RESPONSIVENESS TO INTERVENTION: AN ALTERNATIVE APPROACH TO THE IDENTIFICATION OF LEARNING DISABILITIES

Frank M. Gresham, University of California-Riverside

The process by which public schools identify students as learning disabled often appears to be confusing, unfair, and logically inconsistent. In fact, G. Reid Lyon of the National Institute of Child Health and Human Development has suggested that the field of learning disabilities is a sociological sponge whose purpose has been and is to clean up the spills of general education. Research indicates that substantial proportions of school-identified students with learning disability (LD) fail to meet state or federal eligibility criteria (Lyon, 1996; MacMillan, Gresham, & Bocian, 1998; Shaywitz, Shaywitz, Fletcher, & Escobar, 1990; Shepard, Smith, & Vojir, 1983). In discussing this situation, MacMillan and Speece (1999) noted that although this finding is not in and of itself surprising, the *magnitude* of the percentage of school-identified LD students who fail to meet eligibility criteria ranged from 52 to 70%.

It may be tempting to interpret such findings as a reflection of the failure on the part of school personnel to comply with state special education codes governing eligibility determination. Keogh (1994), however, suggested that classification has three purposes: advocacy, services, and scientific study. "Error rates" in school identification of LD students can be estimated by validating cases of schools for purposes of service delivery against criteria specified in state education codes that are relevant for scientific study.

Unlike diagnosing children with physical or sensory disabilities or those with more severe forms of mental retardation, efforts to detect students exhibiting milder disabilities such as LD or mild mental retardation (MMR) are fraught with much "error" in the sense that children meeting criteria often go undetected. Because diagnosis of these milder disabilities primarily occurs in public schools, only those children referred for assessment are at risk for formal labeling. Previous work examining students whom general education teachers referred has shown that almost half of those referred had IQ scores between 71 and 85 and an additional 16% scored below an IQ of 70 (MacMillan, Gresham, Bocian, & Lambros, 1998). Clearly, teachers perceive low aptitude students as among the most difficult to teach. When MacMillan et al. applied the current IQ cut scores recommended by the American Association on Mental Retardation (IQ < 75), they found that approximately 30% of all referred children scored below that level.

Despite the abundance of children psychometrically eligible for labeling as mildly mentally retarded, only 14% of the 43 children with IQ < 75 were classified by schools as such (MacMillan, Gresham, Siperstein, & Bocian, 1996). More germane to the current topic, 44% of these cases were labeled as LD by the schools, a finding consistent with that of Gottlieb, Alter, Gottlieb, and Wishner (1994) who found school-identified urban LD students to have a mean IQ that was substantially lower than that of suburban LD students and to resemble mildly mentally retarded students of the 1970s.

The LD category now accounts for 52% of all students with disabilities served in special education under the Individuals with Disabilities Education Act (IDEA). Between 1976–77 and 1996–97, the number of students served as LD increased from 797,213 to 2,259,000—a 283% increase. During this same period, the number of students served as MR decreased from 967,567 to 584,000, representing a 60% decrease (U.S. Department of Education, 1998). In commenting on the dramatic increase in LD, MacMillan and colleagues suggested, "Were these epidemic-like figures interpreted by the Center [sic] for Disease Control, one might reasonably expect to find a quarantine imposed on the public schools of America" (MacMillan, Gresham, Siperstein, & Bocian, 1996, p. 169).

Frankly, there is neither a completely accurate nor universally accepted explanation for these data. However, the increase in LD, in part, is attributable to school practices of classifying LD on the basis of absolute low achievement regardless of IQ level or a discrepancy between IQ and achievement—and

including in substantial numbers children who meet criteria for MMR (MacMillan et al., 1998). In fact, an analysis of current classification practices suggested the following: (a) a small minority of such children are classified as mildly mentally retarded, (b) a substantial proportion of these children are served (erroneously) in special education as LD, and (c) some unknown proportion avoid detection, are overlooked by teachers, or are not referred by teachers despite concerns about the child's academic performance (MacMillan, Gresham, Bocian, & Siperstein, 1997; MacMillan, Siperstein, & Gresham, 1996).

PARADIGMS OF LD CLASSIFICATION

The process employed by public schools can be conceptualized as consisting of three steps: (a) the decision to refer by a child's general education teacher, (b) the psychological evaluation of the child, which yields a combination of psychometric scores corresponding to criteria specified in the state as a prerequisite for eligibility, and (c) the team placement decision arrived at after review and discussion of all evidence by a school placement team. It is significant that these three steps occur in a set sequence as presented above. As a result, a student who is not referred by his or her general education teacher is not at risk for being identified as LD. Only those students passing through this first gate—referral—are even considered for psychological evaluation. In addition to the steps in the above sequence, consideration must be given to the fact that at each gate there are differences in the weighting given to various factors that result in three competing paradigms for the identification process. These factors are (a) the nature and role of professional judgment permitted at a specific gate, (b) the concept or question addressed by those involved in the decision making at a particular gate, (c) the use of local versus national norms employed at various gates, and (d) the extent to which sociocultural and contextual factors are considered (Bocian, Beebe, MacMillan, & Gresham, 1999).

Viewing the identification process through the lens of competing paradigms may serve to clarify the process by which schools identify children as LD and why there is often a gap between who is identified by schools and research criteria. The following sections expand on how each of the four factors just noted operate in concert or in competition with each of the three paradigms or gates in the identification process.

Referral

Being referred by a general education teacher is a necessary but insufficient requirement for being school-identified as LD. Although teachers refer students to prereferral teams for academic and/or behavioral difficulties, the referral issue with LD is almost always academic deficiencies. The child's academic performance relative to the modal performance of the class or the gap between the target child's reading level and that of members of the lowest reading group is more salient in reaching the referral decision than are standardized test scores. This perspective reflects what has been referred to in the literature as "teachers as imperfect tests" (Bahr, Fuchs, Stecker, & Fuchs, 1991; Gerber & Semmel, 1984; Gresham, MacMillan, & Bocian, 1997; Gresham, Reschly, & Carey, 1987). The principle guiding the teacher at this step is one of *relativity*—that is, what is the likelihood that this teacher will be able to close the gap in achievement relative to the target child's peers in both the classroom and grade level, given class size, past responsiveness of the child to intervention, and the resources available in the classroom? When the teacher concludes that this relative gap cannot be substantially narrowed without assistance, the decision to refer is highly probable.

Although the referral decision is almost never influenced by information from nationally normed scales, the decision can be and sometimes is tempered by sociocultural and contextual factors. Even in cases where the teacher judges a child's academic performance to be deficient, he or she might refrain from referring because of circumstances involving the home, the facility of both regular and special education teachers with the child's native language, or health concerns. The point here is that although local norms are employed to determine academic performance at the referral step, sociocultural and contextual factors are considerations that sometimes influence the referral decision.

Testing

It is likely that children who were referred and fail to respond to prereferral efforts will ultimately be subjected to the second gate in the referral process—psychoeducational evaluation. MacMillan and Speece (1999) characterized this gate as representing a cognitive paradigm intended to detect or document the existence of a within-child problem. It is through psychoeducational assessment that the referred child's eligibility for special education as LD is established as 98% of the states include a discrepancy in either their definition of or criteria for identifying students with LD (Mercer, Jordan, Allsopp, & Mercer, 1996).

The concept guiding the decision to pass the child through this gate and on to the school-placement team is one of *acceptability*. Through the assessment with standardized tests, one can determine whether the referred child's low level of academic performance is acceptable. If it is severely discrepant from the aptitude score, a low performance in reading is unacceptable (i.e., the child should be doing better). This situation reflects the concept of LD as unexpected underachievement. Conversely, if a child with a very low reading score performs equally low on an individually administered measure of intelligence, he or she is doing about as well as can be expected. This situation reflects the notion of expected underachievement. Finally, although teachers weigh sociocultural and contextual factors in deciding whether to refer the child, the testing step is devoid of such factors.

Team Recommendation

Multidisciplinary teams (MDTs) are responsible for determining eligibility and recommending placement. These teams are permitted to exercise judgment, but unlike the teacher in the referral step, it is a "team judgment," not an individual one. It brings together the two major players involved in referral and testing—that is, the general education teacher who referred the child and the school psychologist or educational diagnostician who performed the psychoeducational assessment.

The decision reached by the MDT reflects a considerable amount of team judgment, as opposed to individual judgment which is reflected in the referral process. The general education teacher assesses the child's academic performance relative to local norms and the school psychologist assesses the child's academic performance discrepancy relative to aptitude and national norms. However, although local norms predominate at the referral step and national norms predominate at the testing step, all three perspectives are considered by the MDT in arriving at a placement decision: local norms, national norms, and sociocultural and contextual factors.

The concept guiding the team decision regarding placement is *profitability,* which reflects the collective perception that the specific special education services provided at that school site will or will not benefit the child. As such, the anticipated profitability gauges the interaction between child characteristics (derived from the comparisons of this child's level of performance to both local and national norms) and the quality of special education services on site. Parental wishes and concerns also will factor into the ultimate decision regarding placement.

The dynamics of specific MDTs will result in assignment of differential weighting to local norms, national norms, and sociocultural and contextual factors in arriving at placement decisions. Thus, team decisions are likely to vary, even in the face of hypothetically identical information, because of the relative forcefulness of particular players serving on the team. Any effort to understand school-identified LD students must consider the importance of these three steps (referral, testing, and team recommendation) and the relative weighting given to available data at each step.

Implications of Competing Paradigms in LD Identification

Presently, research on LD students often examines a group of students who are screened according to criteria for only one of the gates. For example, a "sample" will sometimes be selected from children with a certain psychological profile reflective of the testing gate even though a referring teacher did not initially screen the sample. Such sampling results in a group that overlaps with but is not identical with children

who will be school-identified as LD. Findings over the past 15 years have pointed out the lack of consistent definition in policy or practice in the identification of LD students, a circumstance that has been a major stumbling block to effective research and practice (Lyon, 1996). Response to this challenge has ranged from impugning the concept of LD as neither valid nor instructionally relevant, to criticizing teachers and schools for failing to implement criteria correctly. Some researchers have suggested that schools seek flexibility and the opportunity to exercise professional judgment rather than being held to a rigid code of precise formulas (Keogh & Speece, 1996; MacMillan et al., 1997; McLeskey & Waldron, 1991).

A second implication of the competing paradigm model is the accuracy with which teachers identify within-child variables relevant to the classroom that are later validated by psychoeducational assessments. Teachers' accurate evaluations of students' abilities should be sought after rather than continually challenged. Teachers may be "imperfect tests," but in terms of classroom relevance, their perceptions often outrank students' performances on psychoeducational assessments on isolated tasks conducted under ideal, pristine conditions.

A third implication is recognizing the severe limitations and the ability of the discrepancy concept of LD to both plan instruction and identify students for early intervention. The recent national downward trend in reading achievement and the public pressure for student outcomes and accountability have led to an enhanced focus in the field on reading disabilities (Lyon, 1996). This approach surely holds more promise for students and teachers alike, particularly given the ability of teachers to identify reading disabilities based on curriculum-achievement discrepancy or an achievement discrepancy relative to peers. Perhaps of greater import is the need to train and encourage teachers to exercise their judgment at even earlier points in a student's career. The research field should work to validate that judgment with operationalized criteria, particularly with reading problems.

Although the competing paradigm multiple-gate system now in place does work to identify students in need of services, the competition between expensive, time-consuming assessments at three different steps could be streamlined and articulated in a fashion more respectful of both teacher and school professional judgment to meet students' need for immediate intervention services. The most serious flaw in the current process is the absence of a direct link between assessment procedures used for identification and subsequent interventions that might be prescribed based on these assessment procedures (i.e., treatment validity). In fact, it is clear that most reading difficulties exhibited by students now classified as LD are caused by inadequate literacy experiences, inadequate instruction, or some combination of both (Clay, 1987; Vellutino, Scanlon, & Tanzman, 1998). This being the case, an alternative approach to the identification of students with LD is justified. Therefore, the focus of the current paper is to describe how such an assessment process can be developed and used in identifying and instructing students with LD.

Definitions of LD and the Discrepancy Approach

The purpose of this section is to provide a very brief overview of the history of and difficulties in defining LD and some of the issues inherent in using a discrepancy approach to operationalize the LD construct. Other chapters in this book provide a much more detailed analysis of the issues involved in the definition of LD. This overview is intended to provide a context for discussing a different approach to LD definition: *responsiveness to validated intervention procedures.*

Brief Recent History of LD

Kirk (1962) first used "learning disabled" to describe a group of children who have retardation, disorder, or delayed development in one of more of the processes of speech, language, reading, writing, arithmetic, or other school subjects. This definition was the first to introduce the concept of psychological process disorders and how these processing deficits adversely affect academic achievement (Kavale & Forness, 2000). Shortly thereafter, Bateman (1965) proposed the notion of underachievement as a fundamental aspect of LD. In Bateman's definition, the idea of an "educationally significant discrepancy" between intellectual potential and actual level of academic performance was emphasized. This definition did not specify what constituted an "educationally significant discrepancy" and did not provide information on how

to measure intellectual potential and actual level of performance (Kavale & Forness, 2000). More than three decades later, the field of LD still has not arrived at a consensus in terms of resolving these definitional and measurement issues.

Rutter and Yule (1975) defined two types of reading underachievement difficulties: general reading backwardness (GRB) and specific reading retardation (SRR). GRB is defined as reading below the level expected of a child's chronological age, whereas SRR is defined as reading below the level predicted from a child's intelligence. Rutter and Yule estimated the prevalence of GRB in the school-age population to be 7% and 20% (rural and inner-city settings, respectively), whereas the prevalence rate of SRR was 4% and 10%, respectively. It should be noted that, according to Hinshaw (1992), almost all children with SRR could be classified as GRB, but only half of children with GRB are classifiable as SRR.

Children such as those described by Rutter and Yule (1975) as having SRR may be considered as having LD in most states using a discrepancy-based definition of LD (Mercer et al., 1996). In fact, the prevalence rate of SRR of 4–10% in Great Britain is consistent with the 5% prevalence rate of children served as LD in the United States. Moreover, children who might be described as low achievers might meet the definition of GRB. SRR and GRB capture the concepts of unexpected and expected reading underachievement, respectively.

Issues in Defining LD: The LD/LA Disputes

Differentiation among groups of children having mild disabilities such as LD and MMR as well as low achievement (LA) has always been problematic. Children functioning around the margin of what might be considered a disability group create special problems in assessment, measurement, and eligibility determination for special education programs. At what point, for instance, is low academic achievement considered to be due to MMR and not to LD? How is MMR different from LA? Is LD different from LA, and if so, how is it different? Are LD and LA primarily reflective of differences in degree or kind of academic underachievement? Although these questions remain fundamental to the identification of students having difficulties in school, definitive answers to these questions have not been forthcoming.

Researchers have debated the similarities and differences between students classified as LD (discrepant low achievers) and those classified as LA. The heart of these debates centers on the degree to which LD can be differentiated from LA and the extent to which distributions of these groups' intellectual, academic achievement, and social behavior functioning overlap (Epps, Ysseldyke, & McGue, 1984; Fuchs, Mathes, Fuchs, & Lipsey, 2001; Kavale, Fuchs, & Scruggs, 1994; Ysseldyke, Algozzine, Shinn, & McGue, 1982). Perhaps the most widely cited study in this debate was reported by Ysseldyke et al. (1982) in which school-identified children with LD were compared to a group of LA children on a variety of psychoeducational measures. This study suggested that LD could not be differentiated from LA, with 96% of the scores on psychoeducational measures being in a common range. Ysseldyke et al. argued that LD and LA are essentially identical constructs, and they questioned the diagnostic validity of the term "learning disabilities."

Kavale et al. (1994) criticized the interpretation and analyses of Ysseldyke et al. (1982), indicating that the data had been misused to support policies from the Regular Education Initiative. Kavale et al. reanalyzed Ysseldyke et al.'s original data, using a meta-analytic statistic (Cohen's d) that compares the means of each group relative to the groups' variability (pooled standard deviation [SD]). On the basis of 44 comparisons, Kavale et al. showed that 63% of the LD group could be differentiated from the LA group (effect size = 0.338), with 37% overlap between the groups. This 37% overlap figure is substantially less than the 96% overlap reported by Ysseldyke et al. With respect to academic achievement, almost 80% of the LD group could be differentiated from the LA group, with LD children scoring lower than the LA group.

The results of the Connecticut Longitudinal Study added further fuel to the debate concerning the differentiation of LD and LA (Shaywitz, Fletcher, Holahan, & Shaywitz, 1992). This investigation compared children with LD (defined as a 22-point discrepancy between aptitude and reading achievement) with low achievers (defined as children scoring below the 25th percentile in reading, but who did not show

a severe discrepancy). Using a variety of child-, teacher-, and parent-based measures, these authors found more similarities than differences between LD and LA groups, suggesting that both groups could be considered eligible for special education services.

The separate analyses and interpretations of the same data set by Ysseldyke et al. (1982) and Kavale et al. (1994), coupled with the longitudinal study by Shaywitz et al. (1992), leave a fundamental question unresolved: Are LD and LA quantitatively or qualitatively different? The studies and analyses by Ysseldyke et al. and Shaywitz et al. suggest that LD and LA groups are more alike than different. The analyses by Kavale et al. suggest these groups are more different than alike, particularly in the area of academic achievement. Researchers and practitioners are left with the decision of deciding which group's analyses and conclusions to believe. This distinction is important, given that important educational decisions are made for children with these characteristics and that these decisions have rather substantial economic and legal consequences for school districts.

Recently, a meta-analysis of 79 studies on this topic was completed by Fuchs et al. (2001), the purpose of which was to determine whether LD and LA reflect differences in *degree* of underachievement or differences in *kind* of underachievement. That is, is LD *quantitatively* different or *qualitatively* different from LA in terms of reading achievement? On the basis of 112 effect sizes, the mean weighted effect size was 0.61 (95% CI: 0.56 to 0.65); however, there was considerable heterogeneity among studies concerning the magnitude of differences in reading between LD and LA groups.

Fuchs et al. (2001) interpreted the 0.61 effect size as being large, thereby suggesting that LD could be differentiated from LA (LD < LA in reading achievement). However, this 0.61 effect size translates into only a 9-point ($M = 100$, $SD = 15$) standard score difference between LD and LA groups. In fact, Cohen and Cohen (1983) would define a 0.61 effect size as moderate and a large effect size as being 0.80 or greater. Assuming a median reliability coefficient of 0.90 for reading domain measures used in calculating effect sizes and a standard error of measurement of 4.74 ($SD = 15$), a 95% confidence interval calculates to +9.48 standard score points. Clearly, this difference is not large, particularly when taking into account measurement error of the dependent measures.

It is difficult to make the case that a standard score difference which is within the range of measurement error represents a substantial difference in kind rather than degree and therefore somehow validates the LD construct. Certainly, these data do not support a two-groups approach to LA like that found in the field of mental retardation (Zigler, Balla, & Hodapp, 1984). For the sake of argument, the average IQ scores of students with MMR is around 70 and the average IQ score of students with profound mental retardation is about 25. Few would argue that these two groups do not differ in *kind* on a number of variables such as identification prior to school entry, severe deficits in independent functioning, and frequent comorbid biomedical conditions (MacMillan, Gresham, & Siperstein, 1993).

The Fuchs et al. (2001) meta-analysis suggested that a standard-score point difference of 9 (0.61 SD) was sufficient to conclude that LD students differ in kind from LA students, particularly on timed reading tasks reflecting deficits primarily in automaticity of reading skills. By comparison, in the area of sensory disabilities, there are clear distinctions between hearing impaired and deaf as well as visually impaired and blind based on rather substantial differences in the magnitude of hearing and visual loss, respectively. Clearly, the field of LD must be able to present more convincing evidence to conclude that LD students differ in kind from LA students and thus legitimately deserve special education and related services based on this minimal difference.

IQ-Achievement Discrepancy and LD Definition

There have been a variety of ways of operationalizing the LD construct using some variation of a discrepancy-based notion. Berninger and Abbott (1994) suggested that four major methods have been used to compute discrepancy; all of which have measurement difficulties: (a) deviation from grade level, (b) expectancy formulas, (c) simple standard-score difference, and (d) standard regression analysis. A *deviation from grade-level* approach makes the fallacious assumption that all students should be

functioning on grade level. Of course, this assumption ignores the most fundamental principle of standardized achievement tests: In a normal distribution of test scores, half the students will be above level and half will be below grade level. How far below grade level one must be to qualify for LD using this approach is influenced by a variety of factors such as level of intelligence, socioeconomic status of the school, and measurement problems with grade-equivalent scores.

Another approach is to compare a child's expected and observed grade level in an academic area controlling for IQ (*expectancy approach*). To determine this discrepancy, this approach uses grade-equivalent scores which vary greatly across grade levels in terms of the raw scores underlying these scores and are not comparable across test instruments (Berninger & Abbott, 1994). A third approach uses standard-score differences between IQ and achievement measures (sometimes called the *simple difference method*) to quantify LD. This approach, however, does not account for measurement error in IQ and achievement measures, the unreliability of difference scores, and the attendant effects of regression toward the mean. In a final method, the *regression discrepancy approach,* the measurement errors using the simple difference method are accounted for by calculating aptitude-achievement discrepancies using the parameters of reliability of aptitude, reliability of achievement, and reliability of aptitude-achievement difference scores (Reynolds, 1984). This approach, like the standard-score difference approach, assumes that IQ is the exclusive and self-limiting determinant of achievement.

The aforementioned approaches to quantifying LD have been used to qualify students for special education and related services. Each method, as briefly reviewed, has a number of conceptual and statistical drawbacks. A major controversy in discrepancy-based notions of defining LD is the central importance assigned to IQ tests in this process (Gresham & Witt, 1997; Kavale & Forness, 2000; Siegel, 1989). Perhaps the most important criticism of IQ tests is that they contribute little reliable information to the planning, implementation, and evaluation of instructional interventions for children and youth. Moreover, according to the research contrasting LD and LA populations, IQ tests are not particularly useful in diagnostic and classification purposes for students with mild learning problems. What appears to be needed is an approach to defining LD based on how students respond to instructional interventions rather than some arbitrarily defined discrepancy between ability and achievement.

RESPONSIVENESS TO INTERVENTION

Historical Background: Aptitude H Treatment Interaction

The notion of alternative responsiveness to intervention is not a new concept in the field of education and psychology. In his presidential address to the American Psychological Association, Cronbach (1957) called for the integration of correlational and experimental disciplines of scientific psychology by using the concept of aptitude H treatment interactions (ATIs). ATI research focuses on the measurement of valid aptitudes (characteristics or traits) and how these aptitudes interact with various treatments (instructional methods or types of therapy). ATI research originally attempted to provide a hybrid science spliced from the study of individual differences (aptitudes) and experimental psychology (treatments). Interactions occur when treatments or instructional methods have different effects on persons known to differ in measured aptitudes or characteristics.

Cronbach and Snow (1977) defined an aptitude as any characteristic of a person that predicts the probability of success under a particular treatment condition. These characteristics or aptitudes theoretically can be anything ranging from test-derived aptitudes (verbal-spatial, fluid-crystallized, field dependent-independent) to physical variables (right versus left hemispheric functioning, temporal versus frontal lobe damage). Treatments are defined as any manipulable variable such as instructional method, type of psychotherapy, classroom climate, and so on.

The fundamental logic of ATIs is the matching of instructional treatments to aptitudes. The basic rationale for this matching is based on the belief that learners having strengths in some aptitudes will respond better to treatments capitalizing on these aptitude strengths. Whereas Cronbach and Snow (1977) suggested that

aptitudes and treatments could be matched in several ways (capitalization, compensation, and remediation), most ATI matching studies have been based on capitalization, which adapts instruction to the abilities of the student. For example, students high in verbal comprehension might be expected to learn more under verbal instruction rather than visual instruction.

At its most basic level, an ATI study must have at least two aptitudes and two treatments and thus four data points. For example, one could use scores from the Wechsler Intelligence Scale for Children III (WISC-III) Verbal (Verbal Comprehension) and Performance (Perceptual Organization) scales to define Verbal and Visual learners, respectively. These scores would represent two aptitudes. One could also use phonics and whole-word approaches to reading instruction to define two treatments. To demonstrate an ATI, one could show that Verbal learners respond better to phonics instruction than Visual learners and Visual learners respond better to whole-word instruction than Verbal learners. This example is a *disordinal ATI* and this logic is employed most frequently by school psychologists and special educators to make instructional recommendations based on cognitive ability or aptitude measures (Gresham & Witt, 1997; Reschly & Ysseldyke, 1995). In an ordinal ATI, there is a larger effect on one treatment for one aptitude, but no differences between the two aptitude groups for the other treatment. For instance, phonics may be more effective for Verbal learners with no differences between Verbal and Visual learners using the whole-word treatment.

From a logical perspective, we have every reason to expect that many ATIs exist in teaching students with LD. Ostensibly, "verbal" learners should learn more efficiently and effectively under verbal instruction and "visual" learners should learn more efficiently and effectively under visual instruction. Unfortunately, there is little empirical support for the differential prescription of treatments based on different abilities or aptitudes like these and others found in the literature. This lack of support continues to surprise many professionals who interpret test results and recommend treatments based on the presumption of largely mythical ATIs (Gresham & Witt, 1997; Reschly & Ysseldyke, 1995).

Brief Overview of ATI Research

A comprehensive review of the ATI research literature is far beyond the scope of the current paper; however, a number of reviews of this literature support the unfeasibility of matching aptitudes to treatments for children with LD or other learning difficulties. Comprehensive reviews of the modality matching literature (Arter & Jenkins, 1979; Kavale & Forness, 1987, 1995; Ysseldyke & Mirkin, 1982) fail to consistently show significant ATIs. Studies and reviews conducted within the cognitive style/processing literature fail to consistently demonstrate ATIs (Ayres & Cooley, 1986; Ayres, Cooley, & Severson, 1988; Das, 1995; Das, Naglieri, & Kirby, 1995; Good, Vollmer, Creek, Katz, & Chowdhri, 1993).

Finally, the use of a neuropsychological model within ATI research focuses on inferred brain strengths or functioning. For instance, a child having left hemispheric strength might be presumed to learn more efficiently using methods that capitalize on this strength (e.g., phonics, verbally presented material) whereas children with right hemispheric strengths might perform better using other methods (e.g., holistic, visually presented material). Despite the proliferation of this ATI logic in the neuropsychological literature (see D'Amato, Rothlisberg, & Work, 1999; Hynd, 1989; Reynolds & Fletcher-Jantzen, 1989), I was unable to locate a single, methodologically sound empirical study demonstrating a significant ATI based on neuropsychological assessment, interpretation, and treatment with children having mild learning problems. In fact, reviews by Reschly and Gresham (1989) and Teeter (1987, 1989) question the entire enterprise of applying ATI logic in neuropsychological assessment practices to children with mild learning problems.

Considering the disappointing results of ATI studies using modality matching, cognitive style/processing, and neuropsychological assessment, there is little, if any, empirical support for prescribing different treatments based on the assessment of different aptitudes. Cronbach (1975) expressed his frustration with ATI research by stating: "Once we attend to interactions, we enter a hall of mirrors that extends to infinity" (p. 119). Abandoning the quest for ATIs, Cronbach (1975) suggested context-specific evaluation and short-run empiricism: "One monitors responses to treatment and adjusts it" (p. 126). The approach recommended

by Cronbach forms the conceptual basis for *responsiveness to treatment* as the criterion in making LD eligibility determinations. Yet before describing specific research using this approach for students with LD, I provide a conceptual basis for responsiveness to intervention in the following section.

Responsiveness to Intervention Defined

Responsiveness to intervention can be defined as the change in behavior or performance as a function of an intervention (Gresham, 1991). The concept of responsiveness to intervention uses a discrepancy-based approach; however, the discrepancy is between pre- and postintervention levels of performance rather than between ability and achievement scores. Given that a goal of all interventions is to produce a discrepancy between baseline and postintervention levels of performance, the failure to produce such a discrepancy within a reasonable period (an inadequate response to *intervention*) might be taken as partial evidence for the presence of an LD. Responsiveness to intervention has received a great deal of attention over the past 25 years in the experimental analysis of behavior literature (see Nevin, 1988, 1996 for comprehensive reviews).

In an analogy to Newtonian physics, Nevin (1988) used the term *behavioral momentum* to explain a behavior's resistance to change. That is, a moving body possesses both mass and velocity and will maintain constant velocity under constant conditions. The velocity of an object will change only in proportion to an external force and in inverse proportion to its mass. Considering the momentum metaphor, an effective intervention ("force") results in a high level of momentum ("responsiveness") for the behavior in question (e.g., learning to read).

For example, a reading intervention designed to produce oral reading fluency would be considered successful if it produced reading fluency rapidly and reliably during intervention and if reading fluency persisted after the intervention is withdrawn. In contrast, if oral reading fluency deteriorated after the intervention is withdrawn, teachers would not be satisfied with the rate of oral reading fluency no matter how well a student read during intervention. Also, if oral reading performances occurred at low rates with numerous errors (omissions, substitutions) during intervention, teachers would likely conclude that the student had not established automaticity in oral reading and would seek to extend, intensify, or change the reading instruction.

In the field of LD, the goal for all students is to facilitate the momentum of academic performances, primarily in reading. One can conceptualize *response to intervention* as being determined by response strength ("momentum") in relation to an intervention implemented to change behavior ("external force"). Most children at risk for LD exhibit poor performances in the area of reading (e.g., poor fluency, lack of phonological awareness). That is, their reading behavior has low velocity, which does not change when they are exposed to typical reading instruction in the general education classroom. A *response to intervention* approach to eligibility determination identifies students as having an LD if their academic performances in relevant areas do not change in response to a validated intervention implemented with integrity.

As we shall see later, much sound empirical work has been done on the idea of identifying treatment-adequate and -inadequate responders to intervention in the field of reading disabilities (Fuchs, Fuchs, & Hamlett, 1989a; Vellutino et al., 1996, 1998). The following section describes the concept of treatment validity and how it can be incorporated into the notion of responsiveness to intervention.

Treatment Validity

Treatment validity (sometimes referred to as treatment or instructional utility) is the degree to which any assessment procedure contributes to beneficial outcomes for individuals (Cone, 1989; Hayes, Nelson, & Jarrett, 1987). Although the concept of treatment validity evolved from the behavioral assessment camp, it shares several characteristics and concepts found in the traditional psychometric literature: (a) Treatment validity contains an aspect of Sechrest's (1963) notion of *incremental validity* in that it requires an assessment procedure to improve prediction over and above existing procedures; (2) treatment validity

contains the idea of utility and cost-benefit analysis that is common in the personnel selection literature (Mischel, 1968; Wiggins, 1973); and (c) treatment validity is related to Messick's (1995) evidential basis for test interpretation and use, particularly as it relates to construct validity, relevance/utility, and social consequences. It is possible for a particular test interpretation to have construct validity, but have little or no relevance or utility for a particular use of that test (e.g., recommendations for treatments based on the test). Finally, as previously noted, the whole idea behind ATI research is based on the notion of treatment validity, the matching of instructional treatments to aptitudes.

The ATI literature on modality matching, cognitive style/processing, and neuro-psychological assessment provides little evidence that the information gathered about aptitudes results in "incremental advance information" that helps in recommending instructional interventions for students with learning difficulties. More than 15 years ago in a review in the *Buros Mental Measurement Yearbook* of the Wechsler Intelligence Scale for Children-Revised (WISC-R), Witt and Gresham (1985) wrote: "The WISC-R lacks treatment validity in that its use does not enhance remedial interventions for children who show specific academic skill deficiencies… For a test to have treatment validity, it must lead to better treatments (i.e., better educational programs, teaching strategies, etc.)" (p. 1717). This statement could be extended to *all* cognitive ability measures based primarily on their inability to inform or guide instructional interventions (Gresham & Witt, 1997; Reschly & Grimes, 1995). Voicing a similar sentiment regarding using IQ tests in the diagnosis of reading disability, Share, McGee, and Silva (1989) argued:

> It may be timely to formulate a concept of reading disability that is independent of IQ. Unless it can be shown to have some *predictive value for the nature of treatment outcomes*, consideration of IQ should be discarded in discussions of reading difficulties. (p. 100, emphasis added)

In describing the value of using a treatment validity criterion in the field of LD, Fuchs and Fuchs (1998) suggested that this approach focuses on maximizing regular education's potential effectiveness for *all* students. Judgment about the need for special education is reserved until the effects of instructional adaptations have been assessed in the regular classroom *and* data verify that a special education program would enhance learning. One promising assessment approach that meets the treatment validity criterion and can be used to make eligibility decisions is *curriculum-based measurement* (CBM) (Fuchs & Fuchs, 1997, 1998; Reschly & Grimes, 1995; Shinn, 1995).

Support for a Treatment Validity Approach

There is a great deal of empirical support for adopting a treatment validity approach rather than a discrepancy-based approach to defining LD (Clay, 1987; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Fuchs & Fuchs, 1997, 1998; Torgesen et al., 2001; Vellutino, Scanlon, & Lyon, 2000; Vellutino et al., 1996, 1998). Vellutino et al. (1996) noted that the discrepancy approach to defining LD does not screen out those children whose reading difficulties might be due to either inadequate schooling or limited exposure to effective reading instruction. Instead, Vellutino et al. argued for using exposure to intensive reading instruction as a "first-cut" diagnostic aid in distinguishing between reading problems caused by cognitive deficits versus those caused by experiential deficits (poor or inadequate reading instruction).

Vellutino et al. (1996) conducted a longitudinal study of 183 kindergarten children composed of poor readers (*n* =118) and normal reader controls (*n* = 65). Poor readers were selected on the basis of scoring below the 15th percentile on measures of word identification or letter-sound correspondences using nonsense words. Children in the poor reader group (a subsample of 74 children) were given daily one-to-one tutoring (30 minutes per day) for a total of 15 weeks over 70–80 sessions (35–40 hours of tutoring). Using hierarchical linear regression analyses, Vellutino et al. calculated growth rates for each child from kindergarten to second grade. Slopes from these analyses were rank-ordered and used to place children into 1 of 4 groups: Very Limited Growth (VLG), Limited Growth (LG), Good Growth (GG), and Very Good Growth (VGG). Approximately half of the sample showed VLG (26%) or LG (24%).

If one accepts the proposition that "difficult to remediate" children can be considered LD and easily

remediated children are not LD, then the entire questionable process of calculating ability-achievement discrepancies can be summarily abandoned. Vellutino et al. (1996, 2000) showed that IQ-achievement discrepancy scores did not reliably distinguish between disabled and nondisabled readers, did not distinguish between difficult-to-remediate (VLG and LG) and readily-remediated (VGG and GG) students, and did not predict response to remediation. In short, IQ-achievement discrepancy scores did not have treatment validity.

Requirements for Adopting a Treatment Validity Approach

Adopting a treatment validity approach to the identification of students with LD has several technical requirements. These requirements include (a) ability of measures to model academic growth (Burchinal, Bailey, & Synder, 1994; Fuchs & Fuchs, 1997, 1998; Vellutino, Scanlon, & Tanzman, 1998; Vellutino et al., 1996), (b) availability of validated treatment protocols (Berninger & Abbott, 1994; Torgesen et al., 2001), (c) capability of distinguishing between ineffective instruction and unacceptable individual learning (Fuchs & Fuchs, 1997, 1998), (d) suitability in informing instructional decisions (Fuchs & Fuchs, 1997, 1998; Vellutino et al., 1996, 1998; Witt & Gresham, 1997), and (e) sensitivity to detection of treatment effects (Fuchs & Fuchs, 1997; Marston, Fuchs, & Deno, 1986; Marston, 1987–88; Vellutino et al., 1996, 1998). Each of these requirements for treatment validity will be described in the following sections.

Ability to Model Academic Growth

All intervention investigations attempt to determine whether a change in a dependent variable is due to systematic and controlled changes in an independent (treatment) variable. Traditionally, this question has been addressed using a pretest/posttest design in which an experimental (treatment) and a control group are measured before and immediately after intervention. The effects of treatment in such designs are evaluated by comparing pretest and posttest scores using either repeated measures analysis of variance (ANOVA) or analysis of covariance (ANCOVA, using pretest scores as covariates), or by computing simple differences for groups between pretest and posttest scores (Kirk, 1994). Although these types of analyses can tell us whether or not a given treatment produced mean differences on a dependent variable relative to a control group, these analyses do not supply enough data to model individual change adequately (Burchinal et al., 1994).

A viable alternative to traditional pretest/posttest design comparisons is the use of growth curve analysis (GCA) using hierarchical linear models as a means of modeling academic growth. GCA is used to address three fundamental research questions (Bryk & Raudenbush, 1987; Burchinal et al., 1994). First, GCA is used to determine patterns of change for both individuals and groups. A common example is physicians charting height and weight of children to assess whether or not a child is displaying adequate growth compared to a matched reference group. Second, GCA is used to determine if certain groups show different patterns of change over time. For example, children exposed to a reading intervention emphasizing phonological awareness might be compared to a similar group of children receiving a reading program focusing on orthographic skills. Comparisons between these two groups would be expressed in terms of differences in rate (slopes) and level (intercepts) of change. Third, GCA is often used to study the correlates of change. For instance, a researcher might be interested in contrasting the patterns of change for LD and LA groups who receive the same reading intervention. In addition, the researcher may want to assess whether background characteristics (e.g., gender, ethnicity, socioeconomic status, IQ) moderate these patterns of change over time.

Several assumptions must be met in using GCA to model academic growth (Bryk & Raudenbush, 1987; Burchinal et al., 1994): (1) Growth parameters are assumed to be normally distributed and measured on either an interval or ratio scale; (2) dependent measures are expressed in the same units of measurement over time; (3) structure of the dependent variable does not change over time; (4) each group being compared has homogeneous variances (homogeneity of variance), and (5) an adequate model of change, whether it be linear, quadratic, or cubic, has been selected and fit to the data to model patterns of growth. It should be noted that GCA does not require the same data collection design for each participant in a study;

that is, some individuals may be measured 4 times, others 6 times, and still others 8 times. Moreover, spacing between data collection points for each individual does not need to be equal. In short, GCA allows for a broader representation of the effects of an intervention on growth and is extremely flexible with respect to the number and timing of observations across research participants (Bryk & Raudenbush, 1987; Burchinal et al., 1994).

Fuchs and Fuchs (1998) describe the use of CBM as a promising measurement tool for modeling academic growth within the special education eligibility determination process. CBM meets many of the assumptions of GCA in that it provides equal scaling of the dependent variable for individuals over time, it measures the dependent variable on an interval scale, and the structure of the dependent variable remains constant over time. Use of the CBM model in LD eligibility determination will be described in detail in a subsequent section of this paper.

Validated Treatment Protocols

In order to adopt a responsiveness-to-intervention approach, validated treatment protocols must be implemented for students who might be considered learning disabled. Within both the general education and special education classroom, this may be a daunting task. For example, general education teachers often are not prepared to deal with the normal variation among students in the acquisition of reading and writing skills (Berninger, Hart, Abbott, & Karovsky, 1992). Moreover, a survey of state departments of education revealed that only 29 states require elementary teachers to take academic coursework in reading and no states require coursework in writing (Nolen, McCutchen, & Berninger, 1990). Many students classified as LD may fail to acquire basic academic skills *not* because of some underlying processing disorder, but rather because they have not been given adequate opportunities to learn. There is ample reason to believe that most reading difficulties (and children subsequently labeled as LD) are caused by woefully inadequate preliteracy experience, inadequate instruction, or some combination of both (Vellutino et al., 1996, 1998).

A number of validated treatment protocols can be used to differentiate *adequate* from *inadequate* treatment responders. Recently, Torgesen et al. (2001) compared two carefully designed instructional approaches to facilitate academic growth in reading for 8- to 10-year-old children. One intervention was the Auditory Discrimination in Depth (ADD) program that emphasized discriminations among phonemes, monitoring/representation of sound sequences in spoken syllables, and self-monitoring skills (Lindamood & Lindamood, 1998). The second intervention was Embedded Phonics (EP), which provided direct, explicit instruction in word-level reading skills and providing extensive opportunities to read and write meaningful text (Torgesen et al., 2001). The ADD and EP programs differed in depth and extent of instruction in phonemic awareness and phonemic decoding skills. Both the ADD and EP programs were provided to students on a 1:1 basis, in two 50-minute sessions, 5 days per week for 8–9 weeks and students were assessed at 1-and 2-year followups. Hours of intensive reading instruction for the ADD and EP groups totaled 67.5. Following training, all students received 8 weeks of generalization training consisting of a single 50-minute session each week.

The results of the Torgesen et al. study showed that the ADD and EP programs were equally effective in remediating reading difficulties based on the Woodcock-Johnson Broad Reading Cluster score (slope effect sizes = 4.4 and 3.9, respectively). In fact, these interventions "normalized" the reading skills of approximately one half to two thirds of the students, depending on the outcome measure used. Scores on reading comprehension (Woodcock-Johnson Passage Comprehension) were even better with 80–85% of students performing in the average range. About 40% of the students in this investigation were returned full-time to the general education classroom and were no longer considered in need of special education. Torgesen et al. concluded:

> …the similarities in growth rate of the ADD and EP conditions in our study suggest that given the
> right level of intensity and teacher skill, it is possible to obtain these rates of growth via a variety
> of approaches to direct instruction in reading. We might even suggest that these rates could serve

as a benchmark for "reasonable progress" in reading for students receiving remedial instruction in both public and private settings... [T]hey are clearly much higher than is typically achieved in most current special education settings. (p. 52)

The Torgesen et al. investigation provides insight into how we might define *inadequate responders* based on the responsiveness-to-intervention concept. Approximately 25% of students in this investigation were nonresponders to the intensive reading interventions with mean standard scores of about 70 on Word Attack, Word Identification, and Comprehension. Similarly, the Vellutino et al. (1996) study described earlier suggested that approximately 25% of students exposed to an intensive reading intervention of 37.5 hours showed VLG on measures of word identification and phonological skills. In using this resistance-to-intervention notion to diagnose reading disabilities, Vellutino et al. stated:

...to render a diagnosis of specific reading disabilities in the absence of early and labor-intensive remedial reading that has been tailored to the child's individual needs is, at best, a hazardous and dubious enterprise, given all of the stereotypes attached to this diagnosis... [O]ne can increase the probability of validating the diagnosis if one combines impressions and outcomes derived from early, labor-intensive, and individualized remediation with results of relevant psychological and educational testing in evaluating the etiology of a child's difficulties in learning to read. (p. 632)

Additional information on what constitutes a *validated treatment protocol* can be found in a recent meta-analysis by Swanson and Hoskyn (1999) who summarized 180 intervention studies for students with LD. Interventions were classified into one of four categories: (a) Direct Instruction (DI), (b) Strategy Instruction (SI), (c) Combined DI+SI, and (d) non-DI/non-SI. Swanson and Hoskyn (1999) defined DI as interventions that used fast-paced instruction in small groups; presented well-sequenced, highly focused lessons; provided numerous opportunities to respond; gave frequent performance feedback on accuracy and responses; and used frequent on-topic questions regarding academic material (Englemann & Carnine, 1991; Kame'enui, Jitendra, & Darch, 1995; Lovett, Borden, DeLuca, Lacerenza, Benson, & Brackstone, 1994; Slavin, 1987).

Studies were categorized as SI if they met the following three criteria: (a) They provided elaborate explanations of material (e.g., explanations, elaborations, and plans directing task performance), (b) they used modeling from teachers which included verbal modeling, questioning, and demonstration, and (c) they incorporated prompts or reminders or multiprocess instructions and dialogue between teachers and students (Borkowski & Turner, 1990; Graham & Harris, 1996; Levin, 1986; Pressley & Ghatala, 1990; Rosenshine, 1995). Finally, studies meeting both DI and SI criteria were categorized as Combined DI+SI and studies meeting neither of these criteria were classified as non-DI/non-SI.

On the basis of these 180 studies, a total of 1,537 effect sizes were calculated comparing LD students in the treatment groups with LD students in control groups. Overall, the mean effect size was 0.79 (*SD* = 0.52). Swanson and Hoskyn (1999) described the typical intervention study as including 22.47 minutes of daily instruction, 3.58 times per week, over 35.72 sessions. On average, students received 80 minutes per week over almost 10 weeks of intervention, or approximately 13.3 hours of instruction. With respect to the type of intervention, the Combined DI+SI group had greater effect sizes (*M* = 0.81) than the DI alone (*M* = 0.77), SI alone (*M* = 0.67), and non-DI/non-SI (*M* = 0.62) interventions. There were no significant differences among these latter three intervention groups. Interestingly, studies producing the largest effect sizes reported only minimal discrepancies between IQ and reading achievement (*M* = 0.95) supporting the questionable use of the IQ-achievement discrepancy in predicting responsiveness to intervention described by Vellutino et al. (1998). Also, interventions were less effective with students having reading scores slightly higher than their IQ scores (reading scores > 90 and IQ 85–90).

Swanson and Hoskyn's (1999) meta-analysis suggests that there are several validated intervention approaches in reading for students with LD with effect sizes from 0.58 to 0.81. The Combined DI+SI interventions produced a large effect size (0.81) which indicates that 80% of students in the intervention groups had reading scores equal to or greater than students in control groups. This effect size, however, is substantially lower than those reported by Torgesen et al. (2001) and Vellutino et al. (1996). The lower

effect sizes reported by Swanson and Hoskyn may be due, in part, to differences in the intensity of treatment. Torgesen et al. provided 67.5 hours of instruction over 8 weeks and Vellutino et al. provided 35–40 hours of instruction over 15 weeks. The prototypical intervention in the Swanson and Hoskyn meta-analysis provided only 13.3 hours of instruction over approximately 10 weeks. Regardless of these effect size differences, a substantial body of empirical research supports the validity of treatment protocols for remediating reading deficiencies of students with LD.

Distinguishing Between Acquisition and Performance Deficits

An important decision in using a responsiveness-to-intervention approach to defining LD is the differentiation of skill (acquisition) deficits from performance (motivational) deficits. *Skill deficits* refer to the absence of an academic skill in a student's repertoire ("can't do" problems) and performance *deficits* describe a lack of motivation to perform a given academic skill ("won't do" problems). Skill deficits most often result from inadequate, insufficient, or inappropriate instruction whereas performance deficits result from inadequate, insufficient, or inappropriate arrangement for contingencies for academic performance (Gresham, 1986; Lentz, 1988).

To determine an existing deficit for a particular child, Noell and Witt (1999) have suggested a straightforward process. First, a "test" for a performance deficit is conducted using CBM reading probes (i.e., 100–200-word passages) selected from a child's basal reader as well as two basal readers that immediately precede the current reader. The reading probes are administered under standard (nonreinforced) conditions and under conditions where a preferred reinforcer is given for reading above a prespecified criterion. If performance increases markedly under the reinforcement conditions, then the student is assumed to have a performance deficit rather than a skill deficit. If reinforcement does not markedly improve performance, the student is assumed to have a skills deficit because even under conditions of high motivation, the student still cannot perform the requisite reading skills.

A number of examples in the applied behavior analysis literature have addressed the issue of skill versus performance deficits (Ayllon & Roberts, 1974; Daly & Martens, 1994; Daly, Martens, Dool, & Hintze, 1998; Daly, Martens, Hamler, Dool, & Eckert, 1999; Lovitt, Eaton, Kirkwood, & Pelander, 1971). For instance, Lovitt et al. (1971) gave incentives to improve students' oral reading fluency and to encourage them to read faster. A similar procedure was used by Daly et al. (1998, 1999). Another approach to assess academic performance deficits is to offer students a *choice* among reading materials or a *choice* in the order in which they will complete assignments (silent reading first, followed by vocabulary drill) (Daly, Witt, Martens, & Dool, 1997; Dunlap et al., 1994; Kern, Childs, Dunlap, Clarke, & Falk, 1994). If performance improves dramatically under choice conditions relative to no-choice conditions, then one can assume the student has a performance rather than a skill deficit.

MODELS OF RESPONSIVENESS TO INTERVENTION

Several models of intervention might be considered in adopting the responsiveness-to-intervention approach in defining LD. These models include (a) predictor-criterion models that use and teach those skills that best predict reading competency; (b) a dual-discrepancy model based on children's failure to respond to well-planned and implemented general education interventions, and (c) applied behavior analytic models which focus on manipulation of antecedent and consequent environmental events to improve reading competence.

Predictor-Criterion Models

These models of intervention focus on component skills or processes that represent the best predictors of skill in learning to read. Berninger and Abbott (1994) suggested that oral language skills (e.g., phonemic awareness, phonetic segmentation, rime) and orthographic skills (letter coding, letter cluster, word recognition) are among the best predictors of reading. Criteria used to evaluate reading competence include reading accuracy, reading rate, and reading comprehension. Similarly, direct instruction models (e.g.,

Englemann & Carnine, 1992; Kame'enui et al., 1995) and strategy training models (e.g., Graham & Harris, 1996; Levin, 1986; Pressley & Ghatala, 1990) focus on teaching those skills and strategies that best predict reading performances.

As reviewed previously, reading intervention programs having the most empirical support are those using a combination of direct instruction and strategy training (Swanson & Hoskyn, 1999). In addition, the work of Torgesen et al. (2001) showed strong and equal effects of reading programs focusing primarily on phonemic awareness and phonemic decoding versus programs emphasizing application of these skills in reading meaningful text. The intensity of this treatment may have influenced treatment outcome as well. Recall that these interventions were implemented for 67.5 hours over 8 weeks. Vellutino et al. (1996) used a similar intervention program that included a large component of strategy training. This intervention lasted 30–40 hours over 15 weeks. Swanson and Hoskyn's (1999) meta-analysis showed that the prototypical reading intervention lasted 13.3 hours over approximately 7 weeks.

Clearly, these models of intervention in the literature have produced rather strong effects in the literature with disabled readers. However, a key and unresolved question concerns how these models might be adopted within the LD eligibility process. The purpose of LD identification is to identify students who are *inadequately responding* to a validated intervention after a reasonable period, not to remediate or "normalize" reading skills. What must be determined is what constitutes a "reasonable period" and how to determine inadequate responsiveness. These issues are addressed in the final section of this paper.

Dual-Discrepancy Model

Fuchs and Fuchs (1997, 1998) have suggested using a CBM approach that measures a student's responsiveness (or lack thereof) to intervention delivered in the general education classroom. The logic behind the CBM approach to measure responsiveness to intervention is similar to that in endocrinology in which a child's growth over time is compared to that of a same-age group (Fuchs, 1995). A child who shows a large discrepancy between his or her height and that of a normative comparison group may be considered a candidate for certain types of medical intervention. In education, if a child is showing a discrepancy between the current level of academic performance and that of same-age peers, then that child may be a candidate for special education. It should be noted, however, that a low-performing child who shows growth rates similar to that of peers in the same classroom would not be a candidate for special education because the child is deriving similar education benefits (low though they may be) from that classroom (Fuchs, 1995).

Fuchs and Fuchs (1998) proposed a reconceptualization of the LD identification process based on a *treatment validity* notion. In this approach, students are not classified as LD unless and until it has been demonstrated empirically that they are not benefiting from the general education curriculum. Unlike traditional LD assessment, which assesses a student's status on ability and achievement measures at one point in time, the treatment validity approach repeatedly assesses the student's progress in the general education curriculum using CBM. Fuchs and Fuchs indicate that special education should be considered only when a child's performance shows a *dual discrepancy*—that is, the student *both* performs below the level evidenced by classroom peers and shows a learning rate substantially below that of classroom peers.

Fuchs and Fuchs (1998) state that the dual-discrepancy model is based on three related propositions. First, it assumes that because student ability varies widely, different students will experience different educational outcomes. Second, low academic performance is relative to the classroom in which the student is placed. If a student's growth rate is similar to peers, then that student would not be considered discrepant from peers' learning rates and would not be a candidate for special education placement. Conversely, a student whose growth rate is low relative to classroom peers would be considered a candidate for either an alternative intervention or special education placement. Third, if the majority of students in a general education classroom are demonstrating inadequate growth relative to local or national norms, then one must consider enhancing the educational program for the entire classroom before considering a student's unresponsiveness to intervention.

Use of this CBM dual-discrepancy approach to determine eligibility is a two-stage process: *problem identification* and *problem certification* (Fuchs & Fuchs, 1997; Marston & Magnusson, 1988; Shinn, 1989). *Problem identification* attempts to determine if a student's academic performance is sufficiently deficient to justify further assessment. Shinn (1989) recommended that three to five CBM tests in each academic area of concern be administered on consecutive days using the student's curriculum materials. On the basis of these brief assessments, the student's median score is used as an estimate of performance level. This performance level is then compared to the same assessment data collected from typical peers in the same classroom.

Fuchs and Fuchs (1997) suggest that procedures for sampling "typical peers" vary in completeness, elaboration, and time. Some districts routinely collect local CBM normative data and use this information to gauge progress in the curriculum and/or to determine special education eligibility (Shinn, 1989, 1995; Shinn, Tindal, & Stein, 1988). For districts not collecting normative CBM data, one can assess three same-gender peers selected randomly from students a teacher nominates as having adequate academic achievement in the classroom. With large-scale normative data, a referred student would be identified for further assessment if his/her median score fell at or below the 10th percentile or between 1 and 2 standard deviations below the mean. With data available only at the classroom level, discrepancies between actual and expected performance would be calculated by dividing the expected performance (based on the mean CBM performances of selected peers) divided by the referred student's median CBM score. A ratio of 2.0 or greater would suggest that further assessment is needed.

The *problem certification* phase is designed to determine whether or not the magnitude and severity of the student's academic deficiencies justify special education and related services (Shinn, 1995). In making this determination, three CBM probes are administered at successively lower levels of the student's curriculum. On the basis of these assessments, the highest level at which the student demonstrates successful performance is that student's grade placement. Fuchs and Fuchs (1997) suggest that "success" can be operationalized in two ways. First, if a large CBM normative data base is unavailable, success can be defined relative to fixed standards such as 40–60 words read correctly per minute in second-grade text. Second, if one has access to a large CBM data base, success is based on percentile ranks relative to the student's grade placement. If a student's median score falls between the 25th and 75th percentile for typical students at that grade level, then the student is demonstrating successful performance (Fuchs & Fuchs, 1997).

The longstanding and impressive research program using CBM by Lynn and Doug Fuchs of Peabody College at Vanderbilt University provides empirical support for the dual-discrepancy approach as a decision-making guide in LD eligibility determination (Fuchs, 1995; Fuchs et al., 1989a; Fuchs, Fuchs, & Fernstrom, 1993; Fuchs, Fuchs, Hamlett, Phillips, & Karns, 1995). Similarly, Douglas Marston of Minneapolis Public Schools has successfully used CBM to make eligibility determinations for students with LD (Marston et al., 1986; Marston & Magnusson, 1988; Marston, Mirkin, & Deno, 1984).

A recent investigation by Speece and Case (in press) provided additional data supporting the dual-discrepancy approach to defining LD. These authors identified children as at risk for reading failure if their mean performance on CBM reading probes placed them in the lowest quartile of their class. A contrast group was identified that was composed of five students from each classroom based on scores at the median (2 students) and the 30th, 75th, and 90th percentiles (1 student at each level). At-risk children were placed into one of three groups: CBM dual discrepancy (CBM-DD), regression-based IQ-reading achievement (IQ-DS), and low achievement (LA). Students in the CBM-DD group were given 10 CBM oral reading probes administered across the school year. Slopes (based on ordinary least squares regression) for each child and classroom were calculated, and each student's performance level was based on the mean of the last two data points. Children were placed in the CBM-DD group ($n = 47$) if their slope across the year and level of performance at the end of the year were >1 standard deviation below that of classmates. Students were placed in the IQ-DS group ($n = 17$) if their IQ-reading achievement discrepancy was 1.5 or more standard errors of prediction (approximately a 20-point discrepancy). Children were placed in the LA group ($n = 28$) if their total reading score was <90.

Results of this investigation showed that the CBM-DD group was more deficient on measures of phonological processing and was rated by teachers as having lower academic competence and social skills and more problem behaviors than the IQ-DS and LA groups However, the CBM-DD and IQ-DS groups were *not* different on a standardized measure of reading achievement demonstrating the specificity of the CBM-DD model. These data provided additional support for using the CBM-DD model to identify students with LD, specifically those with a phonological deficit. In summarizing their findings, Speece and Case (in press) suggested:

> Most research on reading disability proceeds from the assumption of failure to learn despite adequate instruction, a tenet of most definitions of learning disability, but this assumption is rarely tested. The dual discrepancy method does not reject the importance of individual differences to reading disability, but, in our view, expands the conceptualization to include the importance of instruction in the expression of the disability. (p. 36)

Fuchs and Fuchs (1997) proposed a three-phase model for determining LD eligibility using the CBM-DD approach. Phase I involves the documentation of adequate classroom instruction and dual discrepancies. It begins with weekly CBM assessments for *all students* in each school. An assessment team composed of a principal, school psychologist, special education teacher, and social worker review these data after 6 weeks to reach two decisions. First, the team decides if the overall classroom performance is adequate relative to other classrooms and district norms. Second, if classroom performance is acceptable, the team reviews individual student data to determine which students meet the dual-discrepancy criteria defined as (a) a difference of 1 standard deviation between a student's CBM median score and that of classmates *and* (b) a difference of 1 standard deviation between the student's CBM slope of improvement (growth) and that of classmates. Assuming students meeting these criteria do not have accompanying low-incidence conditions (e.g., mental retardation, sensory disabilities, autism), they proceed to Phase II of the process.

Phase II involves a prereferral intervention in which one member of the assessment team works with the general education teacher to design an intervention to remediate the student's dual discrepancy. CBM data are collected to judge the effectiveness of the intervention with the provision that the teacher implement a minimum of two interventions over a 6-week period. If students do not show adequate progress, they enter Phase III of the process.

Phase III of this process involves the design and implementation of an extended intervention plan. Essentially, this phase represents a special-education diagnostic trial period in which the student's responsiveness to a more intense intervention is measured. This phase lasts approximately 8 weeks, after which the team reconvenes and makes decisions concerning the child's placement. The team could decide that the intervention was successful and an individualized education plan (IEP) would be developed and the plan continued. Or, the team could decide that the intervention was unsuccessful in eliminating the dual discrepancy and consider alternative decisions such as changing the nature and intensity of the intervention, collecting additional assessment information, considering a more restrictive placement, or changing to a school having additional resources that better address the student's needs.

In summary, Fuchs and Fuchs (1997) propose that in order to qualify a student for special education, a three-pronged test must be passed: (a) a dual-discrepancy between the student's performance level and growth (1 standard deviation for each) and that of peers must be documented, (b) the student's rate of learning with adaptations made in the general education classroom is inadequate, and (c) the provision of special education must result in improved growth.

Functional Assessment Models

Another approach to identifying students on the basis of responsiveness to intervention comes from the applied behavior analysis (ABA) camp (Daly, Lentz, & Boyer, 1996; Daly & Martens, 1994; Daly et al., 1997; Haring, Lovitt, Eaton, & Hansen, 1978; Howell, Fox, & Morehead, 1993). This approach attempts to offer a *functional* rather than a *structural* explanation for children's academic difficulties. I also include within the ABA approach the Direct Instruction (Englemann & Carnine, 1991; Gersten et al., 1986) as well

as the Precision Teaching models of intervention (Lindsley, 1991). The field of LD has traditionally offered structural explanations in the form of labels or traits to explain academic problems (e.g., LD, dyslexia, processing disorders). Structural explanations are not particularly useful from an intervention perspective because student traits (inferred from performances) cannot be directly manipulated and because the explanations do not identify environmental factors that might be contributing to academic failure (Daly et al., 1997).

Alternatively, a functional approach to understanding academic failure attempts to relate academic performance to environmental events that precede and follow student performance (e.g., opportunities to respond, reinforcement for accurate responding, time allocated for instruction, modeling and feedback of academic behaviors). From a functional perspective, the job of the interventionist is to analyze those factors that may explain poor performance and implement an instructional intervention to improve academic responding. In a functional approach, academic responding is operationalized using curriculum-based measures of oral reading, mathematics computation, written expression, and spelling such as those recommended in the dual-discrepancy approach of Fuchs and Fuchs (1997, 1998).

Daly et al. (1997) identified five common reasons why students fail and provided rather straightforward methods for testing these hypotheses quickly and efficiently so as to lead to interventions. The reasons are as follows: (a) they do not want to do it ("won't do" problems), (b) they have not spent enough time doing it (lack of practice and feedback), (c) they have not had enough help to do it (insufficient prompting or poor fluency), (d) the student has not had to do it that way before (instructional demands do not promote mastery), and (e) it is too hard (poor match between student skill level and instructional materials).

An extremely important concept in a functional approach to remediating academic difficulties is the instructional hierarchy (Haring et al., 1978). The instructional hierarchy describes the relationship between intervention components and stages of skill mastery. In the instructional hierarchy, students move through states of *acquisition, fluency, generalization,* and *adaptation*. Strategies that use modeling, prompting, and error correction can be expected to improve acquisition (accuracy), and strategies including practice and reinforcement are expected to improve fluency. Generalization training involves discrimination training across stimuli and maintenance activities over time (Daly et al., 1996; Martens, Witt, Daly, & Vollmer, 1999).

There is an extensive research base supporting the ABA model for improving academic performances (Daly et al., 1997, 1999; Elliott, Busse, & Shapiro, 1999; Englemann & Carnine, 1991; Greenwood, 1991; Skinner, 1998). Swanson and Sachs-Lee (2000) summarized 85 studies using single-subject designs across the academic domains of reading, mathematics, writing, and language using direct instruction (DI), strategy training (SI), Combined DI+SI, and non-DI/non-SI described earlier in this paper (see Swanson & Hoskyn, 1999). Based on an analysis of 793 effect sizes, the mean effect size was 0.87 (SD = 0.32), suggesting a strong effect. The average age of participants was almost 11 years and the mean IQ and achievement levels of participants were 95 and 77, respectively (M = 100, SD = 15). Results of this meta-analysis showed that DI and SI were effective in remediating academic deficits (except handwriting) and all interventions were more effective with lower IQ students than higher IQ students in reading.

The use of the ABA approach for eligibility determination creates some measurement challenges because this model relies almost exclusively on single-case experimental design data. Both the predictor-criterion and CBM-DD models use well-established and straightforward quantitative approaches to determine treatment nonresponders. An unresolved issue in the ABA approach concerns the most appropriate way of quantifying the effects of intervention. Gresham and Lambros (1998) identified several methods for quantifying the effects of interventions using single-case experimental design data that are described below. Time-series analysis is not included here because fitting these regression models with relatively few data points often yields inaccurate results and it is often impossible to meet the statistical assumptions of these models in educational practice (Kazdin, 1984).

Visual Inspection

Visual inspection of graphed data is by far the most common way of analyzing data from single-case designs (Johnston & Pennypacker, 1993). Effects of intervention are determined by comparing baseline levels of performance to postintervention levels of performance to detect treatment effects. Unlike statistical analyses, this method uses the "interocular" test of significance. There is a considerable body of research, however, suggesting that even highly trained behavior analysts cannot obtain consensus in evaluating single case data using visual inspection (Center, Skiba, & Casey, 1985–86; DeProspero & Cohen, 1979; Knapp, 1983; Matyas & Greenwood, 1990, 1991; Ottenbacher, 1990). It would appear that visual inspection of graphed data often results in erroneous conclusions regarding the presence or absence of treatment effects, particularly given that the data points are serially dependent or autocorrelated.

A study by Matyas and Greenwood (1990) showed that Type I error rates ranged from 16 to 84% for autocorrelated data, suggesting that researchers often judge the presence of treatment effects where none exist. Given the interpretative problems with graphed data in determining treatment effects and unacceptably high Type I error rates, other procedures should be used to supplement or corroborate interpretation of graphed data (Fisch, 1998). These are described in the following sections.

Reliable Changes in Behavior

Another method of quantifying effects in single-case designs is to calculate the extent to which changes in academic performance are reliable. Nunnally and Kotsche (1983) first proposed a reliable change index (RCI) to determine the effectiveness of an intervention for individuals. The RCI is defined as the difference between a posttest score and a pretest score divided by the standard error of difference between posttest and pretest scores (Christensen & Mendoza, 1986; Jacobson, Follette, & Revenstorf, 1984). The standard error of difference is the spread or variation of the distribution of change scores that would be expected if no actual change had occurred. An RCI of +1.96 ($p < 0.05$) would be considered a reliable change in behavior.

With single-case data, RCIs must be computed for baseline (pretest) and intervention (posttest) phases of the design. For example, in an ABAB withdrawal design, pretest scores would be calculated from the initial baseline (A) and posttest scores from the mean of the two intervention phases (B+B). Similarly, in a multiple baseline design, pretest scores would be calculated from the baselines of each subject (setting or behavior) and posttest scores from the means of the respective intervention phases. The standard error of difference would be based on the autocorrelation and variation of baseline and intervention phases. Although the RCI approach can be used to detect reliable changes in academic performance (relative to baseline) for a single student, it does not provide specific decision rules that might be used in making an LD eligibility determination. Moreover, RCIs are influenced by the reliability of the dependent measures used. If a measure is highly reliable (0.90 or higher), then small changes in behavior could be considered statistically reliable. Conversely, if a measure has low reliability, then large changes in behavior might not be statistically reliable, but could be important.

Effect sizes. Another way of quantifying single-case data is through the use of effect sizes. Although effect sizes typically are used to integrate group design research studies, Busk and Serlin (1992) have proposed two methods for calculating effect sizes in single-case studies. The first approach makes no distributional assumptions and calculates effect sizes by subtracting the treatment mean from the baseline mean and dividing by the standard deviation of the baseline mean. The second approach, based on the homogeneity of variance assumption, is the same, except that it uses the pooled within-phase variances as the error term. Effect sizes calculated in this way are interpreted the same way as traditional effect-size estimates. They can be used to estimate the effects of one or more treatments for an individual or to summarize a body of single-case intervention.

Swanson and Sachs-Lee (2000) used an alternative approach to calculate effect size by using the last three data points in baseline and treatment phases to calculate the means. This difference was then divided by the correlation between baseline and treatment data points, taking into account the average standard deviation for repeated measures. These authors argue that the number of sessions may inflate or deflate effect sizes

and are subject to fluctuations in the dependent variable that are not a result of the treatment (cyclicity).

Effect sizes also can be calculated by computing the percentage of nonoverlapping data points (PNOL) between baseline and treatment phases (Mastropieri & Scruggs, 1985–86). PNOL is computed by indicating the number of treatment data points that exceed the highest baseline data point and dividing by the total number of data points in the treatment phase. For example, if 8 of 10 treatment data points exceed the highest baseline data point, then PNOL is 80%. This method provides for quantitative synthesis of single-case data that is relatively easy. However, the method would be inappropriate in some situations, including unusual baseline trends, floor and ceiling effects, and students in the initial stages of skill acquisition (Strain, Kohler, & Gresham, 1998).

Yet another approach in quantifying the effects of interventions in single-subject designs is to analyze trends over time by using time-structured Markov chains (Fisch, 1998). Markov chains involve the analysis of two-dimensional matrices containing the probabilities of changing from one set of conditions (e.g., preintervention performances) to another set of conditions (postintervention performances). Haccou and Meelis (1992) indicate that Markov chains are used frequently in naturalistic settings to assess changes in "states" of behavior from one time period to the next.

Social Validation

Social validity deals with three fundamental questions faced by professionals in the field of LD: What should we change? How should we change it? How will we know it was effective? There are sometimes disagreements among professionals as well as between professionals and consumers on these three fundamental questions. Wolf (1978) described the social validation process as the assessment of the *social significance* of the goals of intervention, the *social acceptability* of the intervention procedures to attain these goals, and the *social importance* of the effects of the intervention. This last component of the social validation process is most relevant to quantifying a student's responsiveness to intervention in the LD eligibility determination process.

The social importance of the effects produced by an intervention established the practical or educational significance of changes in academic performance. Do the quantity and quality of the change in academic performance make a difference in the student's academic functioning? Does the change in academic performance have habilitative validity (Hawkins, 1991)? Is the student's academic performance now in the "functional" range? All of these questions capture the essence of establishing the social importance of intervention effects.

One means of establishing the social importance of intervention effects is to conceptualize academic functioning as belonging to either a functional or dysfunctional distribution. For example, we could socially validate a reading intervention by demonstrating that a student moved from a dysfunctional to a functional range of reading performance. This result could be established by calculating the probability that the student's reading score belonged to a functional rather than a dysfunctional distribution. We could base these calculations on norm-referenced achievement tests or locally normed CBM measures.

Fawcett (1991) suggested that in evaluating the social importance of effects, we should specify various levels of performance. For example, one could specify *ideal* (the best performance available), *normative* (typical or commonly occurring performance), or *deficient* (the worst performance available). Interventions moving a student from a deficient level of performance to normative or ideal levels of performance could be considered socially important.

CONCLUSION

This paper argues that a child's inadequate responsiveness to an empirically validated intervention can be taken as evidence of LD and should be used to classify children as such. Some might argue that diagnoses in medicine, for example, are not confirmed or disconfirmed on the basis of whether a patient responds to treatment. However, one should always keep in mind that medical diagnoses often have direct treatment

implications and that the causes of many physical diseases (unlike mild disabilities such as LD) are known. Moreover, treatment intensity in medicine is typically matched to the nature and severity of whatever physical malady is present. Obviously, a physician's first choice of treatment for most medical problems is not hospitalization. The point here is that not all children will require the most intense form of treatment of academic difficulties, and treatment intensity, strength, and/or duration should increase only after the child fails to show an adequate response to intervention.

In the current paper, I argue that children who fail to respond to empirically validated treatments implemented with integrity might be identified as LD. The concept of *responsiveness to intervention* appears to be a viable alternative approach to defining LD, particularly in light of the myriad difficulties with discrepancy-based models. This paper defines *responsiveness to intervention* as a change in academic performance as a function of an intervention. In order to employ treatment responsiveness as a criterion for identifying students as LD, assessment procedures should have treatment validity; that is, the assessment should contribute to the planning and implementation of more effective treatments to remediate academic deficits (Fuchs & Fuchs, 1998; Gresham & Witt, 1997; Nelson et al., 1987). Several issues in adopting the responsiveness-to-intervention approach appear to have been resolved, including (a) modeling academic growth, (b) sensitivity of measures to reflect growth, and (c) validated treatment protocols. These were discussed at length in this paper and will not be reiterated here except to say that the validated treatment protocols represent different intensities and durations of treatment. Depending on a student's response to treatment, these treatments may have to be titrated until an acceptable level of academic functioning is achieved. More important, several unresolved issues await further investigation and deliberation before the field can adopt responsiveness to intervention in eligibility determination.

Unresolved Issues in the Alternative Responsiveness-to-Intervention Approach

Five important issues appear to be most important at this time in adopting responsiveness to intervention as the criterion for LD eligibility determination: (a) selecting the "best" intervention available, (b) determining the optimal length and intensity of the intervention, (c) ensuring the integrity of interventions, and (d) conducting cost-benefit analyses. These issues are discussed in the following sections.

Selecting the "best" intervention available. According to available research, there appears to be a consensus on the core components a reading intervention should address for students with reading disabilities. Reading research over the past 20 years indicates that the reading difficulties of these students are caused by weaknesses in the ability to process the phonological aspects of language (Liberman, Shankweiler, & Liberman, 1989; Stanovich & Siegel, 1994; Torgesen, 1996). In fact, reading growth is best predicted by initial levels of phonological skill rather than verbal ability or discrepancy between IQ and reading achievement (Torgesen et al., 2001; Vellutino et al., 1996, 1998). Torgesen et al. (2001) suggested that these phonological weaknesses require reading instruction that is more phonemically explicit and systematic than that provided to other children and there are many ways in which this might be accomplished in designing instructional activities.

Given the above consensus regarding the most important skills to target in intervention, what is the "best" intervention to accomplish this end? The meta-analysis by Swanson and Hoskyn (1999) suggested that interventions using a combination of direct instruction and strategy instruction produced the largest effect sizes, with 80% of the treatment groups having mean reading scores equal to or greater than those of control group students. Recall that the typical intervention in this meta-analysis was 13.3 hours over 10 weeks. Vellutino et al.'s (1996) intervention provided 35–40 hours of instruction over 15 weeks whereas the recent study by Torgesen et al. (2001) involved 67.5 hours over 8–9 weeks.

Comparisons among these studies are difficult given the large variability in the intensity and length of interventions (to be discussed below). Interventions based on applied behavior analysis, while effective, typically are of shorter duration, and outcome measures typically are more narrowly defined (Daly et al., 1996; Daly & Martens, 1994; Haring et al., 1978). Given the various effective intervention options available, practitioners must determine what "best practices" will be at the local level in terms of selecting

and implementing a given strategy.

Determining the optimal length and intensity of intervention. Determining the length and intensity of intervention that is implemented is a crucial decision when using responsiveness to intervention as the criterion for identifying LD. Keep in mind a fundamental principle: The length and intensity of intervention will depend entirely on a student's responsiveness to it, which is individually based. Fuchs and Fuchs (1997, 1998) indicated that a general educator should attempt two interventions lasting no longer than 6 weeks before placing the student in a special education trial period. This special education trial period should last no longer than 8 weeks, after which time the assessment team reconvenes to continue and/or enhance the intervention program. Fuchs and Fuchs (1997) suggested that any assessment method must provide adequate data for evaluating treatment effectiveness and should answer the following questions. Is the nonadapted regular education classroom producing adequate academic growth? Have adaptations to the general education classroom produced improved growth? Has the provision of special education interventions improved student learning?

Another insight into this issue of length and intensity of interventions can be found in the meta-analysis of Swanson and Hoskyn (1999). As stated earlier, the typical intervention consisted of 22.47 minutes of daily instruction delivered 3.58 times per week for 35.72 sessions. Thus, the prototypical intervention consisted of about 13.3 hours of instruction distributed over approximately 10 weeks. It should be noted, however, that there was a huge degree of variability in terms of minutes of daily instruction ($SD = 29.71$ minutes), times per week ($SD = 1.58$), and number of sessions ($SD = 21.72$ sessions). Moreover, the samples used in these studies varied greatly regarding criteria used for participant selection, thereby introducing a confounding factor when evaluating responsiveness to intervention.

The prototypical study using (a) direct instruction, (b) strategy training, and (c) combined direct instruction + strategy training produced effect sizes of 0.77, 0.67, and 0.81, respectively. Also, students having the most severe reading deficits (<85) responded better to treatments ($M = 0.71$) than students with less severe reading difficulties (>84 and <91; $M = 0.51$). If one were to use the length and intensity of the prototypical reading study in this meta-analysis with a combination of direct instruction and strategy training, one could expect to produce a standard score point difference of 12 ($M = 100$, $SD = 15$) between pretest and posttest scores. For example, a student entering the intervention with a standard score of 78 could be expected to improve to a score of 90 at posttest, thereby indicating near-normal performance.

Another approach to determining optimal length and intensity of intervention can be found in the Vellutino et al. (1996) investigation. Recall that this study selected children who scored at or below the 15th percentile in reading (Word Identification and Word Attack) and were given 35–40 weeks of intensive one-to-one tutoring in reading. Each session lasted for 30 minutes, and 80 sessions were spread over 15 weeks for a total of 35–40 hours of reading instruction. At posttest, about half of the children showed either Good Growth or Very Good Growth in reading with posttest percentile ranks in the 44th and 64th percentiles, respectively, by the end of second grade. This study suggested that an intensive one-to-one reading intervention could be used to normalize reading performances of poor readers selected in the first grade. It is unknown at this time, however, how much one might change or otherwise deviate from this effective treatment protocol and produce similar results.

Finally, the study by Torgesen et al. (2001) compared two interventions with fourth graders implemented in two, 50-minute daily sessions, 5 days per week over 8–9 weeks (67.5 hours of intervention). The 19 children who were returned to general education subsequent to intervention moved from pretest scores of about 70 (average of Word Attack and Word Identification) to 2-year follow-up scores of approximately 95. In contrast, the students remaining in special education moved from pretest scores of about 67 to posttest scores of 83. Relative to growth made in the regular resource room, the average effect size was approximately 4.15 for the two treatment groups (difference between pretreatment and posttreatment slopes divided by pooled variability of pretreatment slopes). As with the Vellutino et al. (1996) study, we do not know how much this intervention can be modified or diluted and still obtain relatively large treatment effects.

One means of determining the optimal length and intensity of interventions based on the extant literature is to employ a multiple gating procedure similar to that used in the Heartland Area Education Agency (AEA) in Iowa to make special education entitlement decisions (Reschly & Tilly, 1999; Reschly & Ysseldyke, 1995). Figure 1 shows the problem-solving model used in the Heartland AEA for making special education eligibility determinations. Note that I have superimposed examples of interventions varying in intensity (that were reviewed in the current paper) within the Heartland AEA model. The responsiveness-to-intervention approach in this model makes the following assumptions:

1. The intensity (and costs) of intervention is matched to the degree of unresponsiveness to the intervention.

2. Movement through levels of intervention intensity is based on inadequate response to interventions implemented with integrity.

3. Decisions regarding movement through levels are based on an ongoing collection of empirical data collected from a variety of sources.

4. An increasing body of knowledge (data) is collected to inform decision making as a student moves through the levels.

5. Special education and IEP determination should be considered only after a student shows inadequate responsiveness to interventions at the previous levels.

Figure 1. Degree of unresponsiveness and intensity of treatment.

Ensuring the integrity of interventions. Treatment integrity (sometimes called treatment fidelity or procedural reliability) refers to the degree to which a treatment is implemented as intended (Gresham, 1989; Yeaton & Sechrest, 1981). Establishing and maintaining the integrity of treatments is one of the most important aspects of both the scientific and practical application of instructional procedures. It is likely that the ineffectiveness of many instructional interventions can be attributed, in part, to the poor integrity with which these procedures were implemented (i.e., deviations from an established treatment protocol). Adopting a responsiveness-to-intervention approach to identifying LD makes treatment integrity (the reliability of treatment implementation) a central feature of the entire process. In contrast, the entire practice of determining the most appropriate IQ-achievement discrepancy model is based on the reliability of difference scores (e.g., simple difference, predicted difference). In order to determine the degree of responsiveness to intervention, a treatment must be reliably and accurately implemented.

Recently, Gresham, MacMillan, Beebe-Frankenberger, and Bocian (2000) sought to determine the extent to which integrity was assessed in the LD intervention literature by analyzing articles in the three major LD journals from January 1995 to August 1999 (*Journal of Learning Disabilities, Learning Disability Quarterly,* and *Learning Disabilities: Research & Practice*). Of the 479 articles published in these journals, 65 articles (13.6%) were intervention articles. Of these 65 articles, only 12 articles (18.5%) actually measured and reported data on treatment integrity. In their synthesis of the LD intervention literature, Swanson, Carson, and Saches-Lee (1996) reported that less than 2% of the studies provided *any* information about treatment integrity. In spite of the methodological and statistical rigor used in this and other meta-analyses of the LD literature, none of these methodological considerations can answer two fundamental questions: (a) How are treatments implemented, and (b) What is the relation between treatment integrity and treatment outcomes in LD intervention research?

Swanson and Sachs-Lee (2000), in their review of the single-case intervention research with LD, found that only 28% of the studies ($N = 24$ studies) provided any measure of treatment integrity. Of these 28 studies, only 8 studies specified steps used to measure the integrity of the intervention. There appears to be a curious double standard in the LD intervention literature with respect to the measurement and reporting of reliability for the independent and dependent variables. That is, it is almost always the case that reliability data for the dependent variable are presented in published treatment-outcome research. In contrast, this same type of information rarely is required for the independent (treatment) variable.

Given the central importance of assessing treatment integrity in the responsiveness-to-intervention model of LD identification, the following recommendations are offered concerning how researchers and practitioners might conduct integrity assessments:

Specific components of an intervention should be operationally defined and measured much like the operational definition and measurement of dependent measures.

Each component of a treatment should be measured by either direct observation or videotaping using an occurrence-nonoccurrence method. Levels of treatment integrity should be obtained by summing the number of components correctly implemented and dividing this number by the total number of components to yield percentage integrity.

Two estimates of treatment integrity should be calculated. One, the integrity of each component across days or sessions of treatment should be computed to yield *component integrity*. Two, the integrity of all treatment components within days or sessions of treatment should be calculated to yield *daily or session integrity*. Given these two estimates of integrity, failure to find significant treatment effects might be explained by poor component integrity over time, by poor daily or session integrity, or both.

⊞ Indirect methods of assessing treatment integrity such as instructional manuals, permanent products, self-reports, interviews, and behavior rating scales should be used to supplement direct measures of integrity, but they must be interpreted cautiously. There is often low agreement between direct and indirect methods of integrity assessment (Gresham, 1997; Noell & Witt, 1999; Wickstrom, Jones, LaFleur, & Witt, 1998).

Cost-benefit analysis. An important aspect of using the responsiveness-to-intervention approach to LD identification is determining the financial costs to school districts. As mentioned earlier, the average cost of a traditional eligibility determination for a student with a mild disability is around $2,500 per case (Reschly, personal communication, 2001). What costs are incurred by using the CBM–dual-discrepancy model in which local normative data are collected over 20 weeks? What costs are associated with adopting any of the functional assessment models? Currently, we have no published data to assist us in calculating these costs.

Torgesen (personal communication, 2001), however, provided some data regarding the costs of his intensive intervention program described earlier (Torgesen et al., 2001). Torgesen states that a teacher who was doing this kind of intervention with children (two 50-minute sessions per day) could probably work with two children at a time for 8 weeks and the rest of the time could be spent following up on children taught earlier, or working as a teacher consultant, or planning. Given the normal interruptions in schools (assemblies, absences) it takes about 10 weeks of teacher time to deliver the full 80 sessions.

A teacher could work with about six severely LD children a day for 10 weeks. On the basis of a 37-week school year, a teacher could probably go through about three treatment cycles with six students per cycle and thus provide intensive reading intervention services to approximately 18 children per year (6 students H 3 treatment cycles). Remember, however, that Torgesen et al.'s (2001) data suggest that about half of these children will no longer need special education after the intervention. One way Torgesen calculates the cost is to take the cost per session at $50 (more or less depending on local costs for private tutoring) and multiply this figure by 80 sessions of instruction; the cost per student is approximately $2,000. Thus, for a teacher working with 18 students per year, the total cost of an intensive, treatment-oriented approach to LD would be about $36,000 per year. The mere cost of simply identifying, but not treating, 18 LD students using traditional IQ-achievement discrepancies is estimated to be $45,000 (18 H $2500).

One should consider these costs in light of the fact that the cost of educating a student in a resource room placement is 1.7 to 2.0 times the cost of educating a general education student in a regular classroom. In addition, remember that in the Torgesen et al. (2001) study, 40% of the students in the study no longer needed special education. Moreover, one should also note that the efficacy of traditional special-education-delivered interventions, according to meta-analyses, have been somewhat less than impressive (Kavale & Forness, 1999).

Another consideration in calculating these cost-benefits is the cost of LD eligibility determination using the traditional competing paradigm model described in concert with special education costs. Assuming the cost of a typical eligibility process is approximately $2,500 and also remembering that all LD students must undergo 3-year reevaluations, the cost of identifying and providing special education for LD students is almost twice that of educating general education students. As such, there may be long-term cost-benefits in adopting the responsiveness-to-intervention model, particularly in light of the following: (a) The average effect size of special education placement for LD students is about 0.30 (Kavale & Forness, 1999), (b) relatively few students get decertified as LD during their school careers, (c) early intensive reading interventions for poor readers (kindergarten-first grade) leads to GG or VGG in reading for about 50% of this population, and (d) intensive intervention may lead to a decertification of about 40% of children receiving this type of intervention.

The question for the LD field remains: How long do we implement an intervention before we determine that a child is an inadequate responder and thus eligible for more intensive special education services? Further, what is the cost of this intervention-based model relative to the traditional eligibility approach? Is the responsiveness-to-intervention approach more expedient in identifying students as LD so that intervention takes place earlier? How intense should this intervention be and how long should it last? Who should implement the intervention (teachers, paraprofessionals, reading specialists)? These questions must be addressed first when adopting a responsiveness-to-intervention approach to the identification of LD.

One must realize that some individuals have political, personal, financial, and/or other reasons in wanting to maintain the status quo in the classification of students as LD. This position is indefensible in light of the overwhelming evidence in the field that the IQ-discrepancy approach to LD identification is simply not valid and, most important, does not inform treatment decisions. These individuals may argue that a treatment-responsiveness model is analogous to confirming the accuracy of a cancer diagnosis by determining whether or not a treatment regimen of chemotherapy and radiation leads to remission. They might also argue that this approach does not improve the identification of students as LD, that it has some insurmountable measurement problems, that it leads to late identifications, and that it will be extremely expensive. However, it always should be remembered that these arguments are simply red herrings in the sea of abyss of what we now call LD.

It is incumbent upon the LD field to focus on answering the critical questions using empirical findings for assessment and interventions provided in this paper as a foundation. Establishing an effective method for determining eligibility for LD that can be linked to intervention can go a long way toward decreasing, if not eliminating, the probability that learning disabilities will continue to be the sociological sponge that wipes up the spills of general education.

REFERENCES

Arter, J., & Jenkins, J. (1979). Differential-diagnosis-prescriptive teaching: A critical appraisal. *Review of Educational Research, 49,* 517–555.

Ayllon, T., & Roberts, M. (1974). Eliminating discipline problems by strengthening academic performance. *Journal of Applied Behavior Analysis, 7,* 71–76.

Ayres, R., & Cooley, E. (1986). Sequential versus simultaneous processing on the K-ABC: Validity in predicting learning success. *Journal of Psychoeducational Assessment, 4,* 211–220.

Ayres, R., Cooley, E., & Severson, H. (1988). Educational translation of the Kaufman Assessment Battery for Children: A construct validity study. *Journal of Psychoeducational Assessment, 4,* 113–124.

Bahr, M., Fuchs, D., Stecker, P., & Fuchs, L. (1991). Are teachers' perceptions of difficult-to-teach students racially biased? *School Psychology Review, 20,* 599–608.

Bateman, B. (1965). An educational view of a diagnostic approach to learning disorders. In J. Hellmuth (Ed.), *Learning disorders* (Vol. 1, pp. 219–239). Seattle, WA: Special Child.

Berninger, V. W., & Abbott, R. D. (1994). Redefining learning disabilities: Moving beyond aptitude-

achievement discrepancies to failure to respond to validated treatment protocols. In G. Reid Lyon (Ed.), *Frames of reference for the assessment of learning disabilities* (pp. 163–183). Baltimore, MD: Paul H. Brookes.

Berninger, V., Hart, T., Abbott, R., & Karovsky, P. (1992). Defining reading and writing disabilities with and without IQ: A flexible developmental perspective. *Learning Disability Quarterly,* 103–118.

Bocian, K., Beebe, M., MacMillan, D., & Gresham, F. M. (1999). Competing paradigms in learning disabilities classification by schools and variations in the meaning of discrepant achievement. *Learning Disabilities Research & Practice, 14,* 1–14.

Borkowski, J., & Turner, L. (1990). Transsituational characteristics of metacognition. IN W. Schneider & F. Weinert (Eds.), *Interactions among aptitudes, strategies, and knowledge in cognitive performance* (pp. 159–176). New York: Springer-Verlag.

Bryk, A., & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Burchinal, M., Bailey, D., & Snyder, P. (1994). Using growth curve analysis to evaluate child change in longitudinal investigations. *Journal of Early Intervention, 18,* 403–423.

Busk, P., & Serlin, R. (1992). Meta-analysis for single-case research. In T. Kratochwill & J. Levin (Eds.), *Single-case research design and analysis* (pp. 187–212). Hillsdale, NJ: Erlbaum.

Center, B., Skiba, R., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education, 19,* 387–400.

Christensen, L., & Mendoza, J. (1986). A method of assessing change in a single subject: An alteration of the RC index. *Behavior Therapy, 17,* 305–308.

Clay, M. (1987). Learning to be learning disabled. *New Zealand Journal of Educational Studies, 22,* 155–173.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Inc.

Cone, J.D. (1989). Is there utility for treatment utility? *American Psychologist, 44,* 1241–1242.

Cronbach, L. (1957). The two disciplines of scientific psychology. *American Psychologist, 12,* 671–684.

Cronbach, L. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 30, 116*–127.

Cronbach, L., & Snow, R. (1977). *Aptitudes and instructional methods*. New York: Wiley (Halstead Press).

Daly, E., & Martens, B. (1994). A comparison of three interventions for increasing oral reading performance: Application of the instructional hierarchy. *Journal of Applied Behavior Analysis, 27,* 459–469.

Daly, E., Lentz, F. E., & Boyer, J. (1996). The instructional hierarchy: A conceptual model for understanding the effective components of reading interventions. *School Psychology Quarterly, 11,* 369–386.

Daly, E., Martens, B. K., Dool, E., & Hintze, J. (1998). Using brief functional analysis to select interventions for oral reading. *Journal of Behavioral Education, 8,* 203–218.

Daly, E., Martens, B. K., Hamler, K., Dool, E., & Eckert, T. (1999). A brief experimental analysis for identifying instructional components needed to improve oral reading fluency. *Journal of Applied Behavior Analysis, 32,* 83–94.

Daly, E., Witt, J. C., Martens, B. K., & Dool, E. (1997). A model for conducting functional analysis of academic performance problems. *School Psychology Review, 26,* 554–574.

D'Amato, R. C., Rothlisberg, B., & Work, P. (1999). Neuropsychological assessment for intervention. In C. Reynolds & T. Gutkin (Eds.), *Handbook of school psychology* (3rd ed., pp. 452–475). New York: Wiley.

Das, J. P. (1995). Neurocognitive approach to remediation: The PREP Model. *Canadian Journal of*

School Psychology, 9, 157–173.

Das, J. P., Naglieri, J., & Kirby, J. (1995). *Assessment of cognitive processes*. Needham, MA: Allyn & Bacon.

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis, 12,* 573–579.

Dunlap, G., DePerczel, M., Clarke, S., Wilson, D., Wright, S., White, R., & Gomez, A. (1994). Choice making to promote adaptive behavior for students with emotional and behavior challenges. *Journal of Applied Behavior Analysis, 27,* 505–518.

Elliott, S., Busse, R., & Shapiro, E. (1999). Intervention techniques for academic performance problems. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd ed., pp. 664–685). New York: Wiley.

Englemann, S., & Carnine, D. (1991). *Theory of instruction: Principles and application.* Eugene, OR: ADI.

Epps, S., Ysseldyke, J., & McGue, M. (1984). Differentiating LD and non-LD students: "I know one when I see one." *Learning Disability Quarterly, 7,* 89–101.

Fawcett, S. (1991). Social validity: A note on methodology. *Journal of Applied Behavior Analysis, 24,* 235–239.

Fisch, G. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst, 21,* 111–123.

Foorman, B., Francis, D., Fletcher J., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90,* 37–55.

Fuchs, D., Fuchs, L., & Fernstrom, P. (1993). A conservative approach to special education reform: Mainstreaming through transenvironmental programming and curriculum-based measurement. *American Education Research Journal, 30,* 149–178.

Fuchs, L., & Fuchs, D. (1997). Use of curriculum-based measurement in identifying students with disabilities. *Focus on Exceptional Children, 30,* 1–16.

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

Fuchs, L., Fuchs, D., & Hamlett, C. (1989a). Effects of alternative goal structures within curriculum-based measurement. *Exceptional Children, 55,* 429–438.

Fuchs, L., Fuchs, D., & Hamlett, C. (1989b). Effects of instructional use of curriculum-based measurement to enhance instructional programs. *Remedial and Special Education, 10,* 43–52.

Fuchs, L., Fuchs, D., & Hamlett, C. (1989c). Monitoring reading growth using student recalls: Effects of two teacher feedback systems. *Journal of Educational Research, 83,* 103–111.

Fuchs, L., Fuchs, D., Hamlett, C., Phillips, N., & Karns, K. (1995). General educators' specialized adaptation for students with learning disabilities. *Exceptional Children, 61,* 440–459.

Fuchs, D., Mathes, P., Fuchs, L., & Lipsey, M. (2001). *Is LD just a fancy term for underachievement? A meta-analysis of reading differences between underachievers with and without the label.* Nashville, TN: Vanderbilt University.

Gerber, M., & Semmel, M. (1984). Teacher as imperfect test: Reconceptualizing the referral process. *Educational Psychologist, 14,* 137–146.

Gersten, R., Woodward, J., & Darch, J. (1986). Direct Instruction: A research-based approach to curriculum and teaching. *Exceptional Children, 53,* 17–31.

Good, R., Vollmer, M., Creek, R., Katz, L., & Chowdhri, S. (1993). Treatment utility of the Kaufman Assessment Battery for Children: Effects of matching instruction and student processing strength. *School Psychology Review, 22,* 8–26.

Gottlieb, J., Alter, M., Gottlieb, B., & Wishner, J. (1994). Special education in urban America: It's not justifiable for many. *The Journal of Special Education, 27,* 453–465.

Graham, S., & Harris, K. (1996). Self-regulation and strategy instruction for students who find writing and learning challenging. In C. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 347–360). Mahwah, NJ: Erlbaum.

Greenwood, C. (1991). A longitudinal analysis of time, engagement, and achievement in at-risk versus non-risk students. *Exceptional Children, 57,* 521–535.

Gresham, F. M. (1986). Conceptual issues in the assessment of social competence in children. In P. Strain, M. Guralnick, & H. Walker (Eds.), *Children's social behavior: Development, assessment, and modification* (pp. 143–179). New York: Academic Press.

Gresham, F. M. (1989). Assessment of treatment integrity in school consultation and prereferral intervention. *School Psychology Review, 18,* 37–50.

Gresham, F. M. (1991). Conceptualizing behavior disorders in terms of resistance to intervention. *School Psychology Review, 20,* 23–36.

Gresham, F. M. (1997). Treatment integrity in single-subject research. In R. Franklin, D. Allison, & B. Gorman (Eds.), *Design and analysis of single case research* (pp. 93–117). Mahwah, NJ: Lawrence Erlbaum.

Gresham, F. M., & Lambros, K. (1998). Behavioral and functional assessment. In T. S. Watson & F. M. Gresham (Eds.), *Handbook of child behavior therapy* (pp. 3–22). New York: Plenum.

Gresham, F. M., MacMillan, D. L., Beebe-Frankenberger, M., & Bocian, K. (2000). Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? *Learning Disabilities Research & Practice, 15,* 198–205.

Gresham, F. M., MacMillan, D. L., & Bocian, K. (1997). Teachers as "tests": Differential validity of teacher judgments in identifying students at-risk for learning difficulties. *School Psychology Review, 26,* 47–60.

Gresham, F. M., Reschly, D. J., & Carey, M. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. *School Psychology Review, 16,* 543–563.

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly, 12,* 249–267.

Haccou, P., & Meelis, E. (1992). *Statistical analysis of behavioural data: An approach based on time-structured models.* Oxford, England: Oxford University Press.

Haring, N., Lovitt, T., Eaton, M., & Hansen, C. (1978). *The fourth R: Research in the classroom.* Columbus, OH: Merrill.

Hawkins, R. (1991). Is social validity what we are interested in? Argument for a functional approach. *Journal of Applied Behavior Analysis, 24,* 205–213.

Hayes, S., Nelson, R., & Jarrett, R. (1987). The treatment utility of assessment: A functional approach to evaluating assessment quality. *American Psychologist, 42,* 963–974.

Hinshaw, S. (1992). Externalizing behavior problems and academic underachievement in childhood and adolescence: Causal relationships and underlying mechanisms. *Psychological Bulletin, 111,* 127–155.

Howell, K., Fox, S., & Morehead, M. (1993). *Curriculum-based evaluation: Teaching and decision making* (2nd ed.). Belmont, CA: Brooks-Cole.

Hynd, G. (1989). Learning disabilities and neuropsychological correlates: Relationship to neurobiological theory. In D. Bakker & H. Van der Vlugt (Eds.), *Learning disabilities: Neuropsychological correlates and treatment* (Vol. 1, pp. 123–147). Amsterdam: Swets & Zeitlinger.

Jacobson, N., Follette, W., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behavior Therapy, 15,* 336–352.

Johnston, J., & Pennypacker, H. (1993). *Strategies for human behavioral research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Kame'enui, E., Jitendra, A., & Darch, C. (1995). Direct instruction in reading as contronym and eonomine. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 11,* 3–17.

Kavale, K., & Forness, S. (1987). How not to specify learning disability: A rejoinder to Koss. *Remedial and Special Education, 8,* 60–62.

Kavale, K., & Forness, S. (1995). *The science of learning disabilities.* San Diego: College-Hill Press.

Kavale, K., & Forness, S. (1999). Effectiveness of special education. In C. R. Reynolds & T. B. Gutkin (Eds.), *Handbook of school psychology* (3rd ed., pp. 984–1024). New York: Wiley.

Kavale, K., & Forness, S. (2000). What definitions of learning disability do and don't say: A critical analysis. *Journal of Learning Disabilities, 33,* 239–256.

Kavale, K., Fuchs, D., & Scruggs, T. (1994). Setting the record straight on learning disability and low achievement: Implications for policy making. *Learning Disabilities Research & Practice, 9,* 70–77.

Kazdin, A. (1984). Statistical analysis for single-case experimental designs. In D. Barlow & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (pp. 285–324). New York: Pergamon.

Keogh, B. (1994). A matrix of decision points in the measurement of learning disabilities. In G. R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities* (pp. 15–26). Baltimore: Paul H. Brookes.

Keogh, B., & Speece, D. (1996). Learning disabilities within the context of schooling. In D. Speece & B. Keogh (Eds.), *Research on classroom ecologies: Implications for inclusion of children with learning disabilities* (pp. 1–14). Mahwah, NJ: Lawrence Erlbaum.

Kern, L., Childs, K., Dunlap, G., Clarke, S., & Falk, G. (1994). Using assessment-based curricular intervention to improve the classroom behavior of a student with emotional and behavioral challenges. *Journal of Applied Behavior Analysis, 27,* 293–323.

Kirk, R. E. (1994). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks-Cole.

Kirk, S. (1962). *Educating exceptional children.* Boston: Houghton Mifflin.

Knapp, T. (1983). Behavior analysts' visual appraisal of behavior change in graphic display. *Behavioral Assessment, 5*, 155–164.

Levin, J. (1986). Four cognitive principles of learning strategy instruction. *Educational Psychologist, 21*, 3–17.

Lentz, E. (1988). Effective reading interventions in the regular classroom. In J. Graden, J. Zins, & M. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for all students* (pp. 351–370). Washington, DC: National Association of School Psychologists.

Liberman, I., Shankweiler, D., & Liberman, A. (1989). The alphabetic principle and learning to read. In D. Shankweiler & I. Liberman (Eds.), *Phonology and reading disability: solving the reading puzzle* (pp. 1–33). Ann Arbor: University of Michigan Press.

Lindamood, P., & Lindamood, P. (1998). *The Lindamood phoneme sequencing program for reading, spelling, and speech.* Austin, TX: PRO-ED.

Lindsley, O.R. (1991). Precision teaching's unique legacy from B.F. Skinner. *Journal of Behavioral Education, 1,* 253–266.

Lovett, M., Borden, S., DeLuca, T., Lacerenza, L., Benson, N., & Brackstone, D. (1994). Treating the core deficits of developmental dyslexia: Evidence of transfer of learning after phonologically and strategy-based reading programs. *Developmental Psychology, 30,* 805–822.

Lovitt, T., Eaton, M., Kirkwood, M., & Pelander, J. (1971). Effects of various reinforcement contingencies on oral reading rate. In E. Ramp & B. Hopkins (Eds.), *A new direction for education: Behavior analysis* (pp. 54–71). Lawrence KS: University of Kansas.

Lyon, G. R. (1996). Learning disabilities. *The Future of Children, 6,* 54–76.

MacMillan, D. L., Gresham, F. M., & Bocian, K. (1998). Discrepancy between definitions of learning disabilities and what schools use: An empirical investigation. *Journal of Learning Disabilities, 31,*

314–326.

MacMillan, D. L., Gresham, F. M., Bocian, K., & Lambros, K. (1998). Current plight of borderline students: Where do they belong? *Education and Treatment of Children, 33*, 83–94.

MacMillan, D. L., Gresham, F. M., Bocian, K., & Siperstein, G. (1997). The role of assessment in qualifying students as eligible for special education: What is and what's supposed to be. *Focus on Exceptional Children, 30,* 1–20.

MacMillan, D. L., Gresham, F. M., Siperstein, G., & Bocian, K. (1996). The labyrinth of IDEA: School decisions on referred students with subaverage general intelligence. *American Journal on Mental Retardation, 101,* 161–174.

MacMillan, D. L., Gresham, F. M., & Siperstein, G. (1993). Conceptual and psychometric concerns over the 1992 AAMR definition of mental retardation. *American Journal on Mental Retardation, 98*, 325–335.

MacMillan, D. L., Siperstein, G., & Gresham, F. M. (1996). Mild mental retardation: A challenge to its viability as a diagnostic category. *Exceptional Children, 62,* 356–371.

MacMillan, D. L., & Speece, D. (1999). Utility of current diagnostic categories for research and practice. In R. Gallimore, L. Hernheimer, D. MacMillan, D. Speece, & S. Vaughn (Eds.), *Developmental perspectives on children with high-incidence disabilities* (pp. 111–133). Mahwah, NJ: Lawrence Erlbaum.

Marston, D. (1987–88). The effectiveness of special education: A time-series analysis of reading performance in regular and special education settings. *The Journal of Special Education, 21,* 13–26.

Marston, D., Fuchs, L., & Deno, S. (1986). Measuring pupil progress: A comparison of standardized achievement tests and curriculum-based measures. *Diagnostique, 11,* 71–90.

Marston, D., & Magnusson, D. (1988). Curriculum-based assessment: District-level implementation. In J. Graden, J. Zins, & M. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for all children* (pp. 137–172). Washington, DC: National Association of School Psychologists.

Marston, D., Mirkin, P., & Deno, S. (1984). Curriculum-based measurement: An alternative to traditional screening, referral, and identification. *The Journal of Special Education, 18,* 109–117.

Martens, B., Witt, J., Daly, E., & Vollmer, T. (1999). Behavior analysis: Theory and practice in educational settings. In C. R. Reynolds & T.B. Gutkin (Eds.), *Handbook of school pyschology* (3rd ed., pp. 638–663). New York: Wiley.

Mastropieri, M., & Scruggs, T. (1985–86). Early intervention for socially withdrawn children. *The Journal of Special Education, 19,* 429–441.

Matyas, T., & Greenwood, K. (1990). Visual analysis of single-case time series: Effects of variability, cerial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23,* 341–351.

Matyas, T., & Greenwood, K. (1991). Problems in the estimation of autocorrelation in brief time series and some implications for behavioral data. *Behavior Assessment, 13,* 137–157.

McCleskey, J., & Waldron, N. (1991). Identifying students with learning disabilities: The effect of implementing state guidelines. *Journal of Learning Disabilities, 24,* 501–506.

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

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741–749.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Nevin, J. (1988). Behavioral momentum and the partial reinforcement effect. *Psychological Bulletin, 103*, 44–56.

Nevin, J. (1996). The momentum of compliance. *Journal of Applied Behavior Analysis, 29*, 535–547.

Noell, G. H., & Witt, J. C. (1999). When does consultation lead to intervention implementation? *The Journal of Special Education, 33,* 29–35.

Nolen, P., McCutchen, D., & Berninger, V. (1990). Ensuring tomorrow's literacy: A shared responsibility. *Journal of Teacher Education, 41,* 63–72.

Nunnally, J., & Kotsche, W. (1983). Studies of individual subjects: Logic and methods of analysis. *Journal of Clinical Psychology, 22,* 83–93.

Ottenbacher, K. J. (1990). When is a picture worth a thousand p values? A comparison of visual and quantitative methods to analyze single case data. *The Journal of Special Education, 23*, 436–449.

Pressley, M., & Ghatala, E. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist, 25,* 19–34.

Reschly, D. J., & Gresham, F. M. (1989). Current neuropsychological diagnosis of learning problems: A leap of faith. *Handbook of clinical neuropsychology* (pp. 503–519). New York: Plenum.

Reschly, D. J., & Grimes, J. (1995). Intellectual assessment. *Best practices in school psychology III* (pp. 763–774). Washington, DC: National Association of School Psychologists.

Reschly, D., & Tilly, W. D. (1999). Reform trends and system design alternatives. In D. Reschly, W. D. Tilly, & J. Grimes (Eds.), *Special education in transition: Functional assessment and noncategorical programming* (pp. 19–48). Longmont, CO: Sopris West.

Reschly, D. J., & Ysseldyke, J. (1995). School psychology paradigm shift. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology III* (pp. 17–32). Washington, DC: National Association of School Psychologists.

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

Reynolds, C. R., & Fletcher-Janzen, E. (Eds.) (1989). *Handbook of clinical child neuropsychology.* New York: Plenum.

Rosenshine, B. (1995). Advances in research on instruction. *Journal of Educational Research, 88,* 262–268.

Rutter, M., & Yule, W. (1975). The concept of specific reading retardation. *Journal of Child Psychology and Psychiatry, 16,* 181–197.

Sechrest, L. (1963). Incremental validity: A recommendation. *Educational and Psychological Measurement, 23,* 153–158.

Share, D., McGee, R., & Silva, P. (1989). IQ and reading progress: A test of the capacity notion. *Journal of the American Academy of Child and Adolescent Psychiatry, 28,* 97–100.

Shaywitz, B., Fletcher, J., Holahan, J., & Shaywitz, S. (1992). Discrepancy compared to low achievement definitions of reading disability: Results from the Connecticut Longitudinal Study. *Journal of Learning Disabilities, 25,* 639–648.

Shaywitz, S., Shaywitz, B., Fletcher, J., & Escobar, M. (1990). Prevalence of reading disability in boys and girls: Results of the Connecticut longitudinal study. *Journal of the American Medical Association, 264*, 998–1002.

Shepard, L., Smith, M., & Vojir, C. (1983). Characteristics of pupils identified as learning disabled. *American Educational Research Journal, 20,* 309–331.

Shinn, M. (1989). *Curriculum-based measurement: Assessing special children.* New York: Guilford.

Shinn, M. (1995). Best practices in curriculum-based measurement and its use in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology III* (pp. 547–567). Washington, DC: National Association of School Psychologists.

Shinn, M., Tindal, G., & Stein, S. (1988). Curriculum-based assessment and the identification of mildly handicapped students: A research review. *Professional School Psychology, 3,* 69–85.

Siegel, L. (1989). IQ is irrelevant in the definition of learning disabilities. *Journal of Learning Disabilities, 22,* 469–478.

Skinner, C. (1998). Prevention of academic skill deficits. In T. S. Watson & F. M. Gresham (Eds.), *Handbook of child behavior therapy* (pp. 61–82). New York: Plenum Press.

Slavin, R. (1987). Grouping for instruction in the elementary school. *Educational Psychologist, 22,* 109–122.

Speece, D., & Case, L. (in press). Classification in context: An alternative to identifying early reading disability. *Journal of Educational Psychology.*

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

Strain, P., Kohler, F., & Gresham, F. M. (1998). Problems in logic and interpretation with quantitative syntheses of single-case research: Mathur and colleagues (1998) as a case in point. *Behavioral Disorders, 24,* 74–85.

Swanson, H. L., Carson, C., & Saches-Lee, C. (1996). A selective synthesis of intervention research for students with learning disabilities. *School Psychology Review, 25,* 370–391.

Swanson, H. L., & Hoskyn, M. (1999). Definition X treatment interaction for students with learning disabilities. *School Psychology Review, 28,* 644–658.

Swanson, H. L., & Sachs-Lee, C. (2000). A meta-analysis of single-subject intervention research for students with LD. *Journal of Learning Disabilities, 33*, 114–136.

Teeter, P. A. (1987) Review of neuropsychological assessment and intervention with children and adolescents. *School Psychology Review, 16*, 582–593.

Teeter, P. A. (1989). Neuropsychological approaches to the remediation of educational deficits. In C. Reynolds & E. Fletcher-Jantzen (Eds.), *Handbook of clinical child neuropsychology* (pp. 357–376). New York: Plenum Press.

Torgesen, J. (1996). A model of memory from an information processing perspective: The special case of phonological memory. In G. Reid Lyon (Ed.), *Attention, memory, and executive function: Issues in conceptualization and measurement* (pp. 157–184). Baltimore: Brookes.

Torgesen, J., Alexander, A., Wagner, R., Rashotte, C., Voeller, K., & Conway, T. (2001). Intensive remedial instruction for children with severe reading disabilities: Immediate and long-term outcomes from two instructional approaches. *Journal of Learning Disabilities, 34,* 33–58.

United States Department of Education. (1998). *Twentieth annual report to Congress on implementation of the Individuals With Disabilities Education Act.* Washington, DC: Author.

Vellutino, F., Scanlon, D., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-achievement discrepancy definition of reading disability. *Journal of Learning Disabilities, 33,* 223–238.

Vellutino, F., Scanlon, D., Sipay, E., Small, S., Pratt, A., Chen, R., & Denckla, M. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. *Journal of Educational Psychology, 88,* 601–638.

Vellutino, F., Scanlon, D., & Tanzman, M. (1998). The case for early intervention in diagnosing reading disability. *Journal of School Psychology, 36,* 367–397.

Wickstrom, K., Jones, K., LaFleur, L., & Witt, J. (1998). An analysis of treatment integrity in school-based behavioral consultation. *School Psychology Quarterly, 13,* 141–154.

Wiggins, J. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Witt, J. C., & Gresham, F. M. (1985). Review of the Wechsler Intelligence Scale for Children-Revised. In J. Mitchell (Ed.), *Ninth mental measurements yearbook* (pp. 1716–1719). Lincoln, NE: Buros Institute.

Witt, J. C., & Gresham, F. M. (1997). Utility of intelligence test for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology*

Quarterly, 12, 249–267.

Wolf, M. (1978). Social validity: The case for subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203–214.

Yeaton, W., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156–167.

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

Ysseldyke, J., & Mirkin, P. (1982). The use of assessment information to plan instructional interventions: A review of the research. In C. Reynolds & T. Gutkin (Eds.), *Handbook of school psychology* (pp. 395–435). New York: Wiley.

Zigler, E., Balla, D., & Hodapp, R. (1984). On the definition and classification of mental retardation. *American Journal of Mental Retardation, 89,* 215–230.

NOTES

Figure 1. Degree of unresponsiveness and intensity of treatment.
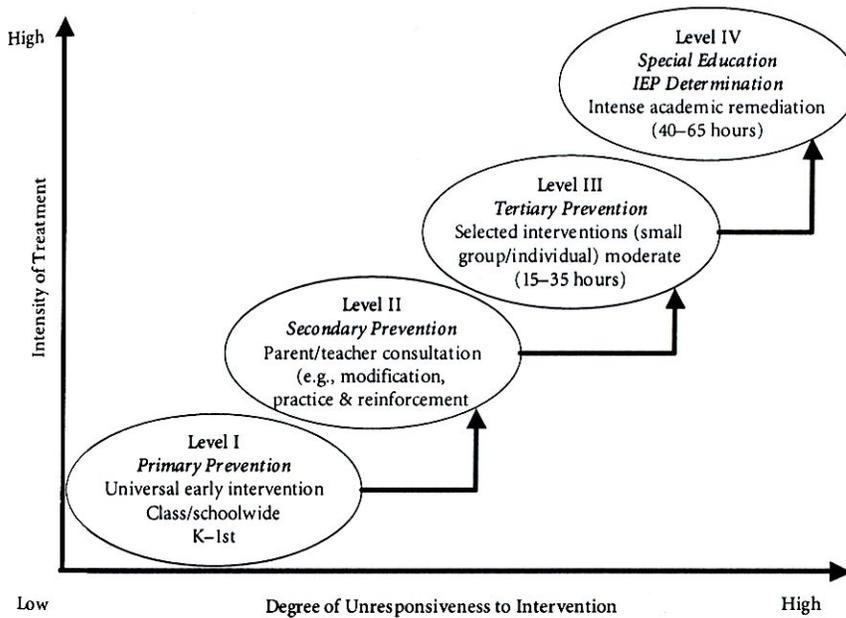
Five Fundamental Principles

1. Intensity of intervention is matched to the degree of unresponsiveness to the intervention.
2. Movement through levels is based on inadequate response to intervention.
3. Decisions regarding movement through levels are based on empirical data collected from a variety of sources.
4. An increasing body of data is collected to inform decision making as a student moves through the levels.
5. Special education and IEP determination should be considered only after a student shows inadequate responsiveness to intervention.