Evidence-Based Assessment of Learning Disabilities in Children and Adolescents

Jack M. Fletcher

*Department of Pediatrics and the Center for Academic and Reading Skills,
University of Texas Health Science Center at Houston*

David J. Francis

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

Robin D. Morris

Department of Psychology, Georgia State University

G. Reid Lyon

Child Development and Behavior Branch, National Institute of Child Health and Human Development

The reliability and validity of 4 approaches to the assessment of children and adolescents with learning disabilities (LD) are reviewed, including models based on (a) aptitude–achievement discrepancies, (b) low achievement, (c) intra-individual differences, and (d) response to intervention (RTI). We identify serious psychometric problems that affect the reliability of models based on aptitude–achievement discrepancies and low achievement. There are also significant validity problems for models based on aptitude–achievement discrepancies and intra-individual differences. Models that incorporate RTI have considerable potential for addressing both the reliability and validity issues but cannot represent the sole criterion for LD identification. We suggest that models incorporating both low achievement and RTI concepts have the strongest evidence base and the most direct relation to treatment. The assessment of children for LD must reflect a stronger underlying classification that takes into account relations with other childhood disorders as well as the reliability and validity of the underlying classification and resultant assessment and identification system. The implications of this type of model for clinical assessments of children for whom LD is a concern are discussed.

Assessment methods for identifying children and adolescents with learning disabilities (LD) are multiple, varied, and the subject of heated debates among practitioners. Those debates involve issues that extend beyond the value of specific tests, often reflecting different views of how LD is best identified. These views reflect variations in the definition of LD and, therefore, variations in what measures are selected to operationalize the definition (Fletcher, Foorman, et al., 2002). Any focus on the "best tests" leads to a hopeless morass of confusion in an area such as LD that has not successfully addressed the classification and definition issues that lead to identification of who does and who does not possesses characteristics of LD. Definitions always reflect an implicit classification indicating how different constructs are measured and used to identify members of the class in terms of similarities and differences relative to other entities that are not considered members of the class (Morris & Fletcher, 1988). For LD, children who are members of this class are historically differentiated from children who have other achievement-related difficulties, such as mental retardation, sensory disorders, emotional or behavioral disturbances, and environmental causes of underachievement, including economic disadvantage, minority language status, and inadequate instruction (Fletcher, Francis, Rourke, Shaywitz, & Shaywitz, 1993; Lyon, Fletcher, & Barnes, 2003). If the classification is valid, children with LD may share characteristics that are similar with other groups of underachievers, but they

should also differ in ways that can be measured and that can serve to define and operationalize the class of children and adolescents with LD.

In this article, we consider evidence-based approaches to the assessment of LD in the context of different approaches to the classification and identification of LD. We argue that the measurement systems that are used to identify children and adolescents with LD are inseparable from the classifications from which the identification criteria evolve. Moreover, all measurement systems are imperfect attempts to measure a construct (LD) that operates as a latent variable that is unknowable independently of how it is measured and therefore of how LD is classified. The construct of LD is imperfectly measured simply because the measurement tools themselves are not error free (Francis et al., 2005). Different approaches to classification and definition capitalize on this error of measurement in ways that reduce or increase the reliability of the classification itself. Similarly, evaluating similarities and differences among groups of students who are identified as LD and not LD is a test of the validity of the underlying classification, so long as the variables used to assess this form of validity are not the same as those used for identification (Morris & Fletcher, 1988). As with any form of validity, adequate reliability is essential. Classifications can be reliable and still lack validity. The converse is not true; they cannot be valid and lack reliability. A valid classification of LD predicts important characteristics of the group. Consistent with the spirit of this special section, the most important characteristic is whether the classification is meaningfully related to intervention. For LD, a classification should also predict a variety of differences on cognitive skills, behavioral attributes, and achievement variables not used to form the classification, developmental course, response to intervention (RTI), neurobiological variables, or prognosis (Fletcher, Lyon, et al., 2002).

To address these issues, we consider the reliability and validity of four approaches to the classification and assessment of LD: (a) IQ discrepancy and other forms of aptitude–achievement discrepancy, (b) low achievement, (c) intra-individual differences, and (d) models incorporating RTI and some form of curriculum-based measurement. We consider how each classification reflects the historically prominent concept of "unexpected underachievement" as the key construct in LD assessment (Lyon et al., 2001), that is, what many early observers characterized as a group of children unable to master academic skills despite the absence of known causes of poor achievement (sensory disorder, mental retardation, emotional disturbances, economic disadvantages, inadequate instruction). From this perspective, a valid classification and measurement system for LD must identify a unique group of underachievers that is clearly differentiated from groups with other forms of underachievement.

Defining LD

Historically, definition and classification issues have haunted the field of LD. As reviewed in Lyon et al. (2001), most early conceptualizations viewed LD simply as a form of "unexpected" underachievement. The primary approach to assessment involved the identification of intra-individual variability as a marker for the unexpectedness of LD, along with the exclusion of other causes of underachievement that would be expected to produce underachievement. This type of definition was explicitly coded into U.S. federal statutes when LD was identified as an eligibility category for special education in Public Law 94–142 in 1975; essentially the same definition is part of current U.S. federal statues in the Individuals with Disabilities Education Act (1997).

The U.S. statutory definition of LD is essentially a set of concepts that in itself is difficult to operationalize. In 1977, recommendations for operationalizing the federal definition of LD were provided to states after passage of Public Law 94–142 to help identify children in this category of special education (U. S. Office of Education, 1977). In these regulations, LD was defined as a heterogeneous group of seven disorders (oral language, listening comprehension, basic reading, reading comprehension, math calculations, math reasoning, written language) with a common marker of intra-individual variability represented by a discrepancy between IQ and achievement (i.e., unexpected underachievement). Unexpectedness was also indicated by maintaining the exclusionary criteria present in the statutory definition that presumably lead to expected underachievement. Other parts of the regulations emphasize the need to ensure that the child's educational program provided adequate opportunity to learn. No recommendations were made concerning the assessment of psychological processes, most likely because it was not clear that reliable methods existed for assessing processing skills and because the field was not clear on what processes should be assessed (Reschly, Hosp, & Smied, 2003).

This approach to definition is now widely implemented with substantial variability across schools, districts, and states in which students are served in special education as LD (MacMillan & Siperstein, 2002; Mercer, Jordan, Allsop, & Mercer, 1996; Reschly et al., 2003). It is also the basis for assessments of LD outside of schools. Consider, for example, the definition of reading disorders in the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994), which indicates that the student must perform below levels expected for age and IQ, and specifies only sensory disorders as exclusionary:

A. Reading achievement, as measured by individually administered standardized tests of read-

ing accuracy or comprehension, is substantially below that expected given the person's chronological age, measured intelligence, and age-appropriate education.

B. The disturbance in Criterion A significantly interferes with academic achievement or activities of daily living that require reading skills.

C. If a sensory deficit is present, the reading difficulties are in excess of those usually associated with it.

The International Classification of Diseases–10 has a similar definition. It differs largely in being more specific in requiring use of a regression-adjusted discrepancy, specifying cut points (achievement two standard errors below IQ) for identifying a child with LD, and expanding the range of exclusions.

Although these definitions are used in what are often disparate realms of practice, they lead to similar approaches to the identification of children and adolescents as LD. Across these realms, children commonly receive IQ and achievement tests. The IQ test is commonly interpreted as an aptitude measure or index against which achievement is compared. Different achievement tests are used because LD may affect achievement in reading, math, or written language. The heterogeneity is recognized explicitly in the U.S. statutory and regulatory definitions of LD (Individuals With Disabilities Education Act, 1997) and in the psychiatric classifications by the provision of separate definitions for each academic domain. However, it is still essentially the same definition applied in different domains. In many settings, this basic assessment is supplemented with tests of processing skills derived from multiple perspectives (neuropsychology, information processing, and theories of LD). The approach boils down to administration of a battery of tests to identify LD, presumably with treatment implications.

Underlying Classification Hypotheses

Implicit in all these definitions are slight variations on a classification model of individuals with LD as those who show a measurable discrepancy in some but not all domains of skill development and who are not identified into another subgroup of poor achievers. In some instances, the discrepancy is quantified with two tests in an aptitude–achievement model epitomized by the IQ-discrepancy approach in the U.S. federal regulatory definition and the psychiatric classifications of the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; American Psychiatric Association, 1994) and the International Classification of Diseases–10. Here the classification model implicitly stipulates that those who meet an IQ-discrepancy inclusionary criterion are different in meaningful ways from those who are underachievers and do not meet the

discrepancy criteria or criteria for one of the exclusionary conditions. Some have argued that this model lacks validity and propose that LD is synonymous with underachievement, so that it should be identified solely by achievement tests (Siegel, 1992), often with some exclusionary criteria to help ensure that the achievement problem is unexpected. Thus, the contrast is really between a two-test aptitude–achievement discrepancy and a one-test chronological age-achievement discrepancy with achievement low relative to age-based (or grade-based) expectations. If processing measures are added, the model becomes a multitest discrepancy model. Identification of a child as LD in all three of these models is typically based on assessment at a single point in time, so we refer to them as "status" models. Finally, RTI models emphasize the "adequate opportunity to learn" exclusionary criterion by assessing the child's response to different instructional efforts over time with frequent brief assessments, that is, a "change" model. The child who is LD becomes one who demonstrates intractability in learning characteristics by not responding adequately to instruction that is effective with most other students.

Dimensional Nature of LD

Each of these four models can be evaluated for reliability and validity. Unexpected underachievement, a concept critically important to the validity of the underlying construct of LD, can also be examined. The reliability issues are similar across the first three models and stem from the dimensional nature of LD. Most population-based studies have shown that reading and math skills are normally distributed (Jorm, Share, Matthews, & Matthews, 1986; Lewis, Hitch, & Walker, 1994; Rodgers, 1983; Shalev, Auerbach, Manor, & Gross-Tsur, 2000; Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992; Silva, McGee, & Williams, 1985). These findings are buttressed by behavioral genetic studies, which are not consistent with the presence of qualitatively different characteristics associated with the heritability of reading and math disorders (Fisher & DeFries, 2002; Gilger, 2002). As dimensional traits that exist on a continuum, there would be no expectation of natural cut points that differentiate individuals with LD from those who are underachievers but not identified as LD (Shaywitz et al., 1992).

The unobservable nature of LD makes two-test and one-test discrepancy models unreliable in ways that are psychometrically predictable but not in ways that simply equate LD with poor achievement (Francis et al., 2005; Stuebing et al., 2002). The problem is that the measurement approach is based on a static assessment model that possesses insufficient information about the underlying construct to allow for reliable classifications of individuals along what is essentially an unobservable dimension. If LD was a

manifest concept that was directly observable in the behavior of affected individuals, or if there were natural discontinuities that represented a qualitative breakpoint in the distribution of achievement skills or the cognitive skills on which achievement depends, this problem would be less of an obstacle. However, like achievement or intelligence, LD is a latent construct that must be inferred from the pattern of performance on directly observable operationalizations of other latent constructs (namely, test scores that index constructs like reading achievement, phonological awareness, aptitude, and so on). The more information available to support the inference of LD, the more reliable (and valid) that inference becomes, thus supporting the fine-grained distinctions necessitated by two-test and one-test discrepancy models. To the extent that the latent construct, LD, is categorical, by which we mean that the construct indexes different classes of learners (i.e., children who learn differently) as opposed to simply different levels of achievement, then systems of identification that rely on one measurable variable lack sufficient information to identify the latent classes and assign individuals to those classes without placing additional, untestable, and unsupportable constraints on the system. It is simply not possible to use a single mean and standard deviation and to estimate separate means and standard deviations for two (or more) unobservable latent classes of individuals and determine the percentage of individuals falling into each class, let alone to classify specific individuals into those classes. Without constraints, such as specifying the magnitude of differences in the means of the latent classes, the ratio of standard deviations, and the odds of membership in the two (or more) classes, the system is under-identified, which simply means that there are many different solutions that cannot be distinguished from one another.

When the system is under-identified, the only solution is to expand the measurement system to increase the number of observed relations, which in one sense is what intra-individual difference models attempt by adding assessments of processing skills. Other criteria are necessary because it is impossible to uniquely identify a distinct subgroup of underachieving individuals consistent with the construct of LD when identification is based on a single assessment at a single time point. Adding external criteria, such as an aptitude measure or multiple assessments of processing skills, increases the dimensionality of the measurement system and makes latent classification more feasible, even when the other criteria are themselves imperfect. But the main issues for one-test, two-test, and multitest identification models involve the reliability of the underlying classifications and whether they identify a unique subgroup of underachievers. In the next section, we examine variations in reliability and validity for each of these models, fo-

cusing on the importance of reliability, as the validity of the classifications can be no stronger than their reliability.

Models Based on Two-Test Discrepancies

Although the IQ-discrepancy model is the most widely utilized approach to identifying LD, there are many different ways to operationalize the model. For example, some implementations are based on a composite IQ score, whereas others utilize either a verbal or nonverbal IQ score. Other approaches drop IQ as the aptitude measure and use a measure such as listening comprehension. In the validity section, we discuss each of these approaches. The reliability issues are similar for each example of an aptitude–achievement discrepancy.

Reliability

Specific reliability problems for two-test discrepancy models pertain to any comparison of two correlated assessments that involve the determination of a child's performance relative to a cut point on a continuous distribution. Discrepancy involves the calculation of a difference score (D) to estimate the true difference (Δ) between two latent constructs. Thus, discussions about discrepancy must distinguish between problems with the manifest (i.e., observed) difference (D) as an index of the true difference (Δ) but also must consider whether the true difference (Δ) reflects the construct of interest. Problems with the reliability of D based on differences between two tests are well known, albeit not in the LD context (Bereiter, 1967). However, there is nothing that fundamentally limits the applicability of this research to LD if we are willing to accept a notion of Δ as a marker for LD. There are major problems with this assumption that are reviewed in Francis et al. (2005). The most significant is regression to the mean. On average, regression to the mean indicates that scores that are above the mean will be lower when the test is repeated or when a second correlated test is used to compute D. In this example, individuals who have IQ scores above the mean will obtain achievement test scores that, on average, will be lower than the IQ test score because the achievement score will move toward the mean. The opposite is true for individuals with IQ scores below the mean. This leads to the paradox of children with achievement scores that exceed IQ, or the identification of low-achieving, higher IQ children with achievement above the average range as LD.

Although adjusting for the correlation of IQ and achievement helps correct for regression effects (Reynolds, 1984–1985), unreliability also stems from the attempt to assess a person's standing relative to a cut point on a continuous distribution. As discussed in the

following section on low achievement models, this problem makes identification with a single test—even one with small amounts of measurement error—potentially unreliable, a problem for any status model.

None of this discussion addresses the validity question concerning Δ. Specifically, does Δ embody LD as we would want to conceptualize it (e.g., as unexpected underachievement), or is Δ merely a convenient conceptualization of LD because it is a conceptualization that leads directly to easily implemented, operational definitions, however flawed they might be?

Validity

The validity of the IQ-discrepancy model has been extensively studied. Two independent meta-analyses have shown that effect sizes on measures of achievement and cognitive functions are in the negligible to small range (at best) for the comparison of groups formed on the basis of discrepancies between IQ and reading achievement versus poor readers without an IQ discrepancy (Hoskyn & Swanson, 2000; Stuebing et al., 2002), findings similar to studies not included in these meta-analyses (Stanovich & Siegel, 1994). Other validity studies have not found that discrepant and nondiscrepant poor readers differ in long-term prognosis (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Silva et al., 1985), response to instruction (Fletcher, Lyon, et al., 2002; Jiménez et al., 2003; Stage, Abbott, Jenkins, & Berninger, 2003; Vellutino, Scanlon, & Jaccard, 2003), or neuroimaging correlates (Lyon et al., 2003; but also see Shaywitz et al., 2003, which shows differences in groups varying in IQ but not IQ discrepancy). Studies of genetic variability show negligible to small differences related to IQ-discrepancy models that may reflect regression to the mean (Pennington, Gilger, Olson, & DeFries, 1992; Wadsworth, Olson, Pennington, & DeFries, 2000). Similar empirical evidence has been reported for LD in math and language (Fletcher, Lyon, et al., 2002; Mazzocco & Myers, 2003). This is not surprising given that the problems are inherent in the underlying psychometric model and have little to do with the specific measures involved in the model except to the extent that specific test reliabilities and intertest correlations enter into the equations.

Despite the evidence of weak validity for the practice of differentiating discrepant and nondiscrepant students, alternatives based on discrepancy models continue to be proposed, and psychologists outside of schools commonly implement this flawed model. However, given the reliability problems inherent in IQ discrepancy models, it is not surprising that these other attempts to operationalize aptitude–achievement discrepancy have not met with success. In the Stuebing et al. (2002) meta-analysis, 32 of the 46 major studies had a clearly defined aptitude measure. Of these studies,

19 used Full Scale IQ, 8 used Verbal IQ, 4 used Performance IQ, and 1 study used a discrepancy of listening comprehension and reading comprehension. Not surprisingly, these different discrepancy models did not yield results that were different from those when a composite IQ measure was utilized. Neither Fletcher et al. (1994) nor Aaron, Kuchta, and Grapenthin (1988) were able to demonstrate major differences between discrepant and low achievement groups formed on the basis of listening comprehension and reading comprehension.

The differences in these models involve slight changes in who is identified as discrepant or low achieving depending on the cut point and the correlation of the aptitude and achievement measures. The changes simply reflect fluctuations around the cut point where children are most similar. It is not surprising that effect sizes comparing poor achievers with and without IQ discrepancies are uniformly low across these different models. Current practices based on this approach to identification of LD epitomized by the federal regulatory definition and psychiatric classifications are fundamentally flawed.

One-Test (Low Achievement) Models

Reliability

The measurement problems that emerge when a specific cut point is used for identification purposes affect *any* psychometric approach to LD identification. These problems are more significant when the test score is not criterion referenced, or when the score distributions have been smoothed to create a normal univariate distribution. To reiterate, the presence of a natural breakpoint in the score distribution, typically observed in multimodal distributions, would make it simple to validate cut points. But natural breaks are not usually apparent in achievement distributions because reading and math achievement distributions are normal. Thus, LD is essentially a dimensional trait, or a variation on normal development.

Regardless of normality, measurement error attends any psychometric procedure and affects cut points in a normal distribution (Shepard, 1980). Because of measurement error, any cut point set on the observed distribution will lead to instability in the identification of class members because observed test scores will fluctuate around the cut point with repeated testing or use of an alternative measure of the same construct (e.g., two reading tests). This fluctuation is not just a problem of correlated tests or simply a matter of setting better cut scores or developing better tests. Rather, no single observed test score can capture perfectly a student's ability on an imperfectly measured latent variable. The fluctuation in identifications will vary across different tests, depending in part on the measurement

error. In both real and simulated data sets, fluctuations in up to 35% of cases are found when a single test is used to identify a cut point. Similar problems are apparent if a two-test discrepancy model is used (Francis et al., 2005; Shaywitz et al., 1992).

This problem is less of an issue for research, which rarely hinges on the identification of individual children. Thus, it does not have great impact on the validity of a low achievement classification because, on average, children around the cut point who may be fluctuating in and out of the class of interest with repeated testing are not very different. However, the problems for an individual child who is being considered for special education placement or a psychiatric diagnosis are obvious. A positive identification in either example often carries a poor prognosis.

Validity

Models based on the use of achievement markers can be shown to have a great deal of validity (see Fletcher, Lyon, et al., 2002; Fletcher, Morris, & Lyon, 2003; Siegel, 1992). In this respect, if groups are formed such that the participants do not meet criteria for mental retardation and have achievement scores that are below the 25th percentile, a variety of comparisons show that subgroups of underachievers emerge that can be validly differentiated on external variables and help demonstrate the viability of the construct of LD. For example, if children with reading and math disabilities identified in this manner are compared to typical achievers, it is possible to show that these three groups display different cognitive correlates. In addition, neurobiological studies show that these groups differ both in the neural correlates of reading and math performance as well as the heritability of reading and math disorders (Lyon et al., 2003). These achievement subgroups, which by definition include children who meet either low achievement or IQ-discrepancy criteria, even differ in RTI, providing strong evidence for "aptitude by treatment" interactions; math interventions provided for children with reading problems are demonstrably ineffective, and vice versa.

Despite this evidence for validity, concerns emerge about definitions based solely on achievement cut points. Simply utilizing a low achievement definition, even when different exclusionary criteria are applied, does not operationalize the true meaning of unexpected underachievement. Although such an approach to identification is deceptively simple, it is arguable whether the subgroups that remain represent a unique group of underachievers. For example, how well are underachievers whose low performance is attributed to LD differentiated from underachievers whose low performance is attributed to emotional disturbance, economic disadvantage, or inadequate instruction (Lyon et al., 2001)? To use the example of word recognition,

there is little evidence that these subgroups vary in terms of phonological awareness or other language tasks, RTI, or even neuroimaging correlates. In this respect, the validity is weak because the underlying construct of LD is not adequately assessed. Additional criteria are needed, but simply adding a single aptitude measure *decreases* reliability and does not add to the validity of a low achievement definition.

Models Based on Intra-Individual Differences

A commonly proposed alternative to models based on aptitude–achievement discrepancies or low achievement involves an examination of individual differences on measures of cognitive function. Thus, for example, a recent consensus article from 10 major advocacy groups organized by the National Center for Learning Disabilities (2002) stated that "while IQ tests do not measure or predict a student's response to instruction, measures of neuropsychological functioning and information processing could be included in evaluation protocols in ways that document the areas of strength and vulnerability needed to make informed decisions about eligibility for services, or more importantly, what services are needed. An essential characteristic of LD is failure to achieve at a level of expected performance based upon the student's other abilities" (p. 4).

This statement proposes intra-individual differences as a marker for unexpected underachievement. As opposed to a single marker such as IQ discrepancy or low achievement, unexpectedness is operationalized as unevenness in scores across multiple tests. The person identified as LD (by definition) has strengths in many areas of cognitive or neuropsychological function but weaknesses in core attributes that lead to underachievement. The LD is unexpected because the weaknesses lead to selected and narrow difficulties with achievement and adaptive functions. Proponents of this view believe that such approaches identify children as LD based on profiles across tests that differentiate types of LD and also differentiate LD from other childhood disorders, such as mental retardation and behavioral disorders such as attention deficit hyperactivity disorder (ADHD). This approach leads to definitions based on inclusionary criteria in which children are identified as LD based on characteristics that relate to intra-individual differences (Lyon et al., 2001).

Reliability

In essence, the intra-individual difference model employs a multitest discrepancy approach and carries with it the problems involved with estimation of discrepancies and cut points. These problems are inherent in any attempt to identify a person as LD (Fletcher et

al., 2003). However, examining patterns of test scores has long been favored by clinical neuropsychologists, largely because it seems to correspond more closely with clinical practice and because it adds information to the decision-making process (see the elegant discussion of differential test scores, discrepancies, and profiles in Rourke, 1975). The unique reliability issue involves the idea that LD is represented by unevenness in test profiles. This may be true, but does this observation mean that children with flatter profiles are not LD? Severity is correlated with the shape of a profile due to the lack of independence of different tests that might be used to construct the profile (Morris, Fletcher, & Francis, 1993). Children with increasingly severe reading problems, for example, will show increasingly flat profiles across processing measures (e.g., phonological awareness, rapid naming, and vocabulary) in direct correspondence to severity because all these measures are moderately correlated. Thus, if the inclusionary criterion for the presence of LD is evidence of a discrepancy in neuropsychological or processing skills, such an approach may exclude the most severely impaired children, irrespective of global measures such as IQ, because more severely impaired children are less likely to show skill discrepancies due to the intercorrelation of the tests (Morris et al., 1993, 1998).

Validity

A major assumption of a multitest intra-individual differences model is that identification based on performance patterns will lead to enhanced treatment of children with LD. It is commonly assumed that such tests point out areas that need intervention. However, there is little evidence that strengths and weaknesses in processing skills are related to intervention outcomes. It is well established that training in underlying processes does not usually generalize into the related academic area (Lyon & Moats, 1988; Reschly, Tilly, & Grimes, 1999; Vellutino, 1979). For example, training on phonological awareness skills without explicit transfer to a letter component produces gains in phonological awareness but not in reading (National Reading Panel, 2000). Training in auditory or visual perceptual skills does not lead to better outcomes for children identified as "auditory" or "visual" learners (Lyon, Fletcher, Fuchs, & Chhabra, in press; Vellutino, Fletcher, Scanlon, & Snowling, 2004).

There is support for the idea that intra-individual differences identify some children as LD, epitomized by the link of dyslexia with word recognition and phonological processing (Vellutino et al., 2004). Even here the intra-individual differences model focuses on skills that are only correlated with the achievement domain. Simply identifying children with LD based solely on processing skills is questionable and would likely yield many false positive identifications of children as LD without achievement difficulties (Torgesen, 2002). The reliability of many processing measures is lower than those associated with the achievement (or IQ) domain, so such false positives should be expected. Other than the word recognition–phonological processing link, relations of processing and other forms of LD are not well established (Torgesen, 2002). Finally, what do we learn about variability in processing skills that is not apparent in profiles across achievement domains (Fletcher et al., 2003)? In fact, the model has the most validity at the level of achievement markers but simply collapses into a low achievement model in the absence of processing measures. Thus, if we accept the notion that specific discrepancies in cognitive domains are a unique marker for LD, given that the processing measures are usually linked to an achievement domain, what is unique about variations in processing skills that is not apparent in variations in achievement domains? Would we eliminate as LD students who have difficulties in reading, math, and writing? This is not viable, as impairments in all domains often occur in non-mentally retarded children with language-based difficulties.

Models Incorporating RTI

An alternative approach to status models that would increase the reliability of these would increase the number of time points whereby a child was assessed. Shepard (1980), for example, proposed that IQ discrepancies could be assessed more reliably if a child was tested four times. The impracticality of such an approach, which would require about 10 to 12 hr per child, is obvious, not to mention that even more resources would be devoted to determination of eligibility, taking away funds and time needed for intervention.

Another approach to increasing the number of time points would involve much shorter assessments of key achievement skills over time. These approaches, or RTI models, typically involve identification practices based in part on multiple short assessment probes of knowledge and performance in a specific academic domain, such as reading or math (Fuchs & Fuchs, 1998). By linking multiple assessments to specific attempts to intervene with the child, the construct of unexpected underachievement can be operationalized, in part, on the basis of nonresponsiveness to instruction to which most other students respond (Gresham, 2002). In fact, this is still a variation of a discrepancy model, but the advantage is that the model is better identified because of multiple short assessments of a key attribute (e.g., reading, math) over time.

Such models have been proposed in several recent consensus reports that address LD identification (Bradley, Danielson, & Hallahan, 2002; President's

Commission on Excellence in Special Education, 2002), most notably in a recent report of the National Research Council (Donavon & Cross, 2002). These reports suggest that one criterion for LD identification is when a student does not respond to high-quality instruction and intervention. The implementation of this approach requires frequent monitoring of progress as the student receives the intervention (Fuchs & Fuchs, 1998). Speece and Case (2001) found that serial assessments of growth and level of performance in reading fluency predicted reading problems in at-risk children better than a single assessment of fluency. This approach is anchored in a system known as curriculum-based measurement, where the assessments themselves have adequate reliability and are constantly improving (Fuchs & Fuchs, 1999; Shinn, 1998).

Reliability

Are RTI approaches that involve multiple assessments over time psychometrically more reliable than traditional approaches to LD identification? An approach based on multiple measures over time has the potential to reduce the difficulties encountered with reliance on a single assessment at a single time point. Certainly the reliability of the multiple assessment approach is greater than if the single assessment is used to form a discrepancy, because typically the discrepancy will be a poorer (i.e., less reliable) measure of the true difference than the observed measures of their respective underlying constructs. Focusing on successive measurements over time has the effect of moving the identification process from "ability–ability" comparisons (two different abilities compared at one point in time) to "ability change" models (same ability over time). Such approaches have the potential to ameliorate the difficulties associated with ability–ability discrepancies, whether univariate or bivariate, because they involve the use of more than two assessment time points. Generally, the more information that is brought to bear on any eligibility or diagnostic decision, the more reliable the decision, although it is certainly possible to create counterexamples by combining information from irrelevant or confounding sources. Such irrelevancies are not likely to be introduced by assessing the same skill over time as in a model that incorporates RTI, when that skill was previously deemed relevant to assess at a single time point.

Conceptually, the study of change is made more feasible by the collection of multiple assessments because the precision by which change can be measured improves as the number of time points increases (Rogosa, 1995). When more than two assessment time points are collected, the reliability of estimated change can also be estimated directly from the data, and the imprecision inherent in individual estimates can be used to provide improved estimates of growth parameters for individual students as well as for groups of students. If change is not linear, the use of four or more time points can map the form of growth. And for those who favor status models over change or learning models, it remains possible to use the intercept term in the individual growth model as an estimate of status. This intercept provides a more precise estimate of true status at any single point in time than would any single assessment.

These approaches are not without difficulty. The introduction of serial assessments has not eliminated the necessity of indirect estimation of the parameters of interest. In the discrepancy model, D is used to estimate Δ. A model incorporating RTI uses a complex function of the observed data for individual i as well as the data from many other individuals to estimate each of the π_{ij}, the j true learning parameters for individual i. Different approaches to this estimation problem have varying strengths and weaknesses but all will make assumptions about the arithmetic form of the model, the distribution of the learning parameters, and the distributions of the errors. The ramifications of these assumptions for inferences about individual learning parameters must be studied in the LD context.

Models based on RTI also involve imperfect measures that include measurement error (Fletcher et al., 2003). However, this problem is reduced because of the use of multiple assessments and the borrowing of precision from the entire collection of data to provide a more precise estimate of the growth parameters of each individual. Thus, it becomes possible to estimate a child's "true" status more precisely as well as to estimate the rate of skill acquisition and to use these estimates as indicators of LD. In addition, this approach to estimation makes assumptions about the distribution of errors of measurement. In some cases, errors might be assumed to be uncorrelated. Again, this assumption must be examined in terms of its importance to inferences about individual status and rates of learning. In many cases, the inclusion of multiple assessment time points will allow this assumption to be relaxed, and the correlation among errors of measurement can be estimated and taken into account in forming inferences about individual status and rates of learning.

There still could be a need to identify individual children as LD based on cut points unless the entire process devolves to clinical judgment. Models that include RTI do not solve the issue of the dimensional versus categorical nature of LD. Determining cut points and benchmarks, for example, will continue to be an arbitrary process until cut points are linked to functional outcomes (Cisek, 2001), an issue never really addressed in LD identification for any identification model. However, models that include RTI have the promise of incorporating functional outcomes because they are tied to intervention response.

513

Validity

The introduction of serial assessments has an advantage beyond any statistical advantage it may confer for the estimation of individual's true status. Specifically, the introduction of serial assessments brings learning and the measurement of change to the forefront in conceptualizations of LD. The collection of serial assessments under specified conditions of effective instruction simultaneously focuses the definition of LD on a failure to learn, where learning can be measured more directly. Moreover, the specific instructional elements and the conditions under which they are implemented can be described, thereby providing a clearer basis for the expectation of learning and the unexpectedness of any failure to learn. Finally, focusing on multiple assessments in a RTI model has the advantage of clearly tying the identification process to the most important component of the construct of LD, which is unexpected underachievement. Models that incorporate RTI may identify a unique group of children that can be clearly differentiated from other low achievers in terms of cognitive correlates, prognosis, and even neurobiological factors.

Studies of children defined using different methods as responders and nonresponders clearly show significant differences in cognitive skills. For example, Stage et al. (2003), Vellutino et al. (2003), and Vaughn, Linan-Thompson, and Hickman-Davis (2003) found that nonresponders to early intervention differed from responders in both preintervention achievement scores and preintervention cognitive tasks. Nonresponders typically had more severe deficits in both reading-related factors (e.g., phonemic awareness, fluency) and reading skills. In recent imaging studies involving both early intervention and remediation of older students (see Fletcher, Simos, Papanicolaou, & Denton, 2004), we likewise found that individuals who were nonresponders showed more severe reading difficulties prior to intervention. More dramatic were the differences in neuroimaging correlates between those who responded to intervention and those who did not. We have found that nonresponders persist with a brain activation pattern that generally demonstrates a failure to activate left hemisphere areas known to be involved in the development of reading skills. In fact, those who were nonresponders showed predominant right-hemisphere activity much like that observed in children and adults with identified reading disabilities (Fletcher et al., 2004).

Implications for Clinical Assessments

This review of classification models may seem removed from the question of how to conduct clinical assessments of children suspected of LD. In fact, when a psychologist conducts any assessment for LD, the selection of tests reflects the underlying classification model and the constructs it specifies. If the psychologist or educator adopts an aptitude–achievement discrepancy model, the primary tools will be the tests used to operationalize aptitude (e.g., IQ) and achievement. If the clinician adopts a low achievement model, aptitude will not be measured—just achievement. An intra-individual differences model will require neuropsychological or cognitive processing measures. If a model is used that incorporates RTI, assessments of the integrity of the implementation of the intervention and progress monitoring assessments are necessary.

In evaluating models, we found little evidence that supports the aptitude–achievement and intra-individual difference models. Both involve the assessment of cognitive processes that do not contribute to the identification of a unique group of underachievers with LD and have serious reliability problems. The low achievement model has more reliability and validity but does not identify a unique group of underachievers. RTI criteria may permit identification of a unique group of underachievers but by themselves are not sufficient for identification of LD. Combining the strengths of the low achievement and RTI models leads to a hybrid model that invokes concepts of low achievement and RTI. This model can be expanded to incorporate assessment of contextual factors and other disorders that should be evaluated because of the need for differential treatment (Fletcher, Foorman, et al., 2002).

Learning disorders attributable to mental retardation, sensory problems (blindness, deafness), language status (e.g., English as a second language), or transient factors (adjustment difficulties, disruption of the home or school environment) should not be identified as LD. We have not included economic disadvantage, comorbid emotional and behavior disorders, or established neurological disorders as exclusionary criteria and would stipulate that the only way to exclude LD in children with these associated conditions is to provide an intervention that is appropriate and evaluate RTI. A classification of LD may exclude children with emotional or neurological disorders, or those who are economically disadvantaged from the LD category because of policy or resource issues—all are eligible for special education—but children with these associated conditions have forms of underachievement that are difficult to distinguish from those in children with LD. In the end, LD should be identified only after adequate opportunity to learn has been systematically evaluated. Those who do not respond to intervention need more specialized, individualized, and intensive treatments, as well as the probable conferment of disability status and the civil rights protections that come with identification. It is the intractability as indicated by an inadequate response to quality instruction that must be present to identify a child as LD. If a child responds, LD is

not indicated. Any child who has achievement difficulties should receive intervention, whether it is tutoring with support by a college student or intensive intervention by an experienced, well-trained academic therapist.

This model is quite different from the one that child clinical and other psychologists have utilized for the past few decades, and some may respond by suggesting that this model can only be implemented in schools. In fact, we argue that in the absence of an evaluation of RTI, LD should not be identified in any setting—school, clinic, hospital, and so on. We conceptualize traditional clinical evaluations as an opportunity to identify children as "at risk" for LD and to intervene with any child who is struggling to achieve. In schools, screening for reading problems can be done on a large-scale basis in kindergarten and Grade 1 as advocated in Donovan and Cross (2002) and implemented in states such as Texas (Fletcher, Foorman, et al., 2002). Those who are identified as at-risk have their progress monitored and receive increasingly intense, multitiered interventions that may eventuate in identification for special education in LD (Vaughn & Fuchs, 2003). In a multitiered intervention approach, children are screened for risk characteristics, such as weaknesses in letter-sound knowledge and phonological awareness in kindergarten and word reading in Grade 1, with immediate monitoring of progress (Torgesen, 2002). Depending on the rate of progress, interventions are intensified and modified in an effort to accelerate the rate of development of an academic skill. Children are not identified with disabilities until the final tier of the process.

Evaluations outside of schools should utilize a similar approach based on measurement of the three components of the hybrid model proposed by the consensus group in Bradley et al. (2002): (a) low achievement, (b) RTI, and (c) consideration of contextual factors and exclusions. Any psychological evaluation of a child or adolescent should consider the relevant achievement constructs (see following section) that represent the different types of LD. If there is evidence of low achievement, the focus should not be on extensive assessments of processing skills but on referral to an appropriate source for intervention. The psychologist should expect to have a working relationship with the intervention source so that RTI will be measured. This means that clinical child psychologists must be knowledgeable about educational interventions and prepared to develop a treatment plan that incorporates this form of intervention, just as they may be prepared to work with a physician around medication for problems with attention or anxiety. It is perfectly reasonable to ask the child to return every 4 to 6 months to repeat achievement tests and independently evaluate progress in conjunction with more frequent assessments of progress obtained by the intervention source.

The psychologist should also evaluate for other problems that may be associated with low achievement to adequately plan treatment. If mental retardation is suspected, IQ, adaptive behavior, and related assessments consistent with this classification can be administered. But note that if the child or adolescent has achievement scores in reading comprehension or math that are within two standard deviations of the mean (consistent with traditional legal definitions of mental retardation), or development of adaptive behavior obviously inconsistent with mental retardation, assessment of IQ is not necessary as such levels of performance preclude mental retardation. Some children may have oral language disorders that require speech and language intervention that will require referral and additional evaluation. Screening with vocabulary measures and through interacting with the child will help identify these children; the vocabulary screen will also help identify children who may benefit from additional intellectual screening. Many children with achievement difficulties or LD also have comorbid difficulties with attention and both internalizing and externalizing psychopathology. These disorders need to be assessed and treated, as simply referring a child for educational intervention without addressing comorbidities will surely increase the probability of a poor RTI. We believe that no clinical evaluation of a child should be conducted without a documentation of achievement levels through direct assessment or school report of such an assessment. If achievement deficits are apparent, intervention of some sort should be provided. It is not likely that treating a child for a comorbid disorder, such as ADHD, will result in improved levels of achievement in the absence of educational intervention.

Altogether, we are suggesting that from the perspective of LD, psychologists should perform assessments for emotional and behavioral disorders consistent with other articles in this special section. For LD, they need to administer achievement tests and evaluate RTI. This is regardless of subdiscipline (e.g., school psychologist, child clinical psychologist, neuropsychologist) or setting. To evaluate achievement, individualized norm-referenced assessments should be conducted. RTI requires assessments of intervention integrity and monitoring of progress.

Evaluating Achievement

Identifying specific achievement tests is not difficult, although tests for some domains are better developed than others. Lyon et al. (2003) suggested that LD represented six major achievement types, including (a) word recognition; (b) reading fluency; (c) reading comprehension; (d) mathematics computations; (e) reading and math, which is not really a comorbid association but a more severe reading problem with distinct math difficulties; and (f) written expression, which

could involve spelling, handwriting, or text generation. These patterns were drawn from the research literature (e.g., Rourke & Finlayson, 1978; Siegel & Ryan, 1988; Stothard & Hulme, 1996), but an extensive discussion of the evidence for these types is beyond the scope of this article (see Lyon et al., 2003). The assessment implications are straightforward. Many children and adolescents will have difficulties in more than one domain, so a thorough assessment of academic achievement is very important.

A set of achievement tests should be used. It is helpful to use tests from the same battery because the normative group is the same, which facilitates comparisons across tests. However, the battery from which these tests are chosen is less important than the constructs that are measured. Any single battery has strengths and weaknesses that can be supplemented with measures from other assessments. Given the suggestion that six types of LD may exist, the important constructs are word recognition, reading fluency, reading comprehension, math computations, and written expression. We usually assess spelling as a screen for written expression and handwriting difficulties and math and writing fluency as supplemental assessments.

Table 1 outlines these constructs and how they can be assessed with the commonly used Woodcock–Johnson Achievement Battery–III (WJ; Woodcock, McGrew, & Mather, 2001) or the Wechsler Individual Achievement Test–II (WIAT; Wechsler, 2001). We use the WJ and WIAT because they meet established criteria for reliability (internal consistency and test–retest) and validity (construct and concurrent) and were developed to account for variations in ethnicity and socioeconomic status. In particular, the normative sampling took into account this type of variation, and analyses (differential item functioning) were conducted to identify items that were not comparable across these sources of normative variation. There are also other norm-referenced assessments that can be used to supplement the WJ or WIAT, which we discuss later. Few of these supplemental measures have been developed with the

care of the WJ or the WIAT, particularly with regard to the adequacy of the normative base and attempts to address different forms of normative variation.

Table 1 should not be taken to indicate that there are 11 different types of LD, one for each test. To reiterate, many children have problems in multiple domains. The pattern of academic strengths and weaknesses is an important consideration (Fletcher, Foorman, et al., 2002; Rourke, 1975). Table 1 identifies constructs and core tests that would be administered to every child and supplemental tests that would be used if there were concerns about a particular academic domain. If the referral indicated concerns about a particular area, additional tests from other measures would be used. Most children with significant academic problems where LD may eventually be a concern have difficulty with word recognition and consequently tend to have problems across domains of reading. Going beyond the core tests is usually not necessary if the child has problems with word recognition. Isolated problems with reading comprehension and written expression occur infrequently. If the problem is specifically math, using assessments in addition to the WJ or WIAT is helpful in ensuring that the deficiency is not just a matter of attention difficulties.

An advantage of the WJ and WIAT is the assessment of word recognition for both real words and pseudowords, the latter permitting an assessment of the child's ability to apply phonics rules to sound out words. Most achievement batteries assess recognition of real words, which is the essential component. These measures tend to be highly intercorrelated across different assessment batteries, including the Wide Range Achievement Test–III (Wilkinson, 1993), and the Accuracy measure from the Gray Oral Reading Test–Fourth Edition (Wiederholt & Bryant, 2001).

The WJ also has a silent reading speed subtest that, in our assessments, is highly correlated with other fluency measures despite the fact that it is not simply oral reading speed, requiring the child to answer some questions while reading a series of passages for 3 min.

Table 1. *Achievement Constructs in Relation to Subtests From the WJ and the WIAT*

Construct	WJ Subtest	WIAT Subtest
Core Tests		
Word Recognition	Word Identification	Word Reading
	Word Attack	Pseudoword Decoding
Reading Fluency	Reading Fluency	—
Reading Comprehension	Passage Comprehension	Reading Comprehension[a]
Math Computations	Calculation	Numerical Operations
Written Expression	Spelling	Spelling
Supplemental Tests		
Math Fluency	Math Fluency	—
Writing Fluency	Writing Fluency	—
Math Concepts	Quantitative Concepts	—
Written Expression	Writing Samples	Written Expression[a]

[a]Also assesses fluency

The WIAT permits assessment of reading speed during silent-reading comprehension. Both assessments are easily supplemented with the Test of Word Reading Efficiency (Torgesen, Wagner, & Raschotte 1999), which involves oral reading of real words and pseudowords on a list. The Test of Reading Fluency (Deno & Marston, 2001) is an option that requires text reading. Both measures are quick and efficient, and the former was designed with item analyses addressing differential item responses across ethnic groups. Whenever text is read out loud, fluency can be assessed as words read correctly per minute. The Gray Oral Reading Test–Fourth Edition includes a score for fluency of oral text reading.

Reading comprehension can only be screened with the WJ Passage Comprehension subtests which is a cloze-based assessment in which the child reads a sentence or passage and fills in a blank with a missing word. The Reading Vocabulary subtest is used to create a reading comprehension composite, but it places such a premium on decoding that we usually do not administer it. The WIAT also does not demand much reading of text. Some children who struggle to comprehend text in the classroom do not have difficulties on these subtests because the level of complexity rarely parallels what children are expected to read on an everyday basis. Supplemental assessments using the Group Reading Assessment and Diagnostic Education (Williams, Cassidy, & Samuels, 2001), the Gray Oral Reading Test–Fourth Edition, or even one of the well-constructed reading comprehension assessments from the group-based Stanford Achievement Test–10th Edition (Harcourt Assessment, 2002), Iowa Test of Basic Skills (Hoover, Hieronymous, Frisbie, & Dunbar, 2001), or similar instrument is essential. Often children have had these assessments in school, and it is useful to review results as part of the overall evaluation.

Reading comprehension is a difficult construct to assess (Francis, Fletcher, Catts, & Tomblin, in press). In evaluating comprehension skills, the assessor should attend to the nature of the material the child is asked to read and the response format. Reading comprehension tests vary in what the child reads (sentences, paragraphs, pages), the response format (cloze, open-ended questions, multiple-choice, think aloud), memory demands (answering questions with and without the text available), and how deeper aspects of meaning are evaluated (understanding of the essential meaning vs. literal understanding, vocabulary knowledge and elaboration, ability to infer or predict). It may be difficult to determine the source of the child's difficulties based on a single measure. Thus, if the issue is comprehension and the source is not in the child's word recognition or fluency skills, multiple measures that assess reading comprehension in different ways are needed.

For math, the Calculations subtest of the WJ and Numerical Operations subtest of the WIAT are paper-and-pencil tests of math computations (Table 1). Low scores on this type of task predict variation in cognitive skills depending on other academic strengths and weaknesses (Rourke, 1993). However, low scores could reflect problems with fact retrieval and verbal working memory if word recognition is comparably lower, as opposed to problems with procedural knowledge if word recognition is significantly higher and not deficient. Deficient scores can also reflect problems paying attention, especially in children with ADHD. The math computations subtests from the Wide Range Achievement Test–III is also frequently used and is useful because it is timed and the problems are less organized. The key is the paper-and-pencil assessment of math computations, which is how difficulties in math are typically manifested in children who do not have reading problems. As in reading, assessments of fluency are helpful, although there is no evidence suggestive of a math fluency disorder. In Table 1, the WJ Math Fluency subtest is identified as a supplemental measure, representing a timed assessment of single-digit arithmetic facts that may be helpful for identifying children who lack speed in basic arithmetic skills. Such difficulties make it difficult to master more advanced aspects of mathematics. If an assessment of math concepts is needed, which we would do only if math was an overriding concern, the Quantitative Concepts subtest of the WJ is more useful than the WJ Applied Problems or WIAT Math Reasoning subtests, which introduce word problems that are difficult for children with reading difficulties.

Written expression is most difficult to assess, partly because it is not clear what constitutes a disorder of written expression—spelling, handwriting, or text generation (Lyon et al., 2003). Obviously problems with the first two components will constrain text generation. Spelling should be assessed as it may represent the primary source of difficulty with written expression for children, especially if they also have word-recognition difficulties. The analysis of spelling errors (Rourke, Fisk, & Strang, 1986) can be informative in understanding whether the problem is with the phonological component of language or with the visual form of letters (i.e., orthography). Spelling also permits an informal assessment of handwriting.

Table 1 identifies WJ and WIAT measures of written expression. The utility of these measures is not well established, and the significant generation of text in terms of construction and writing of passages and stories is not really required. As with reading comprehension, it may be important to supplement or even replace this assessment with a test such as the Thematic Maturity subtest of the Test of Written Language (Hammill & Larsen, 1998). Measuring fluency with a measure such as the WJ Writing Fluency subtest may also be

useful. As in reading and math, fluency of writing predicts the quality of composition.

From this type of assessment, characteristic patterns emerge that will demarcate the classification and indicate a need for specific kinds of intervention. For each of the six types of LD, there are interventions with evidence of efficacy that should be utilized in or out of a school setting (Lyon et al., in press). The goal is not to diagnose LD, which is not feasible in a one-shot evaluation for the psychometric and conceptual reasons outlined previously, but to identify achievement difficulties that can be addressed through intervention. If the assessor is knowledgeable about these patterns, very specific intervention recommendations, as well as the need for other assessments, can be made.

Table 2 summarizes achievement patterns that are well established in research (Fletcher, Foorman, et al., 2002; Lyon et al., 2003). Intervention should be considered for any child who performs below the 25th percentile on a well-established assessment, with an understanding that these are not firm cut points and should be evaluated across all the measures. We are not indicating that 25% of all children have a LD, only that scores below the 25th percentile are commonly associated with low performance in school, assuming the cut point is reliably attained. In examining Table 2, the decision rules should not be rigidly applied and are simply guidelines to assist clinicians. Identifying LD is always based on factors beyond just the test scores. The decision process should focus on what is needed for intervention, which requires an assessment of contextual

variables and the presence of comorbid disorders that influence decisions about what sort of plan will be most effective for an individual child. Low achievement is related to many contextual variables, which is why the flexibility in special education guidelines allows interdisciplinary teams to base decisions on factors that go beyond test scores. The purpose of assessment is ultimately to develop an intervention plan.

Evaluating Response to Instruction

Once a child is screened or tested for achievement deficits, progress should be monitored if a problem is identified. It is astonishing that U.S. special education guidelines do not require at least yearly readministration of the achievement tests that were used to justify the placement as one method of assessing the efficacy of the intervention plan. If a child is responding to intervention, his or her rate of development should be accelerated relative to the normative population (i.e., the achievement gap is closed). As part of this assessment of RTI, progress should be monitored on a frequent basis if the problem is with word recognition or fluency, math computations, or spelling. Reading comprehension and higher forms of written expression will show less rapid change and progress, as monitoring tools for these types of problems have not been adequately developed.

Most of the tests mentioned here have alternative forms. But some have been developed to permit assessments with even more frequency and are referred to as "curriculum-based assessments" (Fuchs & Fuchs,

Table 2. *Eight Achievement Patterns Associated With Intervention*

1. Decoding and Spelling < 90; Arithmetic one half standard deviation higher than word recognition and spelling and at least 90. This is a pattern characterized by problems with single word decoding skills and better arithmetic ability. Reading comprehension will vary depending on how it is assessed but is usually impaired. Children with this pattern have significant phonological language problems and strengths in spatial and motor skills (Rourke & Finlayson, 1978).
2. Arithmetic < 90, Decoding and Spelling > 90 and at least 7 points higher. Children with difficulties that only involve math show this pattern, which is associated with problems with motor and spatial skills, problem-solving deficiencies, and disorganization (Rourke & Finlayson, 1978). It usually represents problems with math procedures as opposed to math facts (Lyon et al., 2003).
3. Decoding, Comprehension, Spelling, and Arithmetic < 90. This pattern represents a problem with word recognition characterized by language and working memory problems more severe than in children with poor decoding and better development of math skills (Rourke & Finlayson, 1978). The math problem involves learning and retrieving math facts (Lyon et al., 2003).
4. Spelling and Arithmetic < 90, Decoding > 90 and 7 points higher. Essentially the same pattern as Number 3 except the motor (and writing) component is more severe.
5. Reading Comprehension < 90 and 7 points below decoding. This pattern often reflects long-term oral language disorder. Problems with receptive language, short-term memory, and attention are apparent, with strengths in phonological language skills (Stothard & Hulme, 1996).
6. Decoding skills 7 points lower than Comprehension skills and < 90. This pattern reflects a phonological language problem with usually better than average semantic language and spatial skills (Stothard & Hulme, 1996). The pattern is not apparent for reading comprehension measures that are timed or require significant amounts of text reading.
7. Reading Fluency < 90 and < Decoding by one half standard deviation will reflect a problem where accuracy of word reading is less of a problem than automaticity of word reading (Lyon et al., 2003).
8. Spelling < 90. This pattern reflects (a) motor deficits in a young child or (b) residuals of earlier phonological language problems that have been remediated or compensated in older children and adults. The pattern is common in adults with a history of word recognition difficulties. Fluency is often impaired.

Note: The patterns are based on relations of reading decoding, reading fluency, reading comprehension, spelling, and arithmetic. It is assumed that any score below the 25th percentile (standard score = 90) is impaired and that a difference of one half standard deviations is important (± 7 standard score points). The patterns should be considered prototypes and the rules loosely applied (adapted from Fletcher, Foorman, Boudousquie, Barnes, Schatschneider, & Francis, 2002). These patterns are not related to IQ scores.

1999). Such measures are often used by the intervener (e.g., teacher) to document how well a child is responding to instruction. Typically a child would read a short reading passage appropriate for grade level (or do a set of math computations) for 2 to 3 min. The number of words (or math problems) correctly read (or computed) would be graphed over time and compared against grade-level benchmarks, representing a criterion-referenced form of assessment. A child may be screened with these measures, and those performing below the benchmark may be candidates for intervention, especially in schools.

Such assessments should also be accompanied by observations of the integrity of the implementation of the intervention, including the amount of time spent on supplemental instruction, especially if the child does not appear to be making progress. School psychologists are often well prepared in this area of assessment. Although a psychologist operating outside of a school may not be in a position to do curriculum-based assessments or to personally evaluate the intervention, such assessments should be expected, especially if the referral is to a private academic therapist.

A variety of methods have been developed, and the assessments with the most widespread utilization are the Monitoring Basic Skills Progress (Fuchs, Hamlett, & Fuchs, 1990), which assesses reading, math, and spelling fluency, and the Dynamic Indicators of Basic Early Reading Skills (Good, Simmons, & Kame'enui, 2001), a battery of different reading fluency measures. Some of these tools are focused primarily on the lower grades, but the norm-referenced assessments of fluency identified previously—especially if they have alternative forms—can be used with older students,. These measures meet accepted psychometric criteria for reliability and validity. The curriculum-based assessment measures have not been assessed as formally for differential item functioning but have been widely employed with school populations that are quite diverse (Fuchs & Fuchs, 1999; Shinn, 1998).

Conclusions

Based on our evaluation of models, we propose a hybrid model that incorporates features of low achievement and RTI models for the identification of children as LD. We specifically do not find evidence to support extensive assessments of cognitive, neuropsychological, or intellectual skills to identify children as LD. Although some may view this model as only for schools, we reject the idea that the routine evaluations done in the past by psychologists and educators outside of school settings are useful for LD. We find little value in the idea of evaluating a child in a single assessment and concluding that the child has LD based on an IQ-achievement discrepancy, low achievement, or profiles on neuropsychological tests, largely because such assessments are not directly related to treatment and the diagnosis itself is not reliable. As soon as it is apparent that the child has an achievement problem, a referral for intervention should be made and the resources that might be spent on diagnosis should be spent on intervention. Children should not be diagnosed as LD until a proper attempt at instruction has been made. Assessment of achievement skills should be a routine part of any psychological evaluation of a child and cannot be seen as the province of just the schools. Serial monitoring of RTI and the integrity of instruction should be completed before children are identified as LD. There are issues involved in the intervention component, estimation of slope and intercept effects, as well as decisions that have to be made about cut points that will differentiate responders and nonresponders (Gresham, 2002). For these reasons alone, RTI cannot be the sole criterion for identification, and flexibility in decision making is required. At the same time, there appears to be considerable validity to this approach, implying that it is indeed possible to reliably identify nonresponders as a group with unexpected underachievement.

In addition to the evidence for validity (and the greater reliability of the underlying psychometric model), the model does not require the use of exclusionary criteria (especially emotional disturbance and economic disadvantage) to operationalize unexpected underachievement, thus capturing the construct of LD. This is an important consideration given the lack of evidence validating classifications that utilize these particular exclusions (Kavale, 1988; Lyon et al., 2001). The model does operationalize the concept of opportunity to learn, which is rarely directly assessed as part of LD identification. It is also a model that can only be implemented in an instructional setting, such as a school, or in clinical settings outside of public schools where remediation is utilized, such as an academic therapy setting. But it is not consistent with the traditional approach to LD identification based on a single administration of a test battery and consideration of a diagnosis, which we believe is an outmoded model that detracts from intervention. In the absence of an attempt to systematically instruct the child, LD cannot be "diagnosed," obviating the traditional "test and treat" model, as identifying LD must be the end product of an attempt to instruct the child (i.e., "treat and test"). This is not a post hoc approach but rather an argument that in the absence of the opportunity to learn exclusion, the concept of LD has no basis in evidence, and low achievement per se is not adequate evidence for LD. Such an approach ties the concept of LD to treatment, which is important. It may be that a single assessment may indicate "risk" or even an achievement disorder. But such an assessment cannot indicate a "disability" in the absence of functional criteria that would include opportunity to learn.

A final comment involves what some will see as equating LD with measurable deficits on achievement tests. Some will argue that the mere presence of a deficit on a measure of processing skills means that person should be identified by LD, in part because of the belief that such deficits indicate a brain anomaly. The most common example is the linking of "executive function" deficits with LD. We argue that the concept of LD is empty without a focus on achievement, largely because it becomes more difficult to identify a unique subgroup representing LD that would be different from other classes of childhood disorders. Executive functions, for example, are often linked to ADHD, but classifications of ADHD based on executive function deficits as assessed by cognitive tests do not have much validity (Barkley, 1997). Moreover, executive function deficits characterize many childhood populations.

More fundamentally, consider an overarching classification of childhood learning and behavioral difficulties. For LD, achievement deficits represent markers for the underlying classification. What distinguishes the LD prototype from, for example, a behavioral disorder such as ADHD is the presence of an achievement deficit. If a child with ADHD has an achievement deficit, it is usually reflective of a comorbid association (Fletcher, Shaywitz, & Shaywitz, 1999). If we expand our classification to mental retardation, the key for differentiating mental retardation from LD (or ADHD) is not just the intelligence test score. Rather, the major difference is in adaptive behavior, where mental retardation should reflect a pervasive deficit in adaptive behavior and LD as a relatively narrow deficit (Bradley et al., 2002). So a classification of these three major disorders requires markers for achievement, attention-related behaviors, and adaptive behavior. In the absence of these types of markers, and a focus on classification, all children with problems are simply disordered and there is no need for assessment because they would all require the same interventions. When LD is tied to levels and patterns of achievement, an evidence base for differential interventions focused on learning in specific academic domains emerges. This is the strongest evidence for the validity of the concept of LD, its classification, and the source of evidence-based approaches to assessment and identification.

References

Aaron, P. G., Kuchta, S., & Grapenthin, C. T. (1988). Is there a thing called dyslexia? *Annals of Dyslexia, 38,* 33–49.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Barkley, R. A. (1997). *ADHD and the nature of self-control.* New York: Guilford.

Bereiter, C. (1967). Some persisting dilemmas in the measurement of change. In C. W. Harris (Ed.), *Problems in the measurement of change.* Madison: University of Wisconsin Press.

Bradley, R., Danielson, L., & Hallahan, D. P. (Eds.). (2002). *Identification of learning disabilities: Research to practice.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Christensen, C. A. (1992). Discrepancy definitions of reading disability: Has the quest led us astray? *Reading Research Quarterly, 27,* 276–278.

Cisek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cisek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 3–18). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Deno, S. L., & Marston, D. (2001). *Test of Oral Reading Fluency.* Minneapolis: Educators Testing Service.

Donavon, M. S., & Cross, C. T. (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

Fisher, S. E., & DeFries, J. C. (2002). Developmental dyslexia: Genetic dissection of a complex cognitive trait. *Nature: Review of Neuroscience, 10,* 767–780.

Fletcher, J. M., Foorman, B. R., Boudousquie, A. B., Barnes, M. A., Schatschneider, C., & Francis, D. J. (2002). Assessment of reading and learning disabilities: A research-based, intervention-oriented approach. *Journal of School Psychology, 40,* 27–63.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, B. A., & Shaywitz, S. E. (1993). Classification of learning disabilities: Relationships with other childhood disorders. In G. R. Lyon, D. Gray, J. Kavanagh, & N. Krasnegor (Eds.), *Better understanding learning disabilities* (pp. 27–55). New York: Brookes.

Fletcher, J. M., Lyon, G. R., Barnes, M. Stuebing, K. K., Francis, D. J., Olson, R., et al. (2002). Classification of learning disabilities: An evidence-based evaluation. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 185–250). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Fletcher, J. M., Morris, R. D., & Lyon, G. R. (2003). Classification and definition of learning disabilities: An integrative perspective. In H. L. Swanson, K. R. Harris, & S. Graham. *Handbook of learning disabilities* (pp. 30–56). New York: Guilford.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., et al. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 85,* 1–18.

Fletcher, J. M., Shaywitz, S. E., & Shaywitz, B. A. (1999). Comorbidity of learning and attention disorders: Separate but equal. *Pediatric Clinics of North America, 46,* 885–897.

Fletcher, J. M., Simos, P. G., Papanicolaou, A. C., & Denton, C. (2004). Neuroimaging in reading research. In N. K. Duke & M. H. Mallette (Eds.), *Literacy research methods* (pp. 252–286). New York: Guilford.

Francis, D. J., Fletcher, J. M., Catts, H., & Tomblin, B. (in press). Dimensions affecting the assessment of reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Current issues in reading comprehension and assessment.* Mahwah, NJ: Lawrence Erlbaum Associate, Inc.

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of learning disabilities: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38,* 98–108.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88,* 3–17.

Fuchs, L., & Fuchs, D. (1998). Treatment validity: A simplifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research and Practice, 4,* 204–219.

Fuchs, L., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *The School Psychology Review, 28,* 659–671.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1990). *Monitoring basic skills progress.* Austin, TX: PRO-ED.

Gilger, J. W. (2002). Current issues in the neurology and genetics of learning-related traits and disorders: Introduction to the special issue. *Journal of Learning Disabilities, 34,* 490–491.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). 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, 5,* 257–288.

Gresham, F. M. (2002). Responsiveness to intervention: An alternative approach to the identification of learning disabilities. In R. Bradley, L. Danielson, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 467–564). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Hammill, D. D., & Larsen, S. C. (1996). *Test of Written Language.* Austin, TX: PRO-ED.

Harcourt Assessment. (2002). *Stanford Achievement Test* (10th ed.). New York: Author.

Hoover, A. N., Hieronymous, A. N., Frisbie, D. A., & Dunbar, S. B. (2001). *Iowa Test of Basic Skill.* Itasca, IL: Riverside.

Hoskyn, M., & Swanson, H. L (2000). Cognitive processing of low achievers and children with reading disabilities: A selective meta-analytic review of the published literature. *The School Psychology Review, 29,* 102–119.

Individuals With Disabilities Education Act, 20 U. S. C. 1400 *et seq.* Stat. 34 C. F. R. 300 (1997).

Jiménez, J. E., del Rosario Ortiz, M., Rodrigo, M., Hernández-Valle, I., Ramirez, G., Estévez, A., et al. (2003). Do the effects of computer-assisted practice differ for children with reading disabilities with and without IQ-achievement discrepancy? *Journal of Learning Disabilities, 36,* 34–47.

Jorm, A. F., Share, D. L., Matthews, M., & Matthews, R. (1986). Cognitive factors at school entry predictive of specific reading retardation and general reading backwardness: A research note. *Journal of Child Psychology, 27,* 45–54.

Kavale, K. A. (1988). Learning disability and cultural disadvantage: The case for a relationship. *Learning Disability Quarterly, 11,* 195–210.

Lewis, C., Hitch, G. J., & Walker, P. (1994). The prevalence of specific arithmetic difficulties and specific reading difficulties in 9- to 10-year-old boys and girls. *Journal of Child Psychology Psychiatry, 35,* 283–292.

Lyon, G. R., Fletcher, J. M., & Barnes, M. C. (2003). Learning disabilities. In E. J. Mash & R. A. Barkley (Eds.), *Child psychopathology* (2nd ed.., pp. 520–586). New York: Guilford.

Lyon, G. R., Fletcher, J. M., Fuchs, L., & Chhabra, V. (in press). Treatment of learning disabilities. In E. Mash & R. Barkley (Eds.), *Treatment of childhood disorders.* New York: Guilford.

Lyon, G. R., Fletcher, J. M., Shaywitz, S. E., Shaywitz, B. A., Torgesen, J. K., Wood, F. B., et al. (2001). Rethinking learning disabilities. In C. E. Finn, Jr., R. A. J. Rotherham, & C. R. Hokanson, Jr. (Eds.), *Rethinking special education for a new century* (pp. 259–287). Washington, DC: Thomas B. Fordham Foundation and Progressive Policy Institute.

Lyon, G. R., & Moats, L. C. (1988). Critical issues in the instruction of the learning disabled. *Journal of Consulting and Clinical Psychology, 56,* 830–835.

MacMillan, D. L., & Siperstein, G. N. (2002). Learning disabilities as operationally defined by schools. In R. Bradley, L. Daniel-son, & D. P. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 287–368). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Mazzocco, M. M. M., & Myers, G. F. (2003). Complexities in identifying and defining mathematics learning disability in the primary school age years. *Annals of Dyslexia, 53,* 218–253.

Mercer, C. D., Jordan, L., Allsop, D. H., & Mercer, A. R. (1996). Learning disabilities definitions and criteria used by state education departments. *Learning Disability Quarterly, 19,* 217–232.

Morris, R., & Fletcher, J. M. (1988). Classification in neuropsychology: A theoretical framework and research paradigm. *Journal of Clinical and Experimental Neuropsychology, 10,* 640–658.

Morris, R. D., Fletcher, J. M., & Francis, D. J. (1993). Conceptual and psychometric issues in the neuropsychological assessment of children: Measurement of ability discrepancy and change. In I. Rapin & S. Segalovitz (Eds.), *Handbook of neuropsychology* (Vol. 7, pp. 341–352). Amsterdam: Elsevier.

Morris, R. D., Stuebing, K. K., Fletcher, J. M., Shaywitz, S. E., Lyon, G. R., Shankweiler, D. P., et al. (1998). Subtypes of reading disability: Variability around a phonological core. *Journal of Educational Psychology, 90,* 347–373.

National Center for Learning Disabilities. (2002). *Achieving better outcomes—maintaining rights: An approach to identifying and serving students with specific learning disabilities.* New York: Author.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction.* Washington, DC: National Institute of Child Health and Human Development.

Pennington, B. F., Gilger, J. W., Olson, R. K., & DeFries, J. C. (1992). The external validity of age- versus IQ-discrepancy definitions of reading disability: Lessons from a twin study. *Journal of Learning Disabilities, 25,* 562–573.

President's Commission on Excellence in Special Education. (2002). *A new era: Revitalizing special education for children and their families.* Washington, DC: Department of Education.

Reschly, D. J., Hosp, J. L., & Smied, C. M. (2003). *And miles to go ...: State SLD requirements and authoritative recommendations.* National Research Center on Learning Disabilities. Retrieved July 22, 2003, from the World Wide Web: www.nrcld.org

Reschly, D. J., Tilly, W. D., & Grimes, J. P. (1999). *Special education in transition: Functional assessment and noncategorical programming.* Longmont, CO: Sopris West.

Reynolds, C. (1984–1985). Critical measurement issues in learning disabilities. *Journal of Special Education, 18,* 451–476.

Rodgers, B. (1983). The identification and prevalence of specific reading retardation. *British Journal of Educational Psychology, 53,* 369–373.

Rogosa, D. (1995). Myths about longitudinal research (plus supplemental questions). In J. M. Gottman (Ed.), *The analysis of change* (pp. 3–66). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rourke, B. P. (1975). Brain-behavior relationships in children with learning disabilities: A research program. *American Psychologist, 30,* 911–920.

Rourke, B. P. (1993). Arithmetic disabilities specific or otherwise: A nueropsychological perspective. *Journal of Learning Disabilities, 26,* 214–226.

Rourke, B. P., & Finlayson, M. A. J. (1978). Neuropsychological significance of variations in patterns of academic performance: Verbal and visual-spatial abilities. *Journal of Pediatric Psychology, 3,* 62–66.

Rourke, B. P., Fisk, J., & Strang, J. (1986). *Neuropsychological assessment of children.* New York: Guilford.

Shalev, R. S., Auerbach, J., Manor, O., & Gross-Tsur, V. (2000). Developmental dyscalculia: Prevalence and prognosis. *European Adolescent Psychiatry, 9,* 58–64.

521

Shaywitz, S. E., Escobar, M. D., Shaywitz, B. A., Fletcher, J. M., & Makuch, R. (1992). Distribution and temporal stability of dyslexia in an epidemiological sample of 414 children followed longitudinally. *New England Journal of Medicine, 326,* 145–150.

Shaywitz, S. E., Shaywitz, B. A., Fulbright, R. K., Skudlarski, P., Mencl, W. E., Constable, R. T., et al. (2003). Neural systems for compensation and persistence: Young adult outcomes of childhood reading disability. *Biological Psychiatry, 54,* 25–33.

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

Shinn, M. R. (1998). *Advanced applications of curriculum-based measurement.* New York: Guilford.

Siegel, L. S. (1992). An evaluation of the discrepancy definition of dyslexia. *Journal of Learning Disabilities, 25,* 618–629.

Siegel, L. S., & Ryan, E. (1988). Working memory in subtypes of learning disabled children. *Journal of Clinical and Experimental Neuropsychology, 10,* 55.

Silva, P. A., McGee, R., & Williams, S. (1985). Some characteristics of 9-year-old boys with general reading backwardness or specific reading retardation. *Journal of Child Psychology and Psychiatry, 26,* 407–421.

Speece, D. L., & Case, L. P. (2001). Classification in context: An alternative approach to identify early reading disability. *Journal of Educational Psychology, 93,* 735–749.

Stage, S. A., Abbott, R. D., Jenkins, J. R., & Berninger, V. W. (2003). Predicting response to early reading intervention from verbal IQ, reading-related language abilities, attention ratings, and verbal IQ-word reading discrepancy: Failure to validate discrepancy method. *Journal of Learning Disabilities, 36,* 24–33.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. *Journal of Educational Psychology, 86,* 24–53.

Stothard, S. E., & Hulme, C. (1996). A comparison of reading comprehension and decoding difficulties in children. In C. Cornoldi & J. Oakhill (Eds.), *Reading comprehension difficulties: Processes and intervention* (pp. 93–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Stuebing, K. K., Fletcher, J. M., LeDoux, J. M., Lyon, G. R., Shaywitz, S. E., & Shaywitz, B. A. (2002). Validity of IQ-discrepancy classifications of reading disabilities: A meta-analysis. *American Educational Research Journal, 39,* 469–518.

Torgesen, J. K. (2002). Empirical and theoretical support for direct diagnosis of learning disabilities by assessment of intrinsic processing weaknesses. In R. Bradley, L. Danielson, & D. Hallahan (Eds.), *Identification of learning disabilities: Research to practice* (pp. 565–650). Mahwah, NJ: Lawrence Erlbaum. Associates, Inc.

Torgesen, J., Wagner, R., & Raschotte, C. (1999). *Test of Word Reading Efficiency.* Austin, TX: PRO-ED.

U. S. Office of Education. (1977). Assistance to states for education for handicapped children: Procedures for evaluating specific learning disabilities. *Federal Register, 42,* G1082–G1085.

Vaughn, S., & Fuchs, L. S. (2003). Redefining learning disabilities as inadequate response to instruction. *Learning Disabilities Research and Practice, 18,* 137–146.

Vaughn, S. R., Linan-Thompson, S., & Hickman-Davis, P. (2003). Response to treatment as a means of identifying students with reading/learning disabilities. *Exceptional Children, 69,* 391–409.

Vellutino, F. R. (1979). *Dyslexia: Theory and research.* Cambridge, MA: MIT Press.

Vellutino, F. R., Fletcher, J. M., Scanlon, D. M., & Snowling, M. J. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychiatry and Psychology, 45,* 2–40.

Vellutino, F. R., Scanlon, D. M., & Jaccard, J. (2003). Toward distinguishing between cognitive and experiential deficits as primary sources of difficulty in learning to read: A two year follow-up of difficult remediate and readily remediate poor readers. In B. R. Foorman (Ed.), *Preventing and remediating reading difficulties: Bringing science to scale* (pp. 73–120). Baltimore: York.

Wadsworth, S. J., Olson, R. K., Pennington, B. F., & DeFries, J. C. (2000). Differential genetic etiology of reading disability as a function of IQ. *Journal of Learning Disabilities, 33,* 192–199.

Wechsler, D. (2001). *Wechsler Individual Achievement Test* (2nd ed.). San Antonio, TX: Psychological Corporation.

Wiederholt, J. L, & Bryant, B. R. (2001). *Gray Oral Reading Test* (4th ed.). Austin, TX: PRO-ED.

Wilkinson, G. (1993). *Wide Range Achievement Test–3.* Wilmington, DE: Wide Range.

Williams, K. T., Cassidy, J., & Samuels, S. J. (2001). *Group reading assessment and diagnostic education.* Circle Pines, MN: American Guidance Services.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III.* Itasca, IL: Riverside.