameter relative to the slope

parameter in the present study may have affected the outcome in favor of the status parameter.

We are not advocating a return to an IQ-achievement discrepancy scores as the primary means to identify individuals with reading disability, and we are intrigued by the possibility of overcoming the wait to fail nature of both traditional and RTI approaches by replacing reading achievement with print awareness and comparing the performance of both approaches at the preschool level. The main implication of our results is that they reinforce the need to be skeptical of the myth that RTI models provide realistic solutions to the problems associated with the traditional IQ-discrepancy model, until convincing evidence is available to the contrary.

References

- Burgess SR, Lonigan CJ. Bidirectional relations of phonological sensitivity and prereading abilities: Evidence from a preschool sample. *Journal of Experimental Child Psychology*. 1998; 70:117–141. [PubMed: 9729452]
- Case LP, Speece DL, Molloy DE. The validity of a response-to-instruction paradigm to identify reading disabilities: A longitudinal analysis of individual difference and contextual factors. *School Psychology Review*. 2003; 32:557–582.
- Compton DL, Fuchs D, Fuchs LS, Bryant JD. 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.
- Fletcher JM, Francis DJ, Rourke BP, Shaywitz SE, Shaywitz BA. The validity of discrepancy-based definitions of reading disabilities. *Journal of Learning Disabilities*. 1992; 25:555–561. 573. [PubMed: 1431539]
- Fletcher, JM.; Lyon, GR.; Fuchs, LS.; Barnes, MA. *Learning disabilities: From identification to intervention*. New York, NY: Guilford; 2007.
- Fletcher, JM.; Morris, RD.; Lyon, GR. Classification and definition of learning disabilities: An integrative perspective. In: Swanson, HL.; Harris, KR., editors. *Handbook of learning disabilities*. New York: Guilford Press; 2003. p. 30-56.
- 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; 38:98–110. [PubMed: 15813593]
- Francis DJ, Shaywitz SE, Stuebing KK, Shaywitz BA, Fletcher JM. Developmental lag versus deficit models of reading disability: A longitudinal individual growth curves analysis. *Journal of Educational Psychology*. 1996; 1:3–17.
- Fuchs D, Compton DL, Fuchs LS, Bryant J, Davis NG. Making “secondary intervention” work in a three-tier responsiveness-to-intervention model: Findings from the first-grade longitudinal study of the National Research Center on Learning Disabilities. *Reading and Writing Quarterly: An Interdisciplinary Journal*. (in press).
- 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. Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice*. 1998; 13:204–219.
- Fuchs D, Fuchs LS, Compton DL. Identifying reading disability by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disabilities Quarterly*. 2004; 27:216–227.
- Fuchs, D.; Fuchs, LS.; McMaster, KN.; Al Otaiba, S. Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In: Swanson, HL.; Harris, KR., editors. *Handbook of learning disabilities*. New York, NY: Guilford Press; 2003. p. 431-449.
- Fuchs D, Mock D, Morgan PL, Young CL. Responsiveness-to-intervention: Definitions, evidence, and implications for the learning disabilities construct. *Learning Disabilities Research & Practice*. 2003; 18:157–171.

- Good, RH.; Kaminski, RA.; Smith, S.; Bratten, J. Technical adequacy of second grade DIBELS oral reading fluency passages (Technical Report No. 8). Eugene, OR: University of Oregon; 2001.
- Good, RH.; Kaminski, RA.; Smith, S.; Laimon, D.; Dill, S. Dynamic Indicators of Basic Early Literacy Skills. 5th ed.. Eugene, OR: University of Oregon; 2001.
- Good RH, Simmons DC, Kame'enui EJ. The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading*. 2001; 5:257–288.
- Grigorenko EL. Dynamic assessment and response to intervention: Two sides of the same coin. *Journal of Learning Disabilities*. (in press).
- Harcourt Brace. Stanford Achievement Test, Tenth Edition: Technical Data Report. San Antonio, TX: Author; 2003.
- Kirk, SA. Educating exceptional children. Boston: Houghton Mifflin; 1962.
- Lonigan, CJ.; Wagner, R.; Torgesen, J.; Rashotte, C. Test of Preschool Early Literacy. Austin, TX: Pro-Ed; 2007.
- Lyon, GR.; Fletcher, JM.; Shaywitz, SE.; Shaywitz, BA.; Torgesen, JK.; Wood, FB. Rethinking learning disabilities. In: Finn, CE.; Rotherham, AJ.; Hokanson, CR., Jr., editors. Rethinking special education for a new century. 2001. p. 259-287.
- Lyon GR, Shaywitz SE, Shaywitz BA. A definition of dyslexia. *Annals of Dyslexia*. 2003; 53:1–14.
- 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.
- President's Commission on Excellence in Special Education. A new era: Revitalizing special education for children and their families. Washington, DC: U.S. Department of Education; 2002.
- Raudenbush SW, Bryk AS. Hierarchical linear models: Applications and data analysis methods (2nd ed.). 2002
- Reschly DJ. Learning disabilities identification: Primary intervention, secondary intervention, and then what? *Journal of Learning Disabilities*. 2005; 38:510–515. [PubMed: 16392692]
- Shaywitz BA, Fletcher JM, Holahan JM, Shaywitz SE. Discrepancy compared to low achievement definitions of reading disability: Results from the Connecticut longitudinal study. *Journal of Learning Disabilities*. 1992; 25:639–648. [PubMed: 1460385]
- Shaywitz, SE.; Shaywitz, BA. Neurobiological indices of dyslexia. In: Swanson, HL.; Harris, KR., editors. Handbook of learning disabilities. New York: Guilford Press; 2003. p. 514-531.
- Siegel LS. An evaluation of the discrepancy definition of dyslexia. *Journal of Learning Disabilities*. 1992; 25:618–629. [PubMed: 1460383]
- Siegel, LS. Basic cognitive processes and reading disabilities. In: Swanson, HL.; Harris, KR., editors. Handbook of learning disabilities. New York: Guilford Press; 2003. p. 158-181.
- Spear-Swerling, L.; Sternberg, R. Off track: When poor readers become learning disabled. Boulder, Colorado: Westview Press; 1996.
- Speece DL, Case LP. Classification in context: An alternative approach to identifying early reading disability. *Journal of Educational Psychology*. 2001; 93:735–749.
- Stanovich KE. Discrepancy definitions of reading disability: Has intelligence led us astray? *Reading Research Quarterly*. 1991; 26:7–29.
- Stanovich KE, Siegel LS. Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological–core variable–difference model. *Journal of Educational Psychology*. 1994; 86:24–53.
- Stuebing KK, Fletcher JM, LeDoux JM, Lyon GR, Shaywitz SE, Shaywitz BA. Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal*. 2002; 39:469–518.
- Torgesen JK. Individual responses in response to early intervention in reading: the lingering problem of treatment resisters. *Learning Disabilities Research and Practice*. 2000; 15:55–64.
- Torgesen JK, Alexander AW, Wagner RK, Rashotte CA, Voeller K, Conway T, Rose E. 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]

- Vaughn, S.; Chard, D. Three tier intervention research studies: Descriptions of two related projects. *Perspectives*, Winter; 2006. p. 29-34.
- Vaughn S, Fuchs LS. Redefining learning disabilities as inadequate response to instruction: The promise and potential problems. *Learning Disabilities Research and Practice*. 2003; 28:137–146.
- Vaughn S, Linan-Thompson S, Hickman P. Response to instruction as a means of identifying students with reading/learning disabilities. *Exceptional Children*. 2005; 69(4):391–410.
- Vellutino FR, Scanlon DM, Sipay ER, Small S, Chen R, Pratt A, Denckla MB. Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology*. 1996; 88:601–638.
- Wagner, RK. Rediscovering Dyslexia: New Approaches for Identification and Classification. In: Reid, G.; Fawcett, A.; Manis, F.; Siegel, L., editors. *The handbook of dyslexia*. Sage Publications; (in press)
- Wagner RK, Torgesen JK, Rashotte CA, Hecht S, Barker T, Burgess S, Garon T. Causal relations between the development of phonological processing and reading: A five-year longitudinal study. *Developmental Psychology*. 1997; 33:468–479. [PubMed: 9149925]

Table 1Means and standard deviations for the ORF and SAT-10 assessments ($N = 23,438$)

Assessments	Mean	Std	Min	Max
September ORF raw score	13.68	18.91	0	187
December ORF raw score	20.33	22.40	0	197
February ORF raw score	34.27	27.75	0	293
April ORF raw score	49.55	32.26	0	294
September ORF raw score — second grade	57.31	32.63	0	220
SAT 10 reading comprehension percentile score — spring first grade	49.62	27.79	1	97
SAT 10 reading comprehension percentile score — spring second grade	48.52	26.18	0	99

Table 2

Estimates of growth and overall level of performance on ORF

	Linear growth		Quadratic growth	
	Model		Model	
	Estimate	Reliability	Estimate	Reliability
Fixed effects				
Intercept	47.7	0.97	49.8	0.97
Slope	6.1	0.81	9.3	0.57
Quadratic			0.54	0.25
Random effects				
Intercept	993.8		1018.9	
Slope	11.1		29.26	
Quadratic			0.19	
Residual	50.8		35.5	

Note: Estimates for random effects are variances; all fixed and random estimates are significant at $p < .0001$.

Table 3Correlations of the assessment and growth variables ($N = 23,438$)

Assessments	1	2	3	4	5	6
1. April first grade ORF raw score	1.00					
2. Linear growth across the year	.82	1.00				
3. Linear growth at April assessment	.54	.72	1.00			
4. September second grade ORF raw score	.93	.76	.46	1.00		
5. SAT 10 reading comprehension — spring first grade	.79	.69	.43	.79	1.00	
6. SAT 10 reading comprehension — spring second grade	.61	.52	.33	.64	.67	1.00

Note: All correlations are significant at $p < .0001$.

Table 4

Hierarchical regression analyses using parameters estimated from a linear model

<i>Predicting second grade oral reading fluency</i>		
Step	R^2	ΔR^2
A1. ORF slope as sole predictor.	.580***	—
A2. End of year ORF added as second predictor.	.858***	.278***
B1. End of year ORF as sole predictor.	.858***	—
B2. ORF slope added as second predictor.	.858***	.000
<i>Predicting end of first grade SAT 10 reading comprehension</i>		
Step	R^2	ΔR^2
A1. ORF slope as sole predictor.	.472***	—
A2. End of year ORF added as second predictor.	.630***	.158***
B1. End of year ORF as sole predictor.	.626***	—
B2. ORF slope added as second predictor.	.630***	.004***
<i>Predicting end of second grade SAT 10 reading comprehension</i>		
Step	R^2	ΔR^2
A1. ORF slope as sole predictor.	.275***	—
A2. End of year ORF added as second predictor.	.373***	.098***
B1. End of year ORF as sole predictor.	.371***	—
B2. ORF Slope added as second predictor.	.373***	.004***

Table 5

Hierarchical regression analyses using parameters estimated from a quadratic model

<i>Predicting second grade oral reading fluency</i>		
Step	R^2	ΔR^2
A1. ORF slope as sole predictor.	.211***	—
A2. End of year ORF added as second predictor.	.861***	.650***
B1. End of year ORF as sole predictor.	.858***	—
B2. ORF slope added as second predictor.	.861***	.003***
<i>Predicting end of first grade SAT 10 reading comprehension</i>		
Step	R^2	ΔR^2
A1. ORF slope as sole predictor.	.181***	—
A2. End of year ORF added as second predictor.	.626***	.445***
B1. End of year ORF as sole predictor.	.626***	—
B2. ORF slope added as second predictor.	.626***	.000
<i>Predicting end of second grade SAT 10 reading comprehension</i>		
Step	R^2	ΔR^2
A1. ORF slope as sole predictor.	.111***	—
A2. ORF end of year status added as second predictor.	.371***	.260***
B1. End of year ORF as sole predictor.	.371***	—
B2. ORF slope added as second predictor.	.373***	.000