Hindawi Publishing Corporation
ISRN Education
Volume 2013, Article ID 958530, 29 pages
http://dx.doi.org/10.1155/2013/958530

Review Article

Curriculum-Based Measurement: A Brief History of Nearly Everything from the 1970s to the Present

Gerald Tindal

College of Education, University of Oregon, Eugene, OR 97403, USA

Correspondence should be addressed to Gerald Tindal; geraldt@uoregon.edu

Received 8 October 2012; Accepted 13 November 2012

Academic Editors: N. Dumais, F. Jimenez, and L. McCall

This paper provides a description of 30 years of research conducted on curriculum-based measurement. In this time span, several subject matter areas have been studied—reading, writing, mathematics, and secondary content (subject) areas—in developing technically adequate measures of student performance and progress. This research has been conducted by scores of scholars across the United States using a variety of methodologies with widely differing populations. Nevertheless, little of this research has moved from a "measurement paradigm" to one focused on "training in data use and decision making paradigm." The paper concludes with a program of research that is needed over the next 30 years.

1. Introduction

In 2003, Bill Bryson *wrote A Short History of Nearly Everything* and detailed the history of the cosmos, big bang, earth, and the composition and development of life and the population of the planet. The book was massive in detail and comprehensive in scope. In this paper, I take a similar approach and provide a complete guide to curriculum-based measurement, its beginnings, a review of research conducted over the past 30 years, and a glimpse into issues left unattended. Since its beginning in the late 1970s, a number of researchers have come together and provided new insights into understanding how teachers can monitor learning and how learning can inform teaching.

I have divided the paper into sections. Section 1 provides a historical coverage of CBM: the beginnings at the University of Minnesota from 1977 through 1984. The next five Sections 2–6 then address research in various subject areas: oral reading fluency, early literacy, writing, mathematics, and secondary content areas. Finally, Section 7 presents a program of research needed to guide future development, and Section 8 provides a theoretical perspective using the logic of validation from Messick [1] and Kane [2]. An important structural note is that the five subject area sections (reading, early literacy, writing, mathematics, and content in secondary schools) summarize findings from the research and then address different aspects of CBM research. For oral reading fluency, the focus and reflective summary is on the analytical models used to determine the growth within a year and factors influencing reliability of measurement. For early literacy, the focus and reflective summary is on the rich nomological net depicting the interrelations among measures (variables) used to document change in skills as students learn to read. The writing section addresses administration of measures and metrics for scaling behavior in a manner that is sensitive to change for various populations (elementary versus secondary students). In mathematics, the central issue is domain sampling for developing general outcome measures. Finally, secondary content CBM considers access and target skills as students move from learning basic skills to application of them in learning subject specific content.

Each section ends with a reflective interpretation on this issue of measurement methodology that is slightly different than that proposed by Fuchs [3]. In her research analysis, studies were depicted in terms of stages. In the first stage, the focus was on the technical features of static scores (traditional reliability and criterion validity studies). Then, in a second phase, the technical features of growth were considered in documenting the relation between change and proficiency

in an academic domain. Finally, the last stage (three) was reached when the focus of research was on the relevance of the data for making instructional decisions (including teacher acceptance in using these data).

While Fuchs' heuristic is certainly useful and viable, the manner in which research on CBM is analyzed in this paper is quite different. Rather than thinking of the research in a sequential manner (with the earlier phases considered more basic and the last phase more applied), the depiction in this paper is more iterative and targeted at various and more encompassing critical dimensions of a measurement system within a decision making framework, with some dimensions more critical in CBM, given its design for classroom use by teachers in monitoring teaching and learning. Finally, in the last two sections, CBM is depicted in the context of response to intervention (RTI) for evaluating instructional programs and within a theoretical context using an argumentative approach to the process of validation.

The big idea in this document is to summarize the research and findings from the past 30 years to rearrange the focus of curriculum-based measurement for the next 30 years. In short, the past 30 years of research on curriculum-based measurement privileged measurement (as noted in the summary of findings in the five subject area sections) over an argument-based validation scheme on training and data-based decision making (as noted in the last two sections).

2. CBM: The Beginnings at the University of Minnesota

The early research on curriculum-based measurement at the *Institute for Research on Learning Disabilities* at the *University of Minnesota* established two major lines of reports focusing on (a) technical adequacy and (b) teacher decision making. Following is a brief summary of this research, most of which still stands today as relevant and of value in considering continued research that needs to be done in both explicating and extending the findings.

Although not specifically based on the *Standards for Educational and Psychological Testing* [4], the original research addressed the reliability and validity components as defined at that time. In addition to this focus on technical adequacy, several other characteristics were demanded of the measurement system so teachers could use the measures on a regular basis. Deno and Mirkin [5] emphasized specific features of these measures to allow alternate forms so that a time series data display could be used to determine if programs were working or needing to be adjusted. For example, the measures had to be easy to create, quick to administer, usable by all, from the curriculum, and with sufficient technical adequacy. In the end, they laid the groundwork for an experimental view of teaching in which any intervention was considered a hypothesis that needed vindication" [6, p. 32]. This last component was deemed to be a critical component of formative evaluation [7, 8] and development of Individualized Education Programs (IEPs) [9] and assessment of short- and long-term goals [10].

2.1. Institute for Research on Learning Disabilities with Curriculum-Based Measurement. Curriculum-based measurement (CBM) was originally coined in the mid-1970s by Deno and Mirkin [5] in which they articulated two systems for teachers to use in systematically monitoring students' progress over time: (a) mastery monitoring (mm) and (b) general outcome measurement (GOM). To this day, the two forms are considered by the National Center on Response to Intervention (RTI) (http://www.rti4success.org/).

MM is the traditional model in which teachers provide instruction in a specific skill unit and then test students' mastery. As a result, MM has two unique features from a measurement perspective: (a) the domain for sampling items is limited to those taught, and (b) the graphic display reflects a stair step function with time on the x-axis and mastery of a unit (the dependent variable) on the y-axis. The steeper the steps, the faster students are mastering units.

In contrast, GOM is based on comparable alternate forms from long-range goal (LRG) material (skills to be taught over an extended period of time). Because CBM was developed within special education, the LRG is often drawn from the Individual Education Program (IEP). The two unique features of this model are that (a) the domain for sampling items spans the material being taught over an extended time with items that both preview and review skills for any given student and (b) the forms are considered comparable to each other so the graphic display can include slope and variability as the primary dependent variable.

These early features are important to consider as the eventual development of CBM moved on to a more norm-referenced perspective for comparing students to each other in order to identify those at risk of failing-to-learn. Because of the preferred psychometric properties of alternate forms over time leading to slope and variability as outcomes as well as the difficulty in defining and defending "mastery", the field has generally focused on GOM methodologies.

The immediate aftermath of the original publication by Deno and Mirkin [5] was a five-year program of research conducted at the Institute for Research on Learning Disabilities (IRLD) at the University of Minnesota from 1979 through 1983. The focus of the institute was to (a) investigate current practices in identification of learning disabilities, and (b) document the technical characteristics of CBM for monitoring student progress. In total, 23 monographs and 144 technical reports were published through the IRLD with six monographs and 62 research reports written on CBM. The monographs established the critical dimensions of systematic formative classroom measurement system [11, 12].

At that time, few alternative options were present for teachers to use in formatively evaluating their instructional programs. Basal series had mastery tests with little technical adequacy (reliability and validity) in decision making and actually misinformed teachers about skills being learned and practiced by students [13–17]. Informal reading inventories and readability-based estimates of passage difficulty were not (and are still not) reliable [18, 19].

CBMs and time series data on student progress are effective setting events for influencing teachers' decision making and practices as well managing Individualized Education

Programs (IEPs) [7, 8, 10, 20–30]. Some of this research was done because of the unique (statistical) issues present in time series data [31–34] the effects of which are summarized by Marston et al. [35].

Training and structure of instruction is an important component of effective data utilization for systematic use in school settings [30, 36, 37].

The practice of implementing CBM should be planned in its adoption in the schools and may successfully be used by both teachers and students to improve performance [27, 29, 38–40].

CBMs can be effectively used in screening students at risk of failure to learn basic skills [41, 42]. Programs and decisions may be appropriately evaluated at a systems level in a manner that comprehensive ties different decisions together in a manner [35, 42–44].

Domain sampling (breadth and representation) in the curriculum is an important issue to consider in developing progress monitoring probes, influencing sensitivity to change and rate of progress [45–48].

CBMs are reliable and are both related to other critical measures of basic academic skills and can serve as adequate measures of progress in learning basic skills [13, 20–22, 49–56].

In the end, the research is summarized in three monographs focusing on establishing and monitoring progress on IEP goals [57], systemic use of student classroom performance and progress in making multiple decisions [28, 29, 57, 58], and a review of the research over the five years of funding establishing curriculum-based evaluation in reading, spelling, and written expression [59].

The early research also spawned three books in the area of systematic classroom assessment, each addressing unique aspects of this behavioral measurement approach. The books synthesized the research from the institute and researchers responsible for the initial research as follows: (a) Shinn [60], (b) Shinn [61], and (c) Tindal and Marston [62]. The former two books summarized the research completed during the funding of the IRLD and its immediate aftermath, while the latter book extended CBM into a larger, test driven model for measurement development.

2.2. Other Models: Non-General Outcome Measurement. From the beginning, the term begged the question, and several alternative terms were offered. In a special issue of *Exceptional Children*, the term CBM was subjected to a debate. For Gickling and Thompson [63], the term curriculum-based assessment was used to denote administration of mastery monitoring tasks and then using the ratio of known (correct items) to unknown (incorrect items) to diagnostically determine what to teach. At about the same time, Howell [64] developed curriculum-based evaluation in which two types of assessments were administered: survey level and specific level. In addition to the models just noted, these authors added criterion-referenced curriculum-based assessment [65] to describe a mastery-monitoring model with sequentially arranged objectives. Eventually, various models were compared [66], and since those early days, most of the research has been on various dimensions of measurement and decision making. In this early review, the models were compared on several dimensions: relationship to assessment and decision making, underlying premises about the relationship of assessment data to instruction (primarily addressing the prescriptive nature of recommendations to teach), type of student response, focus of material for monitoring student progress (the "shelf life" of the materials), test format (degree of standardization, length, and test construction skills), and technical adequacy (reliability and validity).

With the exception of curriculum-based measurement (CBM), these models have had a limited presence in the research literature in special education. CBM, however, has been extensively researched over the past 30 years, in both general and special education. It has become a term used widely in the special education literature and has become the backbone of response-to-intervention (RTI), serving as the preferred form of progress monitoring used to evaluate instructional programs. Indeed, its staying power has been a function of (testimony to) its compatibility with traditional measurement development and the applications of standards for educational and psychological testing [4]. Indeed, one of the core focuses on CBM was the need for technical adequacy, something which was not embedded into the fiber of the other systems. Indeed, the proliferation of investigations on CBM has been most recently acknowledged in a Festschrift, *A Measure of Success: How Curriculum-Based Measurement has Influenced Education and Learning* [67].

And with this core attention to traditional measurement development, a number of tractable issues can be pursued that would not be possible with non-GOM systems. For example, it is difficult to address slope with mastery monitoring or ratios of correct to incorrect. In contrast, with a solid measurement system based on general outcomes, further research can be conducted on slope, which is addressed next in the context of oral reading fluency. Similarly, major developmental trajectories in skill development are difficult to document with measurement systems across brief periods of time that is typical of non-GOM systems. Yet, with the research on early literacy, it is possible to measure the development of early skills (in letter names and sounds) and their transformation into decoding and fluent reading. With non-GOM systems, research on administration and metrics for scoring and reporting outcomes is restricted, an issue addressed in the area of written expression based on a GOM approach. Ironically, with restricted domains from a non-GOM approach, it is nearly impossible to research the effects of domains; in coverage of mathematics, this area is addressed. Finally, with secondary school measures (of access and target content skill), a non-GOM approach can be used but reflects the traditional manner in which measurement is conducted in the classroom. In fact, a GOM approach in these settings is present with only a tentative foothold accomplished in the past 20 years. These are the topics for the next several sections after which two major issues are addressed: what do we know and need to know in implementing a measurement system for decision making in classrooms and how can a validity argument be deployed in the development of the curriculum-based measurement.

3. Oral Reading Fluency CBM

Oral reading fluency (ORF) has been extensively studied over the past three decades. Since the original study [50]. In a conference paper published by Tindal and Nese [68], the history of research is summarized on slope of improvement for ORF. In the following section, critical text from that publication is summarized. In addition, other topics of generalizability and standard error of measurement are addressed.

3.1. Within Year Average Slope (Growth). The earliest study on average growth in oral reading fluency was by Fuchs et al. [69]. They documented that weekly growth in words correct per minute was 1.5 and 2.0 in first and second grades, 1.0 and 1.5 in third grade, 0.85 and 1.1 in fourth grade, 0.5 and 0.8 in fifth grade, and 0.3 and 0.63 in sixth grade. In 2001, Deno et al. [70] documented growth from nearly 3,000 students tested from four regions of the country. They reported 2.0 words correct per minute growth per week until students achieved 30 WRCM; thereafter, students in the general education population improved at least 1.0 word correct per minute/week.

In the past six years, seven studies have been published on slope of progress in oral reading fluency [71–77]. This recent rise in publications has led to various findings and conclusions.

More growth occurs from the beginning of the year to the middle of the year than from the middle to the end of the year, with overall growth being moderate [74] and more specifically from fall to spring than winter to spring [71]. When quadratic models cannot be computed (because of the limited number of measurement occasions), a piecewise model of growth can be conducted and also fits better than a linear model, with all slopes being positive but negatively accelerated and decreasing as grade level increases [73]. And, as confirmed by Nese et al. [77], slopes are nonlinear for students in grades 3–5 taking benchmark easyCBM measures, again with more growth occurring in the fall than in the winter, at least for students in grades three and four, and more growth for students in earlier grades; students of color, with a disability (SWD), and eligible for free and reduced price lunch perform considerably lower at the beginning of the year with slope only significant for SWD.

Most of this research involves measurement on only two or three occasions, either in the fall and winter or in the fall, winter, and spring. Rarely is instruction considered or progress monitoring being used. In other research on slope that is more oriented to teacher use in the classroom, two explanations have been invoked for different findings on slope.

(1) Steeper slopes may occur because of the lower performance obtained when only one (versus four) baseline measure is used in calculating slope. Furthermore, more accurate estimates occur when measurement is either pre-post or every three weeks [75].

(2) Instructional changes are made more often as a function of weekly goal ambitiousness (1 to 1.5 word per week) but interact with frequency of measurement; when using slope to prompt change, more consistent progress monitoring occurs across conditions of goal [76].

As Tindal and Nese [68] note, the sampling plans for students have been very inconsistent, primarily representing convenience samples for researchers, considerably different in size (from several dozen to thousands), and with different demographic characteristics documented. The results have been reported by grade level and, if special education was noted, it was dummy coded. The measures used in these studies have been equally varied as the students measured and described as generic, standard, or grade appropriate. Likewise, frequency of measurement has been inconsistent from weekly to triannually, with seasonal effects widely interpreted. In general, they report the following.

(1) Growth is not linear *within a year* but likely quadratic. Even in the early research by Deno et al. [70], a large percentage of students showed nonlinear growth.

(2) Growth also is not linear *across years* but reflects much greater improvement in the early grades that tapers off in the later elementary years.

(3) Students receiving special education services enter at a lower level (intercept) and generally grow at a slower rate.

(4) Within year (weekly) growth in oral reading, fluency ranges from .50 words correct per minute per week to 2.0 words correct per minute per week.

An important issue behind this research is the growing sophistication of the estimates of growth. In the original research, growth is simply calculated in a linear manner either as raw score gains or using an ordinary least squares (OLS) regression estimate. Furthermore, this growth is artificially converted to a weekly average by taking the values and dividing by the number of weeks within the span (e.g., year or season). In only two studies were students actually measured on a weekly basis (see the two studies by Jenkins et al. [75] and Jenkins and Terjeson [76]).

In the more recent research, hierarchical linear modeling (HLM) is being used [78] for making growth estimates conditioned on (nested in) various levels [79]. Typically, time is considered level one, student characteristics are considered level two, and teacher or school is considered level three. This kind of estimation provides a more precise estimate as the confounding effects of various levels are controlled. For example, the study by Nese et al. [77] specifically model time in both a linear and nonlinear fashion. This finding of nonlinear growth (within years or over grades) has profound implications for establishing aimlines (or long-range goals) or interpret results. If indeed students grow more from the fall to the winter than they do from the winter to the spring, then teachers would be wise to use this information as mid-year checks on their decision making. More progress should be expected of students in this first four to five months, without which teachers would be making false positive decisions of growth and not changing instruction because they are assuming linear growth. What appears to be an effective

program actually may not be effective, particularly if the teacher is using a normative reference of growth (e.g., how other students are doing).

3.2. Generalizability Theory and Standard Error of Measurement. Separate from research on slope and expected growth for students within the school year, another line of research has developed in the post-IRLD era. In the original IRLD research, the focus was on traditional analyses of reliability (test-retest, inter-judge, and alternate form). In the past decade, however, advancement in the research continues to focus on reliability but with an emaphasis on the conditions under which oral reading fluency is measured. Specifically, two areas of research (with ORF measures) are G-theory and the influence of various data characteristics on the standard error of measurement (SEM).

3.2.1. G-Theory Studies of ORF. In generalizability theory (G-theory), reliability is analyzed in terms of the effects of various measurement facets on estimates of performance [80]. Rather than considering a score on a performance assessment to be comprised of observed score and error, in G-theory, the error is further parsed into constituent components or facets. For example, performance can be a function of the task (or passage in the context of oral reading fluency), the occasion (conditions), or the rater (administrator). These three facets are most typically studied in G-theory; however, the consistent finding is that raters are rarely influential whereas tasks are nearly always influential [81]. Generally, this research involves two phases in which coefficients from these facets are estimated in a G-study, and then a D-study is conducted to ascertain the number of such facets to make a reliable decision (e.g., how many tasks, occasions, or raters are needed?).

While traditional research continued on aspects of ORF measurement as a function of passage difficulty, [82], other perhaps more sophisticated research has employed G-theory to understand the effects of facets on the reliability of performance estimates [82–85]. For example, Hintze et al. [84] used G-theory to examine person, grade, method, occasion, and various interactions. In the first study, participants and developmental changes explained most of the variance with a generalizability coefficient of .90 with two sets of materials and .82 with one set of materials in making intraindividual decisions. For interindividual decisions, the generalizability coefficients were also high (.98) using only three reading passages. In the second study, participants and grade level again explained most of the variance with very little influence from CBM progress monitoring procedures. Generalizability coefficients were .80. Another study by Hintze and Christ [83] used generalizability theory to study passages (both uncontrolled (randomly sampled) and controlled (purposively sampled)) from the reading curriculum. They found an interaction of probe by grade with controlled passages having smaller SE(b).

3.2.2. G-Theory and Standard Error of Measurement on ORF. Poncy et al. [85] documented variability of CBM scores due to student skill, passage difficulty, and unaccounted sources of error reliability coefficients and the SEM given a specified number of probes when making relative and absolute decisions. Using 20 passages from DIBELS, they found that "the largest amount of variation in scores, 81%, was attributable to the person facet. Item, or probe, accounted for 10% of the variance, and 9% of the variation was located in unaccounted sources of error" (p. 331) with the index of dependability ranging from .81 (SEM of 18 WCPM) with one passage to .97 (SEM of 6 WCPM) with nine passages. In a slight variation of analytical research on reliability of oral reading fluency measures, Christ and Ardoin [86] compared four passage construction procedures: "(1) random selection, (2) selection based on readability results, (3) selection based on mean levels of performance from field testing, and (4) use of ED procedures. . .(an estimate of inconsistencies in performance within and across both students and passages)" (p. 59). They reported generalizability coefficients consistent with previous research (.91–.97 for 1–3 passages), but the D study found considerable differences in the passage compositions in favor of the ED procedures in estimating the level of student performance.

For Ardoin and Christ [87], passage stability was investigated (FAIP-R, AIMSweb, and DIBELS) using four dependent variables: intercept, weekly growth, weekly SE(b), and SEE. Statistically significant differences were found "with the FAIP-R passage set resulting in the least amount of error and the DIBELS passage set resulting in the greatest amount of error" (p. 274). In addition, the FAIP-R passages had the lowest intercepts. Finally, measurement error increased with fluency rates. Francis et al. [88] also studied form effects (using DIBELS passages) noting the problems with readability formulae and the lack of actual student performance data to ensure comparability of passages over time. They reported that the passages were not substitutable (nor comparable) in the first assessment and that passages were not equally likely to yield students' median fluency. Finally, they found a significant group by wave interaction: "growth trajectories are significantly influenced by the placement of the more difficult stories (p. 333). Finally, and most recently, Christ [89] reported that the number of data points, quality of data, and method used to estimate growth each influenced the reliability, validity, and precision of estimated growth" (p. 15) with 14 weeks of intervention required for reliable and valid estimates of slope.

3.3. Summary and Reflection on Analytical Models for Understanding CBM. The most significant problem from this last study and all other (previous) analyses on reliability is that nothing about instruction is ever presented or considered. Rather, a limited set of variables for CBM is analyzed: quality of the data set, schedule and administration, and trend line estimation. No questions are asked such as the following, all of which would influence the reliability of score estimation. What is the fidelity of instruction implementation? How well does the intervention match the progress monitoring of oral reading fluency? What grade level of measure is used to progress monitor (e.g., it is hardly ideographic if all students are monitored on grade level passages)? How many data points per intervention phase are present? What

curriculum is used during intervention and how well does this curriculum represent the progress monitoring measures? How often is an intervention changed? How much time are students engaged in what kinds of activities during an intervention? How many students are in instructional groups? How are students identified for Tier II (e.g., what is their initial level of performance on benchmark measures)? What seasonal growth is likely present (i.e., in which season data points are collected, given that several studies have been published or presented at conferences indicating that growth on oral reading fluency is nonlinear)? What cohort effect might be present but is ignored (by data collection year or nested within geography)? How many days are students present during the year? How stable is the sample of students across simulation draws (when used)? How many moves across schools are present by individual students? How many students are with various (and specific) disabilities (or not with any disability)? What are the IEP goals for students with disabilities? If students are with a disability, what is the level and type of instruction implementation using response to intervention (RTI) and how do special and general education teachers plan instruction and interact in its delivery?

The post-IRLD research on the technical characteristics of ORF has made great advancements and moved the field much further along using sophisticated analytical techniques that has uncovered increasingly nuanced issues. Unfortunately, this research has been almost entirely concentrated on ORF and not applied to other CBMs. Furthermore, the research has not yet been contextualized into a nomological net of practice in which important instructional and decision making issues have been presented. For example, all of the previous questions relate equally well to reliability. Finally, in using a Messick frame of validity as argument with a claim supported by warrants and evidence, it appears that the major claims are instructionally free and with little attention to generalizability issues that underlie external and construct validity [90].

4. Early Literacy CBM

A number of skills have been considered in the measurement of early reading, including alphabet knowledge of the names and sounds associated with printed letters, phonological awareness (e.g., ability to distinguish or segment words, syllables, or phonemes), rapid automatic naming (RAN) of letters or digits, rapid automatic naming (RAN) of objects or colors, writing names or letters in isolation, and phonological memory (remembering spoken information). "Six variables representing early literacy skills or precursor literacy skills [have] had medium to large predictive relationships with later measures of literacy development. These six variables not only [have] correlated with later literacy as shown by data drawn from multiple studies with large numbers of children but also maintained their predictive power even when the role of other variables, such as IQ or socioeconomic status (SES) were accounted for" [91, p. vii].

Even though there had been a call to arms on the need to emphasize reading in our public schools with publication of the *Report of the Commission on Reading* [92], it appears that the research to follow was really spurred by three events. First, the *National Reading Panel* [93] emphasized alphabetics (phoneme awareness and phonics), fluency, and comprehension (vocabulary and text), the big five ideas behind effective reading, teaching, and learning. Another reason for the spur in development of such measurement research was the *No Child Left Behind Act* [94], with its emphasis on reading and mathematics in grades three to eight (and a high school year). With this form of high stakes testing, educators realized that reaching proficiency in grade three required attention to the foundation skills in earlier grades. Finally, *Reading First* was funded immediately after NCLB, again providing attention to measures of early literacy. Most of the original research behind CBM did not include measurement of these early reading skills; the landmark review provided by Adams [95] was still six years out from the end of IRLD funding. As Fuchs et al. [96] noted, "in reading, most CBM research has focused on the passage reading fluency task which becomes appropriate for most students sometime during the second semester of first grade. Additional research is needed to examine the tenability of reading tasks that address an earlier phase of reading" (p. 7).

4.1. Evidence on Skill Relations. When students are just beginning to read, students need to master both the graphemic and phonemic components of reading. Furthermore as reading develops, not only are letter names and sounds important, as well as their concatenation, but they form digraphs, rimes, and syllables that are the building blocks of mono- and polysyllabic words. Not only are letter names, letter sounds, and phoneme segmentation critical skills in developing readers, these sublexical features are interrelated, "especially letter sound fluency, may act as the mechanism that connects letter names, letter sounds, and phonemic segmentation to word reading and spelling" (p. 321).

However, it is fluency not accuracy that serves as the essential metric. "Letter name fluency and letter sound fluency, but not phoneme segmentation fluency, uniquely predicted word reading and were stronger predictors than their accuracy counterparts... Fluent recognition of letter-sound associations may provide the mechanism that supports phonological recoding, blending, and accurate word identification" [97, pag 321]. These skills are viewed as sublexical but importantly predictive in laying the groundwork for later word and passage reading fluency [98].

This fluency in the building blocks then becomes important in reading words and passages. Word recognition includes two processes: (a) the ability to decode written words and (b) the ability to decode words instantly and automatically, in addition to psychological and ecological components [99]. Speed in word reading again is a factor in assessment of reading and has long been of interest to researchers [100] and over the years consistently documented its relation with comprehension [101]. Generally, authors invoke the logic of Laberge and Samuels [102] in that automaticity is important from an information processing view.

Letter sound knowledge also has been studied using nonsense word fluency (NWF). In the late 1990s, the research on curriculum-based measurement expanded to include

measures of early reading literacy. Probably the most ubiquitous measure to appear is the *Dynamic Indicators of Early Literacy Skills (DIBELS)* [103]. DIBELS has been used to explore the importance of phonological awareness and phonics as predictors of reading difficulty [104, 105].

In a study by Fuchs et al. [96], higher coefficients were found both predictively and concurrently for word identification fluency (WIF) over nonsense word fluency (NWF) with the Woodcock Reading Mastery Test-Revised. More importantly, they found that "word identification fluency accounts for more unique variance than nonsense word fluency... and word identification fluency slope dominated nonsense word fluency slope in 3 of 4 comparisons" (pp. 16-17). For Ritchey [98], however, both measures were moderately correlated with each other and with word identification and word attack subtests from the Woodcock Reading Mastery Test. Importantly, a ROC analysis showed both measures to have relatively high positive predictive power (in identifying students at risk). NWF was not found to be particularly useful in distinguishing students' emerging skill in learning to blend over time, as the manner in which student respond may change. Likewise, beta weights for DIBELS subtests in kindergarten were moderately correlated with other published measures of reading, and only 32% to 52% of the variance was explained for literacy constructs at the end of first grade [106, p. 349]. Furthermore, letter naming, nonsense word, and phoneme segmentation fluency were as highly correlated with motivation, persistence, and attitude as they were with reading submeasures. Nevertheless, Clemens et al. [107] administered word identification fluency (WIF), letter naming fluency (LNF), phoneme segmentation fluency (PSF), and nonsense word fluency (NWF) as screening measures for 138 first grade students in the fall and another set of reading measures (TOWRE) at the end of first grade. Using a ROC analysis, they reported that the measure with the greatest classification accuracy was the WIF as a significant predictor for each of the outcome variables. Only 3-4 students per classroom were falsely classified. AUC values were shown "ranging from .862 to .909 across outcome measures, followed by LNF (range = .821–.849), NWF (range = .793–.843), and PSF (range = .640–.728)" (p. 238). Only modest improvements were made in combining the measures.

Even though moderate correlations have been reported between various DIBELS measures and published tests, the use of these measures for monitoring progress is really what is critical. And some sensitivity to progress (given instruction) has been reported. Hagan-Burke et al. [108] documented the sensitivity of growth using the DIBELS measures of letter naming fluency, nonsense word fluency, phoneme segmentation fluency, and word use fluency in relation to the Test of Word Reading Efficiency. They reported significant correlations among the measures (with the DIBELS measures loading on a single factor) and significant influence of NWF in predicting the TOWRE (accounting for over 50% of the variance). In another study, both phoneme segmentation fluency (PSF) and nonsense word fluency (NWF) subtests of the DIBELS have been documented to be sensitive to intensive instruction for 47 kindergarten students across four classes in an urban K-5 Title I school [109]. "The results on the DIBELS benchmark assessment scores indicated that the treatment-intensive/strategic students scored significantly lower on the PSF and NWF subtests in the winter (pretest), when compared to the other two benchmark groups with or without treatment" [109, p. 23] and, in a study of the influence of early phonological awareness and alphabet knowledge, Yesil-Dagli Ummuhan [110] studied 2481 ELL students and found, on average, that the ELL students showed a 38 words per minute growth in English ORF throughout the first grade year and were able to read 53 words per minute at the end of the first grade. Finally and more generally, phonemic awareness (PA) instruction has been found to be large and statistically significant for direct measures (d = 0.86), as well as more indirect measures of reading (.53) and spelling (.59). Importantly, PA instruction was effective for students from various backgrounds, including from low socioeconomic backgrounds, at-risk and disabled readers, preschoolers, kindergartners, and first graders, as well as normally developing readers. Instruction was more effective when accompanied with letters, with only a few PA skills, in small groups and over 5–18 hours [111].

4.2. Growth Trajectories in Learning to Read. Sensitivity to instruction and reflection of progress is related to the larger issue of growth trajectories possible (and necessary) in the early years of elementary school. For example, Stage et al. [112] investigated letter sounds and its "relative predictive relationship of kindergarten letter-naming and letter-sound fluency to first-grade growth in ORF" (p. 227). Using four measurement occasions over first grade, they found both letter names and sounds to predict growth in oral reading fluency (even above that of the first ORF occasion to predict later growth in ORF). They also reported that, with eight or fewer letter names, 81% of the students were correctly classified as at risk.

Fluent readers in first grade have also been found to be fluent in second grade [97]. This finding is consistent with other researchers who have reported that letter naming speed appears to be greatest in second grade (accounting for 11% of the variance to just more than 2% in fifth grade on a reading comprehension test [99]. In fact, phonemic awareness in kindergarten has been found to be an important predictor of reading and spelling in fourth and fifth grades [113]. In a study using hierarchical linear models of growth with three levels (time, student, and classroom), Linklater et al. [114] documented growth on initial sound fluency, phoneme segmentation fluency, combined phoneme segmentation, and nonsense word fluency. "ISF in the beginning of kindergarten significantly predicted and accounted for variability on end-of-kindergarten measures of nonsense words, word identification, and reading comprehension" (p. 389). However, considerable variation in growth was also apparent, as a function of time, gender, and initial language status.

In an analysis of growth in reading from kindergarten through third grade, Speece et al. [115] used several measures of oral language, syntax, listening comprehension, phonological awareness, decoding, emergent literacy, and spelling in addition to several background measures (race, gender, SES as measured by free/reduced lunch, family literacy, and

primary language spoken by the child). They found that "the prediction of third grade performance and growth varied by type of reading skill assessed. . .only phonological awareness, emergent reading (TERA-2), and family literacy retained their significance as predictors of the intercept parameter in the conditional model. . .It appears that the unique linguistic roots of word-level reading at third grade are limited to the influence of phonological awareness skill" (p. 328). These findings on the importance of fluency in the basic skills of reading (sublexical features) help explain a finding over 20 years ago: word decoding skills in first grade accounts for 44% of first grade and 12% of fourth grade reading comprehension variance [116].

4.3. Summary and Reflection on Skill Relations in CBM. Measurement of early literacy skills appears to be consistent with the traditional criteria of curriculum-based measures. These skills are straightforward in operationalizing, can generate alternate forms for frequent administration, can be administered by others with a modicum of training, and reflect important general outcomes with sufficient technical adequacy. Unlike oral reading fluency, however, the skills reflect a complex constellation with a relatively brief shelf life. For example, letter naming, one of the earliest skills to develop in kindergarten, is probably not sensitive to progress across years (beyond first grade) for most students. Likewise, in first grade, association of sounds with letters likely has a similar short period for which it is sensitive. Other measures like initial sound fluency as well as phoneme segmentation, nonsense word fluency, or elision tasks are likely to be short lived even though they represent slightly more advanced skills than letter naming or sounding. And, as Vloedgraven and Verhoeven [117] note, three problems still exist with measurement of phonological awareness: (a) the theoretical explanation among various measures as related to a larger coherent construct, (b) the inaccuracy in measurement (and its developmental shelf life), and (c) the difficulty in documenting growth. They reported difficulty parameters that showed the easiest task to be rhyming, then phoneme segmentation, followed by phoneme blending, and finally phoneme segmentation as the most difficult task. They also noted differences in performance on the tasks as a function of grade.

Furthermore, the relation of these skills in development is quite complex. As Mark Twain wrote, "what is needed is that each letter of the alphabet shall have a perfectly definite sound, and that this sound shall never be changed or modified without the addition of an accent, or other visible sound. . .But the English alphabet is pure insanity. It can hardly spell any word in the language with any degree of certainty" [118, pp. 168-169]. Therefore, it is quite uncertain the degree to which the progression of skills advances and the thresholds that are needed for eventual transformation into actual reading (decoding) or fluency (reading accurately with speed). Even though proficiency levels have been identified, the data supporting their application is not inviolate. For example, L. Fuchs and D. Fuchs [119] suggested that students need to complete 35 letter sounds per minute by the end of kindergarten, and Good et al. [104] recommended 50 letter sounds per minute as benchmark goals for nonsense words (with less than 30 indicating the need for intensive instructional support). In the end, oral reading fluency may be more sensitive to monitoring progress than nonsense word fluency [96, 120]. Nevertheless, more research is needed to document the handshake between the relations among the skills and the trajectory supporting normal development.

5. Writing CBM

Written expression curriculum-based measures (administration and scoring) were investigated in the original research at the IRLD for their reliability and criterion relatedness to other published measures [52]. The initial findings indicated that correlations were moderate to high with a three-minute writing time to an unfinished story starter (a prompt providing the student with a scenario to which they were directed to write a narrative) and words written and words spelled correctly. Later, correct word sequences and correct letter sequences were added as metrics for scoring student CBM outcomes [56]. In short order, a number of other researchers expanded nearly all aspects of measurement: administration time, prompts, scoring procedures, and criterion measures. This initial research also was conducted only with students in elementary schools and eventually expanded to secondary students [121–123], who further expanded the outcome indicators to both number and percentage of correct word sequences and correctly spelled words. These later researchers reported stronger correlations (and more obvious group differences) when percentages were used rather than straight counts.

5.1. Administration and Scoring Systems for Elementary School Students. The initial research on written expression [51, 52] was conducted with 82 students from five elementary schools in Minneapolis and St Paul school districts in grades three to six. Like all the initial validation studies at the IRLD in reading and spelling, the focus was on various parameters of a CBM, including amount of time (one to five minutes), directions for administration and scoring, as well as collection of criterion-related evidence. A number of story starters and topic sentences were used, and students' writing samples were scored for average t-unit length (essentially independent clauses), use of mature (infrequent) words, large words, words spelled correctly, and total words written. The latter four measures correlated the highest with the published test (Test of Written Language). In a later study, the emphasis was on growth with significant linear trends reported for words spelled correctly, number of correct letter sequences, and total words written [35]. Finally, the reliability of various outcome indicators were established by Marston and Deno [54], including alternate form, test-retest, inter-judge, and internal consistency. The only research to expand the measures from the original outcome indicators was conducted by Videen et al. [56] who established correct word sequences as a "valid" indicator of writing proficiency in grades three through six. Later research validated these original findings for total words written, words spelled correctly, and correct word sequences with elementary students [124]; ironically,

the countable indices correlated more highly with ratings of story and organization than with conventions.

Nearly a decade later, the focus of research on written expression continued to address various metrics for scoring performance [125]. In this study, total words written, correct word sequences, words spelled correctly, correct punctuation marks, and correct capitalizations, complete sentences, and words in complete sentences were investigated by correlating the measures with teacher ratings (ideas, organization, voice, word choice, sentence fluency, and conventions). Although inter-rater reliability of all measures was adequate (above 90% on most measures and 80% on two of them), test-retest reliability was quite modest except for total words written and words spelled correctly. At best, moderate correlations have been reported among these CBM measures and a standardized test with the relation between teacher ratings and the standardized test much higher.

In a study similar to the original validation research, a number of scoring systems was investigated: words written, words spelled correctly, correct word sequences, and correct minus incorrect word sequences along with different stimulus tasks. The focus was correlating these measures (and tasks) with a standardized test of writing and a state test, as well as language arts GPA. "Students completed two passage-copying tasks and responded to two picture, two narrative, and two expository prompts" [126, p. 554]. Moderately high correlations were found among all the measures. In grade 3, none of the measures showed change from fall to spring, while the 5th grade students (and to a lesser extent 7th grade students) showed noticeable growth on several of the measures.

Very little research has been done on sensitivity to progress with writing CBMs. In one of the few such studies, students were administered traditional story starters and then provided an intervention focusing on the writing process (brainstorming and writing complete sentences) [127]. As in most of the research on written expression, a number of scores were analyzed from the writing samples before and after the intervention: total words written, total punctuation marks, correct punctuation marks, words in complete sentences, correct word sequences, and simple sentences. Ironically, although the number of total words written was the only measure to show a difference from pre-to-post intervention, it did not correlate with a standardized test (in addition to simple sentences).

5.2. Administration and Scoring Systems for Secondary School Students. This research on written expression expanded in both the populations studied and the outcome indicators used to reflect proficiency with research by Parker and Tindal. For example, percentage of words spelled correctly and percentage of correct word sequences were found to be suitable indicators for students in late elementary and middle school grades [121]. In another study with only middle school students [122], a number of production and production-independent indicators were studied: total number of words written regardless of spelling or handwriting legibility, number of letter groupings recognizable as real English words, number of correctly spelled words, number of adjacent and correctly spelled word pairs that make sense together, average length of all continuous strings of correctly spelled and sequenced words, proportion of the total words written that are legible, and the proportion of words written that are correctly spelled. They reported that the percent of legible words, correct word sequences, and mean length of correct word sequences were the strongest predictors of holistic ratings, although growth over six months was limited to total number of words written, number of legible words written, and number of correctly spelled words.

In a later refinement of this work, correct minus incorrect word sequences were added as an outcome indicator with appropriate criterion-related evidence [128, 129]. In the former study, a number of different scoring metrics were used (total words, words correct, correct word sequences, number of characters, and number sentences written). Correlations (with published subtests in both reading and math and English GPA as well as a rating of quality) were moderate (and higher with reading). "In sum, the results of the correlational analyses revealed that four measures—characters per word, sentences, CWS, and MLCWS—had a fairly consistent and reliable pattern of relations with other measures of writing proficiency. Of these four, only sentences and CWS showed divergent validity with correlations higher for the writing measures than for the reading and mathematics measures" (p. 20). Group differences were documented that reflected a sensible pattern with students with learning disabilities significantly below basic, regular, and enriched students. In the latter study [129], a slightly younger group of students was studied (in grades six through eight), an expanded administration time was used (three and five minutes), and story versus descriptive writing was considered. Similar dependent variables were used as the previous study in 1999 (total words, words correct and incorrect, correct and incorrect word sequences, number of characters per word, number words per sentence, as well as correct minus incorrect word sequences) along with teacher ratings (purpose, tone, and voice; main idea, details, and organization; structure mechanics and legibility) and a district writing test used as criterion measures. They reported the highest reliability and criterion validity for correct word sequences and correct minus incorrect word sequences with few differences in administration time or type of writing.

This change in scoring systems across grade levels indicates that "curriculum-based measures need to change as students become older and more skilled" [130, p. 151]. With students in grades four, eight, and ten, the alternate form reliability was investigated with different sample duration or scoring procedures (total words, correct word sequences, and correct minus incorrect word sequences), and criterion-related evidence was collected with a state test. They reported that "alternate-form reliability correlation coefficients decreased with grade level" (p. 159). Although all three scoring systems correlated equally well with the state test (with a slight advantage for CWS-ICWS) and the three administration times (three, five, and ten minutes), they also found "a general pattern of decreasing coefficients with shorter sample lengths for older students" (p. 163).

In another study using a state test as a criterion measure, Espin et al. [131] varied the following variables to investigate

the impact on reliability and validity: time (3 to 10 minutes), scoring procedures, and criterion reference (state writing test). In addition, they investigated impact on primary language of the students. "Although [reliability] correlations were similar across scoring procedure, differences were seen for time-frame" (p. 181) and "the scoring procedure with the strongest reliability and validity coefficients was (CIWS)" (p. 182).

In a study using holistic judgments and a state test as criterion measures, a number of unique scoring procedures were used with high school students, including incorrect word sequences (ICWSs), correct punctuation marks (CPMs), adjectives (ADJs), and adverbs (ADVs), all scored for each 10-minute sample [132]. They reported that "both CPM and ICWS were moderately to strongly related to the holistic scores...[but only] ICWS was moderately and inversely correlated ($r = -.51$) with the norm-referenced scores" (p. 366), though considerable misidentification of learning disabilities was present using various percentile rank cut offs (15th to 30th PR). The number of adjectives or adverbs was not sufficiently reliable or correlated with either holistic judgments or the state test score.

This line of research expanded to include expository essays and a much longer period of time to write (35 minutes versus the traditional three to five minutes). In this last study with a relatively small sample size (only 22 middle school students), "criterion variables in the study were the number of functional essay elements and quality ratings of the essays" [133, p. 210]. They reported moderately strong correlations among correct and correct-incorrect word sequences and both functional elements as well as quality ratings. Interestingly, the correlations decreased only somewhat when only the first 50 words of the essays were scored (particularly with the quality rating and for the students with learning disabilities). Finally, growth over time was documented from pre- to posttest.

To use criterion measures that teachers would find more credible than standardized tests or state tests, Fewster and Macmillan [134] studied student performance from late elementary school through high school and correlated it with teacher grades. Students were given three (annual) administrations of both a reading measure (oral reading fluency) and a three-minute story starter from which to write a composition (scored by counting the total number of words written (TWW) and the number of words spelled correctly (WSC)). In correlating these measures with teacher grades (in English and social studies), they reported that "in nearly every case, WRC [words read correctly] correlated more highly with course grades than WSC (p. 152)" with most in the moderate range and accounting for 15% to 21% of the variance.

Finally, in the most recent study with 447 eighth grade students, Amato and Watkins [135] studied the predictive validity of traditional CBMs in writing. A number of metrics were used: total words written, words spelled correctly (both count and percent, correct word sequences (both count and percent), correct-incorrect word sequences, number of sequences, number of correct capitalizations, and number of punctuation marks (both total and correct). Using the Test of Written Language (TOWL) as a criterion, they reported high reliability for the CBMs (among 10 scorers) and moderate correlations. Only percentage of correct word sequences, number of correct capitalizations, and number of correct punctuation marks were significant.

5.3. Summary and Reflection on Administration and Scoring Systems in Writing CBM. McMaster and Espin [136] reviewed this body of work including the initial IRLD research and the decade in which the research was initially expanded by Tindal and colleagues, as well as Espin and colleagues. Researchers focused on reliability, validity, and sensitivity to growth with the initial research. In their research with elementary students, they addressed students at different skill levels, screening decisions, scoring procedures, and beginning writers. In their coverage of secondary students, the issues were the integration of reading and writing, attention to students requiring remedial and special education, a focus on screening and progress monitoring, consideration of scoring procedures, type task and sample duration, and, finally, predictions of performance on school-based indicators. They concluded by noting the need to address reliability in progress monitoring and attention to generalizability and consequential validity of measures. "Finally, if measures are used by teachers to monitor progress and make instructional decisions, it is necessary to demonstrate that student performance improves as a result" (p. 82).

For the past 30 years, the research on written expression has addressed a number of test administration variables. Clearly, the writing process needs to address the stimulus prompt for soliciting student writing. In addition, the amount of time to write is a critical variable. Finally, the system used to generate a score needs to be carefully considered. Obviously, all three of these variables need to be studied in terms of reliability and validity, particularly criterion-related evidence to be both predictive of performance on other measures but perhaps more importantly to be credible for teachers (and therefore be predictive of their own judgments obtained through grades or holistic ratings). Unlike all other areas in CBM, these test administration variables have been studied, often with similar results.

In summary, by the late 2000s, a sufficient number of studies had been conducted in written expression CBM to indicate that (d) scoring written products needed to be different for middle school students (from the previous use of words written and words written correct with elementary students, (c) percentages of words written correctly (WWC) and correct word sequences (CWS), as well as correct minus incorrect word sequences (CIWS) were more sensitive metrics (in terms of criterion related evidence), (b) amount of writing time influenced reliability but not validity, particularly for older students, and (a) genre for writing was not an influential variable (on either reliability or validity). Furthermore, the various measures of written expression appeared to be correlated with published writing tests, state tests, and teacher grades.

6. Mathematics CBM

As of 2008, most of the research on CBM has been in reading, not mathematics. In fact, "an exhaustive literature search on the development, uses, and implementation of CBM yielded 578 reports, dissertations, chapters, and articles (Foegan et al., [137]). A breakdown by subject area showed that only 32 were conducted in the area of mathematics. Additionally, of the 32 studies exploring CBM in the area of mathematics, 28 were conducted at the early mathematics and elementary levels, whereas only were 4 implemented at the middle school level" [138, p. 236].

The issue of domain sampling is critical in mathematics given the somewhat qualitative shifts in content represented in this subject area over the grades. That is, students in elementary schools learn addition, subtraction, multiplication, and division in a relatively straightforward manner but often without considering the conceptual nature of the number system. This issue is even more pronounced in middle schools with introduction of geometry, algebra, and statistics/probability. As Fuchs notes, "CBM differs from other forms of classroom-based assessment that rely on mastery measurement. With mastery measurement, teachers assess mastery on a problem type, such as adding 2-digit numbers with regrouping and after mastery is demonstrated, move on to the next skill. By contrast, with CBM, every skill in the annual curriculum is represented on each weekly test. So, like a norm-referenced test, CBM samples a broad range of skills, and each repeated measurement is an alternate form of equivalent difficulty, assessing the same constructs" [139, p. 34]. As noted earlier, a mastery monitoring approach is confined to assessing skills from a limited and instructionally delivered domain. In mathematics CBM research, this approach has been labeled as a "subskill mastery measure (SMM)" versus a "general outcomes measure (GOM)" [140]. Another labeling system used by Foegen et al. [141] refers to the domain as "curriculum samples" (from skills taught) or "robust indicators" (reflecting core competence with high likelihood of being related to important curricula). In all of these labels, the key issue is the number of skills represented in the domains (and often hidden in the GOM perspective).

6.1. Domains Sampled in Elementary Schools. In mathematics research, both the domain sampling used in specific studies and the study purposes have been extremely varied, making it difficult to bring consistency to the outcomes. For the most part, this body of research is similar in many ways to that done with any of the other CBM subject areas. A typical range of skills includes computation, applications, and (real life) problem solving [139]. Given the focus of elementary school curriculum on basic operations, it is not surprising that computation has been the focus of considerable CBM research with students in grades 1–5 [142–147]. In a study by Burns et al. [142], single and multiple skills were compared with students in grades three to five. They reported sufficient reliability and moderate predictive validity for older grade band students using three levels of placement (frustration, instructional, and mastery).

However, given that even computation skills require even more fundamental skills without the use of symbols, considerable research also has been conducted on these skills. For example, Clarke et al. [148] included four types of tasks in their research: (a) oral counting, (b) number identification, (c) quantity discrimination, and (d) missing numbers. In addition to analyzing growth, two published standardized tests were used as criterion measures. Although they did not find linear growth, "correlations between the experimental measures in the fall and criterion measures in the spring were in the moderate range" (p. 10). Likewise, Chard et al. [149] used five tasks: counting 1–20, from 3 and 6 and by 2 s, 5 s, and 10 s, identifying numbers (1–20), writing numbers (1–100), discriminating quantity (1–20), and identifying missing numbers (1–20). Lembke et al. [150] had students differentiate quantity (name larger of 2 numbers), name missing numbers, and name randomly ordered numbers. Martinez et al. [151] had students count aloud from 1, name or identify digits from 1 to 20, discriminate the larger of 2 numbers, and identify the missing number in a sequence.

Preschool and kindergarten CBM probes also have included choosing a number (20 items each with four options), counting objects (of specific pictured things with up to 20 pictures), counting (from 1 to the highest number possible up to 30), and visually discriminating among four unique objects (with 30 such trials). All of these tasks have sufficient reliability. Three kindergarten probes used for criterion-referenced evidence included counting circles, counting letters (sounded out), and discriminating the unique item from four pictured [152].

Some of the research includes both computational and requisite skills. For example, Fuchs et al. [144], first grade measures, included fact retrieval (25 addition fact problems with answers from 0 to 12 as well as 25 subtraction fact problems with answers from 0 to 1), computation (problems with two single-digit numbers requiring adding or subtracting), problems with three single-digit numbers requiring adding, and problems with two double-digit numbers (requiring adding or subtracting without regrouping), number identification and counting (four-item tests that present students with number sequences, the last two of which are shown as blanks, for example, 4, 5, 6, ...), and finally concepts and application (25 items including numeration, concepts, geometry, measurement, applied computation, charts and graphs, and word problems). Another example of broad domain sampling includes Leh et al. [153]. On occasion, the tasks vary whether the problem is fully presented (and the student is directed to supply the answer) or the problem is incompletely rendered but with a solution (and the student must create the conditions for the solution shown).

Sensitivity to instructional programs has been studied, often on the efficacy of tutoring. For example, on first grade math (math concepts, procedural calculations, word problems, and math facts), an effect size of .40 has been reported with follow-up research addressing math fluency in first grade and word problems in third grade that showed an "effect size compared to the control condition was 0.27 for math facts tutoring but almost double that for word-problem tutoring (0.53)" [154, p. 261]. A final study showed the effects

of tutoring with "validated instruction" resulting in a very large ES of 1.34.

6.2. Domains Sampled in Middle Schools. Ann Foegen has been the most dominant researcher in middle school mathematics and has addressed a number of problem types. Foegen's [155] early work on middle school math focused on both a fact probe and an estimation probe (both computation only and as word problems). Using teacher ratings and a standardized achievement test as criterion measures, she reported sufficient levels of criterion-referenced correlations over 10 weeks. She also noted that reliability increases when successive probes were aggregated but slope of improvement remained moderate. She extended this work [137] by addressing low achieving students, administering a basic math operations task, a basic estimation task (both computational and word problems), a modified estimation task (with changes in the number types), along with ratings from teachers, grades from students, and performance on a standardized test. Multiple types of reliability were computed (and found to be adequate); in addition, criterion-related evidence (correlations) among the CBMs and teacher ratings were moderate.

Finally, Foegen [156] used *The Monitoring Basic Skills Progress* (MBSP) measures for computation [157] and for concepts and applications [158] (reflecting skills and concepts from the Tennessee state mathematics curriculum); in addition, she used basic facts, estimation, complex quantity discrimination, and missing number. She reported moderately high levels of alternate form reliability (.80–.91) in grades 6–8 (fall, winter, and spring) and test-retest reliability (.70–.85). Using teacher ratings and a standardized achievement test as criterion measures, she reported sufficient levels of criterion-referenced correlations and moderate rates of growth.

6.3. Summary and Reflection of Mathematics Domains with CBM. Four reviews have been published on mathematics CBM. In this chronological summary, the highlights of these reviews are presented.

Foegen et al. [141] comprehensively review two approaches to developing CBMs in math: curriculum sampling versus robust indicators. Rather than a mastery approach, curriculum sampling reflects items from within a year "given the high level of curriculum specificity in mathematics" (p. 122); in contrast, robust indicators reflect problems that "identify aspects of core competence in mathematics [that] enables students' growth to be modeled over multiple years of learning" (p. 122). This review analyzes results from 32 reports on progress monitoring summarizing reliability and validity research (stage 1) as well as student growth (stage 2). Results are summarized for early, elementary, and secondary mathematics. The seven studies on using data to improve student achievement have all been conducted by Fuchs and colleagues with six of them focused on computation.

Lembke and Stecker [159] summarize procedural issues in implementing math CBMs, including steps in data-based decision making (screening and progress monitoring), metrics for summarizing performance (correct digits, problems, or responses), and establishing goals. The review is then structured on identifying reliable and valid measures for screening and progress monitoring, using progress monitoring for students with special needs, analyzing specific skills, and using the measures as part of class wide peer tutoring and consulting with teaching.

Christ et al. [140] review computation CBMs and use Messick's validity framework [1] relating content and criterion-related evidence while also noting the lack of research addressing consequential validity. In particular, two domain samples are considered: sub-skill mastery measure and general outcomes measures as well as stimulus sampling (curriculum and robust indicators); clearly the two features are related. Most reliability studies address internal consistency and alternate forms; no research is presented on test-retest reliability. A key issue is the relative stability of single skill measures versus the greater variance associated with multiple skill measures, particularly when randomly placed on the measure. Finally, duration of administration is summarized. "A single brief assessment of 1 to 2 min is probably sufficient to assess most single skills. Longer and more numerous assessments are necessary to assess multiple skills" (p. 203).

For Fuchs et al. [144], the focus of their review of four studies was on the capability of a response to intention (RTI) system in mathematics to reduce the need for intensive, long-term service. The studies address first grade tutoring, math facts fluency in first grade, math facts tutoring in third grade, and finally tutoring with word problems. Although they describe significant effects from the first grade tutoring (effect size of .40), they also noted the low relative performance level of the students (compared to students not at risk); therefore, another study was implemented with an emphasis on tutoring for fluency. In the third study, transfer from math facts to word problems was the focus for third grade student with effect sizes reported for math facts of .27 and word problems .53, suggesting lack of transfer. Finally, in this third study, the interaction of tutoring with validated classroom instruction was the focus that was found to be significantly more effective than when implemented with teacher-designed instruction.

Following the initial research on reading and writing at the University of Minnesota, there has been a very significant increase in the research on mathematics. Quite likely, the federal legislation (NCLB) that began in 2001 had an influence as the two areas in which districts were held accountable was reading and mathematics. And, like the early literacy skills, waiting until grade 3 made little sense so much of this research spanned the entire grade range of K-8. Also, like the early literacy skills research, mathematics is comprised of multiple domains that eventually need to be stitched together. However, an important difference is that these domains successively build upon each other. Whereas learning to read involves generalizations with many exceptions to the rules (of sounds and syllabication), learning to compute and solve mathematics problems is lawful with successive algorithms used. For example, an early skill in counting (using objects and finger counting) turns into addition and then multiplication. A number of principles are applied in solving problems so that students can learn more general case

strategies (e.g., distributive and associative principles). Therefore, the field of CBM research addressed domains that had to be integrated into long-range goal measurement, a feat more similar to the development of general outcome measurement in secondary settings that is addressed next. That is, to avoid a mastery-monitoring model (with domains of assessment tightly associated with instruction), a more general case sampling was developed, integrating a range of skills.

7. Secondary Content CBM

Tindal and Espin have contributed the most research for CBM in secondary settings, with the essential difference between them being the source for ensuring measurement comparability and applicability to content area instruction. A comprehensive comparison of these two approaches is published by Espin and Tindal [160] addressing critical issues such as the curriculum in secondary classrooms, the skills to be assessed in CBM, research in reading, writing, and mathematics, and finally a comparison of the two approaches. In general, Espin has addressed vocabulary as the key to understanding measurement in secondary settings while Tindal has addressed content concepts.

7.1. Vocabulary Assessment in Secondary Schools. Much of Espin's research laid the groundwork for this area (correlating reading aloud from content text, completing a maze test, and matching vocabulary words with definitions); consistently, low moderate correlations have been reported among these measures, often highlighting the importance of vocabulary.

Quirk [161] notes that words in secondary settings are Graeco-Latin in origin and distinguishes them from English words having an Anglo-Saxon origin, which are typically short (monosyllabic), learned early in life, and form the basis of everyday discourse. When students move from elementary schools to middle schools, the nature of reading changes drastically. The words of secondary content are typically multisyllabic, sound foreign, contain morphophonemic structures that are only somewhat transparent, and are learned later in life [162]. They comprise the primary language of the secondary content classroom and pose a more serious problem for students at risk of failure, "the component parts of Graeco-Latin words no longer carry much meaning for most uses of the words in English. The words have lost their semantic transparency" [162, p. 689]. Consequently, they typically are abstract, do not appear frequently in the English language, and are not easy to understand. "When these features combine in words, they interfere with word use and with word learning" [162, p. 696], and as a result, meaning is derived through whole word search which is slower. Furthermore, in secondary settings, the lens for learning is not spoken language or conversation. "Academic Graeco-Latin words are mainly literary in their use. Most native speakers of English begin to encounter these words in quantity in their upper primary school reading and in the formal secondary school setting. So, the word's introduction in literature or textbooks, rather than conversation, restricts people's access to them" [162, p. 677]. This change has significance in the demands made on students and is likely part of measurement systems that are "curriculum-based."

Initially, Espin's focus was on reading aloud from content area text books [163]. In this study, 124 students in grade 10 were administered a background knowledge vocabulary test (matching a word with its definition with 10 words sampled from the content text), reading passages (both pre and post), and a classroom study task (with a multiple choice test that included literal and inferential questions) in English and Science. They reported high reliability (both alternate form and test-retest) as well as a pattern of performance for most students with disabilities (5 of 7) that was consistent with a general deficit in both content areas (below the 40th percentile rank) with only 28 of 102 general education students in this group; no differences were found in special education comparing general deficit versus content-specific deficit performance. However, when comparing these two groups following a period of study in the text, the two groups were different (with higher performance by content-specific deficit students in the number of words read correctly per minute). Finally, they reported differences between these two groups on background knowledge.

The study was completed by Espin and Deno [164] with 10th grade students ($n = 121$) who were given English and Science texts to read (for one minute) and study (for 30 minutes in preparation for completing a 25-item multiple choice test). They used grade point average and performance on a standardized achievement test as criterion measures. They reported reliable and positive correlations between reading measures and all performance measures (in the range of 40 s to 70 s), accounting for 11% to 28% of the variances. The correlations with reading fluency and study questions and published achievement tests, as well as grade point average, were about the same value.

In a similar study, Espin and Deno [165] measured 120 tenth grade students with four tasks in both English and Science: vocabulary matching (10 items from the study passage), prestudy reading aloud, content area study task (900-word passages with 25 multiple choice (literal or inferential) questions to answer), and poststudy reading aloud (which were not analyzed). They reported moderate correlations among the measures. Text-based reading added 13% of the variance to performance on the content area study task with vocabulary subsequently adding an additional 9% of the variance; this finding was true for both English and Science (though with slightly more positive results for English). When vocabulary was entered into the regression analysis first and subsequent analyses included text-based reading, no additional variance was accounted. Separate analyses for students from both ends of the distribution showed the test-based reading passages to function as well as vocabulary measures. In a follow-up study, they also reported that the vocabulary test was slightly higher in its correlation with a content study task (and accounted for more of the variance) than oral reading of the content text (counting number of words correct per minute).

Espin and Foegen [166] added a maze test to the reading aloud from text and vocabulary test as secondary measures; in this study, 184 students in grades six to eight participated

by taking these three measures and then taking three criterion measures: a 10-item multiple choice comprehension test, daily tests (from five timed reading passages with 10 multiple choice questions), and a posttest comprised of a compilation of multiple choice questions from the daily tests. The results showed moderately high correlations among the three CBM tasks and the three criterion measures though vocabulary was the best predictor measures. Again, in a series of stepwise regression analyses, vocabulary appeared to contribute the most unique variance (when it was added first, no more variance was accounted for by reading aloud or maze).

Espin et al. [167] in a later study that directly addressed the vocabulary measure as a generalized indicator of performance in content area; in this study, text from the classroom of instruction was used. They administered a number of measures to 58 seventh grade students (five of whom were receiving special education services); a number of vocabulary matching measures (with 49 terms and definitions used in each area of sociology, psychology, and geography), both student read and administrator read, were administered along with a standardized achievement test and a classroom knowledge test (36 items using factual and applied knowledge). Alternate form reliability was moderate (but higher when multiple probes were aggregated); the correlation was moderately high between the vocabulary probes (particularly the administrator read measures) and the published test as well as the classroom post-test, though moderate-low with grades.

The study conducted by Espin et al. [168] continued this research on vocabulary matching in social studies with more focus on progress and administration format (student versus administrator). Using the same population as used earlier by Espin et al. [167], they analyzed the data using hierarchical linear modeling to compare growth for the two administration formats. They found more growth in correct matches with student (.65) than administrator (.22) read tests. In a comparison of three criterion measures as level two predictors, they found all to contribute significantly to vocabulary knowledge.

In the most recent study, [169] presented the results from two studies. In the first study, two measures were administered to 236 students in grade eight (a read aloud task administered with timing noted at one, two, and three minutes and a maze task with timing noted at two, three, and four minutes); both tasks were based on human interest stories in the newspaper. Performance on these tasks was then correlated with a state test. Alternate form reliability was quite high and both correlated with the state test.

7.2. Concept-Based Assessment in Secondary Schools. As a basis for framing a concept-based approach, Nolet, and Tindal [170] divide content into three knowledge forms [171]: facts (single dimensional declarative statements), concepts (with a label, attributes, and both examples and nonexamples), and principles (if-then relations that present cause-effect relations). With concepts highlighted as the key knowledge form, attributes and subsequently examples and nonexamples are then articulated for each concept [172]. Attributes are essential for ensuring comparability of alternate forms, a critical feature of CBM. This approach is in contrast to the Direct Instruction approach in which the focus is on the juxtaposition of examples and nonexamples as the basis for defining concepts [173].

The reason for taking this route to curriculum-based measurement in secondary content areas is threefold. First, content (information) becomes much more subject specific in middle and high school grades and less generalizable outside of the subject area. Second, this content specificity is likely to be noteworthy even within a subject area, favoring a mastery monitoring approach rather than a general outcome measurement (GOM) methodology. Third, students with disabilities or those at risk are likely to have difficulty reading in addition to learning subject matter content, requiring a measurement system to do basically double duty, in serving as a barometer of reading AND content learning. The subject specificity (both across and within content areas), however, is the most troubling issue to address in generating a GOM approach. An artifact of this issue is that content information becomes difficult to traverse; every new content area is new with little opportunity to preview and review of information.

7.2.1. General Description of CBA. In the concept-based assessment (CBA), three to four primary attributes per concept are identified in a content area from content experts who use facts in the curriculum to inductively derive them. "A concept-based focus provides the teacher with a template for specifying the domain-specific conceptual knowledge with corresponding explicitly identified attributes and examples and nonexamples" [174, p. 335]. Attributes provide critical rules for organizing the domain of facts with generalizability to novel content in the same subject area. For example, Twyman and Tindal [175] define "nationalism as a concept [with] three attributes as: (a) has an ideology based on devotion and loyalty to a nation, (b) primary emphasis on promotion of its culture and interests, and (c) identifies with the power of historical tradition to oppose those of other nations or supranational groups" (p. 5). As another illustration of a major concept in social studies, Twyman et al. [176] use four critical attributes to characterize civilization: (a) having religious beliefs, (b) forming social groups, (c) generating support systems and activities, and (d) employing writing as a primary form of communication. In both illustrations, the concepts and their corresponding attributes can be applied to novel content with new examples.

7.2.2. Critical Effects from CBA. One of the important side effects from this approach is the immediate reduction in complexity of content information with clear explication of critical information. Furthermore, the concept-based approach leads to assessments that take advantage of constructed responses with explicit guidelines for a partial credit-scoring algorithm [177]. For example, a student's response can be scored for framing an argument, as well as providing both a rational and details that support the rationale with up to seven points possible [174]; alternatively, a multistep flow chart, with binary decisions, can provide a score of 0 to 4 points [172, 175]. In either partial credit system, reliability is enhanced by using an explicit scoring guide [178]. A concept-based approach also provides an explicit link to

instruction through the use of semantic maps [179] or the verbal interactions between teachers and students during instruction [180, 181]. Furthermore, the approach provides a mechanism for special education teachers to collaborate with general education content teachers in an effective manner [182]. Finally, both teachers and researchers can begin to integrate curriculum content, instructional delivery, student perception, and formative test performance that provides essential feedback [183]. The approach is particularly useful in linking assessment with instruction and is universal across content areas with application in the social and physical sciences [181].

7.2.3. Empirical Outcomes from CBA. The approach has empirical support from several studies. For example, Twyman et al. [176] showed that by anchoring both instruction and assessment on concepts, students form a deeper understanding of content. Although students using a concept basis showed no difference on tests of facts (declarative knowledge), they were significantly better on tests of vocabulary and problem solving. This approach has been used in a case study with a seventh grade student for whom English was his second language [176]. In a unit on four Mesoamerican civilizations with a cloze technique using examples and nonexamples of religion, social groups, support systems and activities, and writing, a control group of students taught in a traditional factual approach performed more poorly on a writing task than students in the experimental CBI group. In a similar study on history, Twyman et al. [174] used three types of measures (fact test, vocabulary knowledge, and problem-solving essay) to compare two groups of students taught with CBI and traditionally (as factual information) over 21 days (with each group for 46 minutes). Although both groups improved significantly on the facts test, students in CBI improved differentially more than the traditional group for the other two measures.

7.2.4. Research to Practice. Two training modules provide an overview of concept-based instruction and assessment in middle-high schools (Tindal et al. [184] and Nolet et al. [185]). In the former, a complete description is provided for understanding content information in terms of knowledge forms (facts, concepts, and principles) as well as designing instruction for higher order intellectual operations that addresses curriculum analysis, instructional planning, and interactive teaching. An appendix is included with black line masters for various semantic maps to organize concepts. The latter publication provides companion training in development of assessments for understanding students' ability to reiterate, summarize, illustrate, evaluate, predict, and explain. In addition, a number of important assessment issues are addressed including sampling planning and indexing reliability, both of which are necessary for development of assessment problem-solving tasks. An appendix provides black line masters for a decision-logic scoring guide using constructed problem-solving essays. These initial training materials were later adapted with empirical references and applied to a consultation [182] and learning assessment system [172] for special education teachers in middle schools settings. In addition, a number of training manuals are available in middle school science [186], high school science [187], mathematics [188], language arts [189], and social science [187].

7.3. Summary and Reflection of Access and Target Skills for CBM. Assessment in secondary settings did not develop until well after the initial research at the University of Minnesota IRLD. And the research that did eventually emerge was considerably different. No longer looking at measurement of access skills, the focus directly targeted the curriculum content. For both researchers and programs of research, this target reflected a challenge to a GOM perspective as well as other principles underlying the initial CBM research platform. To maintain a time series (and long-range goal sampling) design, target content had to be approached from a broad perspective that nevertheless maintained relevance for teachers to use in evaluating instructional programs. Espin and colleagues used vocabulary (knowledge) as their content medium while Tindal and colleagues addressed concepts (attributes and examples-nonexamples). Both approaches distilled measurement to a command of academic words and helped remove reading as a necessary access skill which would have presented a tautology. Students could not reach proficiency on content unless they learned to read, and they would not be able to read (content) unless they knew the academic words. However, neither provided quite the elegance attained with other features of CBM (like easy to create and implement as well as administer and score).

8. Analysis of Current Practice and a Proposal for an Alternative Conception

The most dominant use of CBM has been for screening and progress monitoring. Even before response-to-intervention (RTI) was common parlance among educators, CBM offered a system for monitoring students attainment of goals on their IEPs and evaluating instructional programs. As noted in the early history, the measures were validated and used to make a variety of decisions. Of late, however, their use in a formal system such as RTI has become widespread among school districts in the United States.

The National Research Center on Learning Disabilities (NRCLD) defines RTI as "student-centered assessment models that use problem-solving and research-based methods to identify and address learning difficulties in children [through] high-quality classroom instruction, universal screening, continuous progress monitoring, research-based interventions, and fidelity of instructional interventions" [190, p. 86]. RTI depends upon the coordinate relation between implementation of interventions and change on measures over time so that the interventions are either vindicated or modified for students with disabilities or who are at risk of failure. For Wanzek and Cavanaugh [191], fully developed RTI models also integrate general and special education to identify and integrate school resources in providing effective instruction and intervention. Furthermore, scientifically based evidence is used to develop effective instruction including the intensity of instruction (time, frequency, duration, and instructional group size).

8.1. Empirical Results and Needed Research. Systemic research on RTI is more conceptual than actual. For example, Barnett et al. [192] conceptualized technical adequacy of RTI from Messick's [193] evidential (efficacy and effectiveness) and consequential perspective. They present a technical adequacy checklist for each of three instructional tiers "emphasizing technical checks and iterative use, via a fictitious example" (p. 26). More recently, VanDerHeyden [194] added the following to considerations of technical adequacy of RTI implementation in which "assessments must be technically adequate for the purpose for which it is used [and] decision rules must be correctly applied with correct actions occurring in sequence" (p. 336). In the end, her focus was on sensitivity and specificity of decisions (and subsequent errors) using a receiver operating curve (ROC) analysis.

The problem with both previous analyses (and a number of similar publications) is not the researchers' perspectives but the need to follow up on their perspectives. Or, as Gersten et al. [195] note, "fully determining the validity of an assessment process transcends what any one researcher can accomplish. It is a task for a community of researchers and practitioners to consider meanings and utility of assessment procedures in relation to current thinking about how to improve instructional practice and issues raised by studies of implementation" (p. 512).

8.1.1. What We Know. Two factors appear particularly influential in how well teachers implement CBM: adequacy of planning time and teachers' degrees of personal and/or teaching efficacy [196]. Even though five indices of implementation were examined (number of measurement points, ambitiousness of the goal set for the student, number of times student goal was raised, number of times teachers made instructional changes, and timing of changes), she reported that "teachers with high personal efficacy and high teaching efficacy increased the end-of-year goal for students more often than their counterparts with low degrees of efficacy; teachers with high teaching efficacy set goals that were overall more ambitious than those of teachers with low teaching efficacy" (p. 5). In a related and later study, Allinder [197] also reported that when teachers implemented CBM more accurately, their students made significantly greater math gains than the students with teachers who either implemented CBM less accurately or who did not use CBM.

This research on teacher use of CBM is based, in part, on measurement of, and intervention on, teacher behavior. For example, Allinder and BeckBest [198] used an accuracy of implementation scale to investigate teachers' CBM *structure* (measuring baseline, graphing data, writing goals, and drawing goal lines), *measurement* (test administration, scoring, and frequency of measurement), and *evaluation* (describing instructional strategies and the changes to them both in terms of content and timing) with each item rated on a 5-point scale. Their intervention included an initial training and then either university-based consultation or self-monitoring. Though significant gains in math achievement were made, no differences were found between the two treatment conditions. Again, in a related and later study, Allinder et al. [199] studied self-monitoring with CBM versus CBM alone. In their focus on the effects of instructional modifications, they found that teachers with the self-monitoring added to CBM made more significant modification and had significantly greater improvement in their students' performance (pre-post) in math than teachers with CBM alone or no CBM.

Within 20 years, the research on data-based decision making was sufficient to arrive at some generalizations (without and with achievement results) [200]. In particular, five features appeared critical for the use of CBM for elementary or middle school students with mild to moderate disabilities.

(1) Progress monitoring alone is insufficient; rather, instruction should be tailored to student needs.

(2) Both types of changes are needed: (a) raising goals when progress is higher than expected and (b) changing instruction when progress is less than expected.

(3) Decision making is facilitated (made more efficient and with greater satisfaction) with computerized data collection, storage, management, and analysis.

(4) Student strengths and weaknesses can be identified with skills analysis in conjunction with consultation.

(5) Meaningful programmatic changes can be facilitated with consultation.

This research on teacher skills in the use of CBM also is based on the use of various decision rules for identifying risk. For example, four dual discrepancy (DD) cutoffs were compared on DIBELS (winter and spring) in which student growth (in first grade) was set below the 25th percentile, at the 33rd percentile, at the 50th percentile, and less than one standard deviation below the mean [201]. On an end of year reading measure, the percentile rank measures consistently (and increasingly by rank) reflected reading risk; "both the 25th and 33rd percentile criteria moderately differentiated reading skills of responsive and non-responsive students who were at risk, but the 50th percentile and one standard deviation criteria both led to small effect sizes" (p. 402). It appears that teachers need guidance to ensure that students with the greatest need are identified.

And this aspect is important at the systems level. As Mellard et al. [202] determined, in one of the few studies done on systems use of RTI. They addressed four components as "a framework that includes (a) universal screening, (b) tiered levels of high-quality interventions, (c) progress monitoring, and (d) data-based curricular decisions" (p. 186). In articulating the actual practice of RTI, they concluded that schools were screening students in various ways, using norms or percentage of population as cut points for risk assessment, placing students into instructional tiers (in varying proportions), and monitoring progress in tiers two and three. Finally, they noted that "good recordkeeping systems was a recurring theme" (p. 192). In addition, school personnel heeded the need to make screening and progress monitoring results accessible and to share data from year to year.

Ironically, none of this research on data-based decision making included training on single subject methodology. In fact, during the first 20 years of CBM, only one study investigated the effects of decision making using a single

subject design [203]. These researchers reported that, when teachers are trained in a single subject design, they can obtain significant results that transfer well beyond the content of instruction. This same argument appears in consideration of the curriculum in CBM. In terms of content-related evidence and the relation between instructional planning, the curriculum has not been found to be a critical feature that identifies CBM as an effective tool in monitoring progress [204]. That is, the use of CBM does not appear to be dictated by the relation of measurement content with instruction. The studies Fuchs and Deno reference were from some of the early research on curriculum differences conducted at the IRLD in the 1980s. They also reference within curriculum differences in sampling passages being as great as between curriculum differences and the lack of (or potential for) generalization. Rather than limiting sampling plans to the curriculum, they suggest three other criteria as instrumental for instructional utility: (a) comparable difficulty, (b) valid outcome indicators, and (c) qualitative feedback. Therefore, in developing teachers skills for using CBM, these important contexts should be considered so that teachers know how to generalize from measurement results to more broadly considered instructional changes.

8.1.2. What We Do Not Know. Presently, few researchers have directly investigated *training systems*. For example, in a training manual for reading decision making, such topics include benchmark measures, expected growth rates, goal establishment, decision rules, instructional strategies, effective reading instruction, and exemplar case studies [205]. Yet, no information exists on the effectiveness of this training. And, as noted by Barnett et al. [192, p. 20], implementation of RTI requires practitioners to consider prevention, screening, and successive instructional programs that are scientifically based and changed using decision rules. It is this combination of multiple variables that needs to be investigated concurrently and at the systems level.

In summary, more research and development is needed for training teachers on systematic use of data, consideration of goals, skills analysis, and data management systems; it also should include use of single subject designs in practice. Yet, most CBM training systems have little data on the effectiveness of the training, even though they are premised upon the collection of student performance and progress. Another problem is that data are collected only at the individual student level and not on teachers. Therefore, professional development cannot be tailored to the areas in which teachers need the most assistance, whether it is about how to effectively progress monitor students, how to develop effective instructional programs, or how to make decisions for maintaining or changing these programs.

Next generation training needs to allow information to be used from teachers and students through relational databases. Teachers need systematic models of assessment practices with classroom vignettes, exemplary practices, and resources that can be immediately used to develop effective progress monitoring. Teachers also need to determine how many students are being monitored on specific measures, grade levels, and time intervals, how students are being organized into tiers, time, and groups, as well as specific instructional emphases being used (along with curriculum materials and strategies), and how well decisions have been made with subsequent changes in level, variability, or slope. To get teachers with these resources and information, training needs to be based on student reports that can be accessed to evaluate not only the effectiveness of interventions but also the systematic use of data-based decision making. Training needs to focus not only on how to implement best practice but also how to interpret information on student performance and progress. Most importantly, professional development is needed in how to use the information in a systematic manner.

8.2. A Nomological Net Supporting a Theory of Change. To support this "next generation" of training on data use, future research first needs to be based on a nomological net that references three warrants.

Assumption 1 (measurement sufficiency). Students are appropriately placed in long-range goal material to ensure the measures are sensitive to change. What is the type of measure, grade level of measure, and the time interval (density of measures) used during the year?

Assumption 2 (instructional adequacy). Instruction is detailed and explicit, allowing a team of teachers to coordinate various elements such as providing an instructional tier (1–3), allocating sufficient time to teach, grouping students appropriately, deciding on the instructional emphasis using specific curriculum (core and supplemental) materials, and determining what instructional strategies to use.

Assumption 3 (decision making). Interventions need to be introduced for low performing students when data warrant change. Are interventions provided at the right time and in accordance with specific data features (e.g., level, variability, and slope)?

The theory of change can be viewed as the interlocking union of the three components in a chain: measurement sufficiency, instructional adequacy, and decision making. See Figure 1. It is not each link itself that is critical, but it is the intersection of the link with subsequent links that is critical (the time series is important in making causal arguments in which time proceeds from left to right). As teachers collect data (from benchmark to decisions of risk and monitoring of progress), the data used to inform them needs to be sufficient, directed toward instruction, and adjusted as needed. Furthermore, this information needs to be collected into a single database for teachers to monitor their application of RTI as well as policy makers and administrators to use the information in making system decisions. The theory is driven by accessibility as a key ingredient to change. If information is not easily accessible and tractable, then it is unlikely to result in use. This theoretical approach also needs to be holistic. Changing individual components as separate events is unlikely to change systems. Rather, the whole needs to be reflected in the parts that in turn need to connect teachers and administrators.

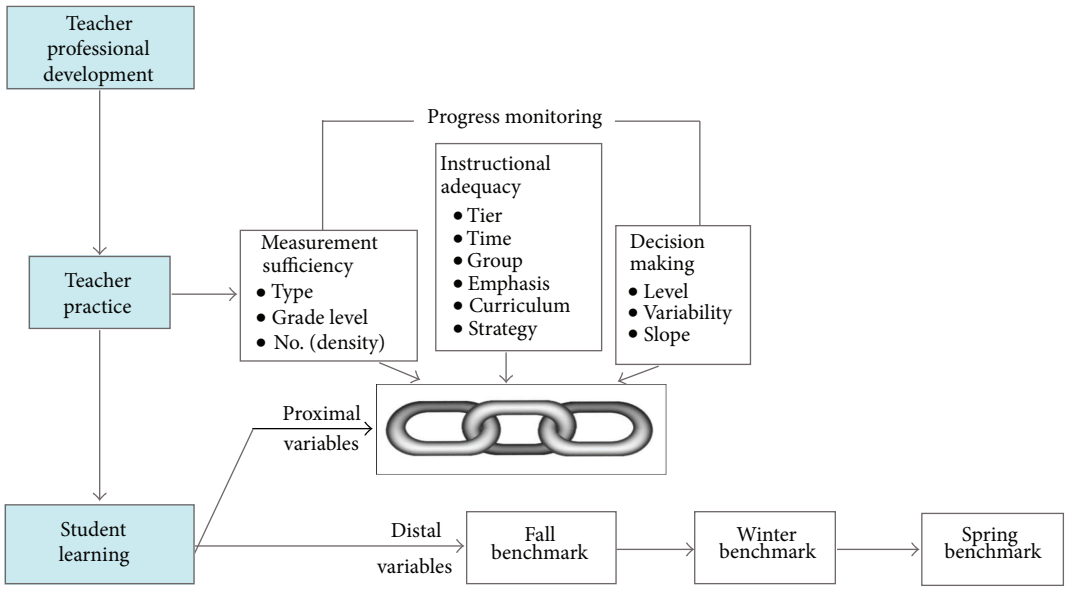

Figure 1: Teacher professional development → Teacher practice → Student learning.

The combined effects from all three components (proximal variables) are critical as well as the relation between them and the outcome (distal variables), which is within-year growth (on benchmark measures) to document change relative to peers. It is not enough to have only one of the proximal variables—the right measures, targeted interventions based on best evidence, and decisions tied to their affect on students. All three are needed. However, they need to work synergistically. And even then, changing these three components is not enough either. Rather, the effect needs to close the achievement gap in which students with disabilities are catching up to their peers on grade level performance measures (e.g., benchmarks). Finally, for systemic change data need to be collected on proximal variables for developing reports on use, allowing for professional development to be tailored and specific.

9. Empirical Research on Three Assumptions (Warrants)

With these assumptions to guide training, and given what we know and do not know about teacher practice, the research agenda for the next 30 years is organized around three critical warrants supporting claims and guiding collection of evidence. In this last section, these three warrants are explicated and used to critique the empirical literature that has been collected in the previous 30 years. However, when the appropriate literature is absent, the analysis is based on easy CBM data extracted from the 2010-2011 school year, with approximately 10,000 students having progress monitoring in kindergarten through grade five. A team of researchers from Behavioral Research and Teaching (BRT) has presented the initial results from these analyses at two national conferences, respectively: *Pacific Coast Research Conference* and *National Council on Measurement in Education* with [206] or Tindal and Nese [68]. Data continue to be harvested on annual benchmarking and progress monitoring data with state test scores in the Oregon Assessment of Knowledge and Skills (OAKS) to document the effects for students with varying characteristics. Most of the research examples in this section are based in reading, but are likely to apply equally well in mathematics.

9.1. Measurement Sufficiency. The focus of this component is to provide teachers and administrators training on student measurement information to improve measurement sufficiency (i.e., increase or decrease the frequency of measurement, change the grade level of measurement, or modify the type or domain of measure being used). These three measurement components would allow policy to reflect an empirical basis for practice.

Many suggestions have been published in the literature on how often a student should be progress monitored, but these reflect few actual studies of different types or schedules of measurement. We know little about the full range of reading measures, as most of the research in reading has focused on oral reading fluency. Yet, teachers need to consider the full range of skills for students, including very early reading (phoneme segmentation and alphabetic principles), as well as vocabulary and comprehension. As Gersten and Dimino [207] note, "equally important is research on RTI that addresses students with problems with vocabulary, listening comprehension, and reading comprehension" (p. 105).

9.1.1. What (Domain) to Measure? Very little research has been completed on the sensitivity of various measures for monitoring reading development. For example, should first grade students with low reading skills be monitored in letter names, letter sounds, phoneme segmentation, or passage reading fluency? At what point in skill development should measures be changed (e.g., from monitoring letter sounds to word reading fluency)? Hintze's work may be the most revealing, but it is limited primarily to oral reading fluency

[82, 208, 209]. In our initial analysis of progress monitoring, we have found considerable variation in the type of reading measure administered over time; see Tindal et al. [206]. In grades three to five, teachers sometimes measured passage reading fluency one week and then either word reading fluency or reading comprehension a few weeks later, followed by word reading fluency. We found little pattern in this cycle of different reading measures being successively administered over time.

9.1.2. What Grade Level to Measure? Recommendations on the grade or domain level to measure are much less prominent than how often to measure. Early research documented differential sensitivity as a function of breadth of the reading domain [47, 210]. Otherwise, very few studies have addressed grade level of the measures but rather consider passage variation [88] or difficulty [82]; again, most of this research is on oral reading fluency. The general finding is that passages differ considerably unless explicitly controlled [83]. Again, our initial analyses of progress monitoring indicate considerable variation in the grade level of the measures used to monitor progress (from 1st to 8th grade). In 3rd to 5th grade, teachers generally (most frequently) monitored progress with grade-level passages or passages that were one grade level below.

9.1.3. How Often to Measure Progress? Frequency of measurement is a function of likelihood of change over time, and for that, the best literature to reference is research on within-year growth for oral reading fluency. As noted earlier, Fuchs et al. [69] documented the average slope of students in grades one to six using a least-squares regression between scores and calendar days with slope converted to a weekly time frame. For the 103 special and general education students, respectively, the weekly growth in words correct per minute was 1.5 and 2.0 in first and second grades, 1.0 and 1.5 in third grade, 0.85 and 1.1 in fourth grade, 0.5 and 0.8 in fifth grade, and 0.3 and 0.63 in sixth grade. Eight years later, Deno et al. [70] conducted a far more reaching study in terms of geographic sampling plan and sample size with nearly 3,000 students tested from four regions of the country. For 2,675 students in general education and 324 in special education classes, they reported 2.0 words correct per minute growth per week until students achieved 30 WRCM; thereafter, students in the general education population improved at least 1.0 word correct per minute/week. In the past decade years, seven studies have been published on slope of progress in oral reading fluency (Deno et al. [70]). In summary, a number of variables have been considered in studying growth of oral reading fluency over the past 15 years. Probably the most remarkable finding is the general consistency in outcomes with very inconsistent sampling plans and research methods.

Note that most of these studies have focused on oral reading fluency with few other reading measures being investigated. The field also needs to address other reading skills (in addition to fluency) that are critical in learning to read for young students in the early elementary grades (e.g., letter naming, letter sounding, and phoneme segmenting) as well as later elementary grades (vocabulary and comprehension). Probably the most significant limitation of prior research in this area is that *benchmark measures*, not *progress measures*, are often used to document growth [68]. The general reason for this is the completeness of the data set for large groups of students and the regularity of data collection (all students in the district take measures in the fall, winter, and spring). Two research exceptions reflecting different schedules report generally similar results to the earlier research on growth: Jenkins and Terjeson [76] and Jenkins et al. [75].

We also know, however, that growth is increasingly being documented as nonlinear, with more growth from fall to winter than from winter to spring [73, 77]. These findings have several implications for policy and practice. If indeed growth is nonlinear, teachers should not use a single goal as part of an Individualized Education Plan (IEP) to create an aim line. This practice would result in teachers underpredicting early progress and overpredicting later progress. From our own (easyCBM) data sets on progress monitoring, we have found that teachers are quite haphazard in the way they monitor students over time. For example, the average length of time between September 1 and the first passage reading fluency (PRF) progress measure in 3rd through 5th grades is more than 10 weeks with a standard deviation greater than seven weeks and students predominantly measured every two to five weeks thereafter. Finally, much of this research fails to include full demographic information for students, particularly those being progress monitored (e.g., students at risk of failing to learn to read) [68].

9.2. Instructional Adequacy. Two primary sources are useful as sources for training on instructional systems: (a) *What Works Clearinghouse* [211–213] and (b) the recently published edition of the *Handbook for Reading Interventions* [214] as well as Wanzek and Cavanaugh [191]. These sources provide a wealth of information that can be used by teachers to organize their instruction, particularly the former two documents. We also use the results from a recent publication on grouping by Schwartz et al. [215] as well as writing from the *National Center to Improve the Tools of Educators* (NCITE).

9.2.1. Instructional Tiers (1–3), Time, and Grouping. Most RTI systems use a 3-tier system [211]. In these three tiers, the following levels of evidence are associated with each tier: (a) tier 1 (differentiated instruction for all students using current reading performance) has *low* levels of evidence, (b) tier two (intensive, systematic instruction on multiple reading skills that is delivered in small groups three to five times per week for 20–40 minutes for students scoring below benchmarks on universal screening) receives *strong* levels of evidence, and (c) tier three (daily intensive instruction on a number of different reading skills for students exhibiting minimum progress in tier two after a reasonable time) receives *low* levels of evidence. An important caveat for tiers two and three is the recommendation for small group instruction to be homogenously configured though the use of a certified teacher or paraprofessional is not deemed critical. Rather, it is the focus on how time is spent and apportioned that is critical, not the total amount of time.

In the easyCBM research, considerable variation exists across districts in their allocations of time in tiers 2 and 3 (from 30 minutes to 60 and 120 minutes extra) though most (51%) of the students receive instruction 4 days per week and another 25% receive instruction five days per week for their first intervention. Fully 22% receive this instruction for fewer than 30 minutes per day, 39% received it for 30–59 minutes per day, another 14% for an hour each day, and only 4% more than an hour every day [216].

The most clear research on grouping comes from Schwartz et al. [215] who used an experimental design to document the effects of student-teacher ratios in teasing out the very significant effects found with 1 : 1 (versus small group instruction) along with professional development (primarily driven by the Reading Recovery model). With teachers randomly assigned to small groups of two, three, and five students in grade two, teachers taught for 20 weeks in daily 30-minute lessons in each condition. Using several pre-post measures, 1 : 1 was compared to all other groups *combined* and found to be significantly different (students scored higher on the post-test after covarying on the pre-test); pairwise comparisons of 1 : 1 with each group were made, and no significant differences were found among the small group comparisons (1 : 2, 1 : 3, and 1 : 5). Finally, post hoc comparisons showed the 1 : 1 condition scoring significantly higher on most of the measures. The problem, then, is simply the resources needed to implement 1 : 1 programs, which is an expensive venture that may be best reserved for students in tier 3. Otherwise, the meta-analysis by Elbaum et al. [217] can be referenced to justify small group instruction, given that 1 : 1 instruction is very resource intensive.

9.2.2. Instructional Emphasis. In the *What Works Clearinghouse* [211], the recommendation is made to "use a curriculum that addresses the components of reading instruction (phonemic awareness, phonics, vocabulary, comprehension, and fluency) and relates to students' needs and development level" (p. 20). In the easyCBM 2001–2011 analysis by Saez [216], four areas were reportedly emphasized as the first intervention (in decreasing use) targeting: fluency (38%), word identification (32%), comprehension (21%), and vocabulary (8%). In addition, we add phonemic awareness and alphabetic principles noted by O' Connor [218].

9.2.3. Curriculum Materials and Instructional Strategies. A myriad of materials and strategies are available for reading instruction. Suffice it to say that "reading instruction should also be explicit. Explicit instruction involves a high level of teacher student interaction that includes frequent opportunities for students to practice the skill with specific corrective feedback" [211, p. 22]. According to the analysis by Saez [216], approximately 66% were *intensive interventions*: CARS/STARS (Comprehensive Assessment of Reading, Strategies/Strategies to Achieve Reading Success), Corrective Reading, Explode the Code, Phonics for Reading, Harcourt Intervention Station, Horizons, Voyager Passport, Open Court Intervention Kit, Triumphs, Read Naturally, Reading Mastery, Rewards, Score4Reading, or SLANT (Structured Language Training). About 31% were *strategic interventions*: Soar to Success, Study Island, Step Up to Writing, various trade consumable workbooks, or various leveled texts. Less than 3% were from the *core curriculum*: Houghton Mifflin or Treasures.

"These instructional factors are encouraging high student engagement in learning, having a strong academic focus and clarity of content coverage, facilitating moderate-to-high success rates, and performance monitoring and informative feedback" [219, p. 582]. In this study, a specific mathematics program (Accelerated Math) was implemented with CBM. The results of the analyses for both the NALT and STAR Math exams indicated that students who participated in AM demonstrated more growth than students who did not participate (p. 532).

In general, and consistent with the *What Works Clearinghouse*, Wanzek and Cavanaugh reported that tier two involved small group instruction and "more frequent programs monitoring (weekly or biweekly) to ensure the effectiveness of instruction for the students" (p. 193). They found group sizes of three to four students, with increasing amounts of time over the grades (though in smaller bites in the lower grades), increasing intensity by offering additional time or smaller groups with instruction more explicit, practice more repeated, feedback more frequent, and more emphasis on high priority skills (also see Saez [216]).

9.3. Decision Making. CBM has generally been viewed as ideographic not nomothetic with students serving as their own control. In this view, the critical currency is how students change over time rather than how they compare to each other. The purpose is to make an individual difference in measurement and evaluation not to document individual differences [220].

Single subject research designs are used to present data on a line graph and analyzed to make data-based decisions [221]. Typically, such designs include a baseline phase prior to implementation of intervention; this phase is then compared to postintervention trajectory [222]. Visual analysis is used to evaluate data of individuals or small groups [221], and decisions rules are then used to signal when instructional changes are needed (not just who should be labeled with a disability). With appropriate labels and particularly timely well-depicted interventions, "cause-effect" relations can be inferred. Using one of three references, teachers can determine who is benefitting and who is not benefitting from the instructional program using (a) the normative distribution of student performance based on all those in the group or a subset of students similar to the targeted student or group, (b) a standard that corresponds with a successful outcome, or (c) individual reference using a comparison to an individual student's prior data [223]. Once sufficient data are collected (e.g., four to five occasions), four indices are summarized to evaluate instructional phases, the first two of which are within phases, and the last two are between phases: (a) slope, (b) variability, (c) level, and (d) overlap (which incorporates all three of these indices).

Decisions can also be based on goal lines or aim lines. Decision rules for intervention change, including service eligibility, are based upon these goal lines in concert with

an empirically sound instructional and decision making sequence [192, 224]. The aim line or goal line refers to the projected amount of weekly growth across time that teachers establish as a minimum for adequate progress [76]. Aim lines are established in research using such techniques as norm references or ordinary least squares (OLS) regression. Deno et al. [70] suggested normative growth rates; others considered slope of improvement relative to aim lines by fitting an OLS regression line to each student's scores, computed the weekly increase based on the obtained parameter, and averaged these weekly increases (across students within grade levels and within general or special education programs). Ardoin [225] developed aim lines by using each student's median baseline score as the aim line start point and calculated the end point of the aim line by multiplying the number of weeks of intervention implementation by the desired gain per week and adding the product to the start point (using an OLS solution). The beginning point of the goal line is placed at the beginning of intervention, the end point is placed at the predicted end of data collection, and a line is drawn to connect the two points. VanDerHeyden et al. [226] developed a local norm in a manner similar to the technique used by Ardoin [225] using students in the "instructional range" as the comparison group. They then used level of performance and trend over time to estimate students' growth by subtracting each student's fall score from his/her spring score and then dividing this difference by the number of intervention weeks. The average of the students' weekly growth estimates was then considered to represent the growth rate of students considered not to be at risk (i.e., those students who scored in the instructional range during the spring administration).

Presently, most decision rules are conceptually not empirically based. Furthermore, a number of methodological problems are inherent in their use: (a) as noted earlier, the assumption of linear growth, which may not be accurate [73, 77], (b) use of gain scores [227], and (c) difficulties in combining statistical and graphic representations of data [192]. Although aim lines may improve decisions, they also have their limitations: disagreement in estimates, distraction from other data qualities, ambiguous results that can lead to misinterpretations [192], and confounding growth with level of baseline performance [194, 224].

In the analysis by Saez [216], three districts were compared in their response to nonresponders with the following options being promulgated: (a) four to six points below aim line or slope is flat/decreasing, (b) measured achievement falls below aim line or flat progress trend, or (c) after four data points to "make an instructional change or continue to monitor." Unfortunately, the vast majority of teachers were implementing few changes, with 355 students (65%) receiving only one intervention, 138 students (25%) receiving two changes, 25 (4.6%) receiving three, and only 29 (5.3%) receiving four or more instructional interventions. Although some of the second interventions were introduced in September ($n = 15$) or October ($n = 13$), many were implemented in the late fall (48 in November and December) and winter (21 in January and February). And of the changes made, most were targeted toward instructional program curricula (50%) or intensity duration (19%); relatively few targeted changes in tier (6%) or group size (6%).

The recommendation is made to monitor progress at least eight times per year by the *What Works Clearinghouse* [211]. "Although no studies have experimentally tested the impact of progress monitoring on outcomes in reading, we still encourage schools to monitor the progress of these students so that personnel possess information on how a student is doing in general education reading proficiency and improving in specific skills" [211, p. 7]. This information would allow decision making to be based on changes in level (average performance), variability (standard deviation of values), and slope (amount of improvement) as indices that can be used to make decisions to change programs. Generally, slope is used most frequently, though it is interpreted relative to the amount of variability.

10. Summary and Parting Comments

Curriculum-based measurement research has a solid 30-year track record that supports its technical adequacy. It represents a standardized form of assessment so that comparability of measures (and outcomes) can be ensured, whether it is comparing a student's performance to other students (norm referenced) or comparing their previous performance to later performance (individual referenced). The former provides effective screening for risk and resource allocation, while the latter provides a system for progress monitoring and evaluation of instruction.

For any measurement system, the first order of business is about technical adequacy, which has been the essential focus throughout the 30 years. The key areas for this research began in reading (fluency), but it was only after major political events that it expanded to early literacy or the foundation of disparate reading skills. Likewise, the initial research on writing was limited, primarily in the populations being monitored (elementary age students but later becoming secondary students), which gave rise to systems for scoring student work (products). Attention to mathematics also was not part of the original research but quickly gained stride across the age ranges of students; given the unique nature of mathematics, the issue mostly addressed domains for sampling items. Finally, in middle and high schools, the focus was on the target skills of content trying to "essentialize" the access skills needed to get to the content. Obviously, more research is needed in all of these subject areas.

However, and perhaps more importantly, further research and development needs to occur in the training of teachers on the use of data and on the reporting systems that teachers are using. Although early research targeted decision making, it was under well-controlled studies being conducted with university support. Little attention was given to full-scale adoption and systematic incorporation into the decision making of the classrooms. Essentially, the research had little verisimilitude. Now, however, it is important to begin this research in typical classrooms that are part of not apart from teachers and students participating in a response-to-intervention system. In order to begin this research, it is best to consider training and reporting. The training can address

the use of norm- or individual-referenced data as well as on smart reports that are electronic and provide prompts for decision making, much like many software applications being developed around scheduling flights, meetings, and weight loss. Given the ubiquitous nature of technology in the classroom, there is no better time for a new generation of curriculum-based measurement embedded into a digital environment.

Acknowledgments

This paper would not have been possible without others at Behavioral Research and Teaching (BRT) Denise Swanson, helped organize the many references cited in this compilation. A Steffani Mast, provided exceptional editing in moving the paper to a final draft. Finally, Dr. Joseph Nese helped craft the logic of the decision making argument at the end of the paper as part of a grant application to fund this line of work.

References

[1] S. Messick, "Standards of validity and the validity of standards in performance assessment," *Educational Measurement*, vol. 14, no. 4, pp. 5–8, 1995.

[2] M. Kane, "Validation," in *Educational Measurement*, R. Brennan, Ed., pp. 17–64, American Council on Education and Praeger, Westport, Conn, USA, 4th edition, 2006.

[3] L. S. Fuchs, "The past, present, and future of curriculum-based measurement research," *School Psychology Review*, vol. 33, no. 2, pp. 188–192, 2004.

[4] American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, *Standards for Educational and Psychological Testing*, Amer Psychological Association, Washington, DC, USA, 1999.

[5] S. Deno and P. Mirkin, *Data Based Program Modification: A Manual*, Leadership Training Institute for Special Education, Minneapolis, Minn, USA, 1977.

[6] G. Tindal and J. F. T. Nese, "Applications of curriculum-based measures in making multiple decisions with multiple reference points," in *Assessment and Intervention: Advances in Learning and Behavioral Disabilities*, M. M. T. Scruggs, Ed., vol. 24, pp. 31–58, Emerald, Bingley, UK, 2011.

[7] P. Mirkin, S. Deno, G. Tindal, and K. Kuehnle, "Formative evaluation: continued development of data utilization systems," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1980.

[8] B. Meyers, J. Meyers, and S. Deno, "Formative evaluation and teacher decision-making: a follow-up investigation," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1980.

[9] S. Deno and P. Mirkin, *Data-Based IEP Development: An Approach to Substantive Compliance*, Monograph, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1979.

[10] G. Tindal, L. Fuchs, S. Christenson, P. Mirkin, and S. Deno, "The relationship between student achievement and teacher assessment of short or long-term goals," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[11] S. Deno, P. Mirkin, and M. Shinn, *Behavioral Perspectives on the Assessment of Learning Disabled Children*, Monograph, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1979.

[12] J. R. Jenkins, S. Deno, and P. Mirkin, *Measuring Pupil Progress Toward the Least Restrictive Environment*, Monograph, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1979.

[13] L. Fuchs and S. Deno, "The relationship between curriculum-based mastery measures and standardized achievement tests in reading," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[14] L. Fuchs, G. Tindal, D. Fuchs, M. Shinn, S. Deno, and G. Germann, "The technical adequacy of a basal reading mastery test: the Holt reading series," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[15] L. Fuchs, G. Tindal, M. Shinn, D. Fuchs, and G. Germann, "Technical adequacy of basal readers' mastery tests: the Ginn 720 series," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[16] G. Tindal, L. Fuchs, D. Fuchs, M. Shinn, S. Deno, and G. Germann, "The technical adequacy of a basal series mastery test: the Scott-Foresman reading program," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[17] G. Tindal, M. Shinn, L. Fuchs, D. Fuchs, S. Deno, and G. Germann, "The technical adequacy of a base reading series mastery test," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[18] D. Fuchs and S. Deno, "Reliability and validity of curriculum-based informal reading inventories," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[19] L. Fuchs, D. Fuchs, and S. Deno, "The nature of inaccuracy among readability formulas," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[20] L. Fuchs and S. Deno, "A comparison of reading placement based on teacher judgment, standardized testing, and curriculum-based assessment," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[21] L. Fuchs, S. Deno, and P. Mirkin, "Direct and frequent measurement and evaluation: effects on instruction and estimates of student progress," Research Report, Minneapolis, Minn, USA, 1982.

[22] L. Fuchs, S. Deno, and P. Mirkin, "Effects of frequent curriculum-based measurement and evaluation on student achievement and knowledge of performance: an experimental study," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[23] L. Fuchs, S. Deno, and A. Roettger, "The effect of alternative data-utilization rules on spelling achievement," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[24] L. Fuchs, C. Wesson, G. Tindal, P. Mirkin, and S. Deno, "Instructional changes, student performances, and teacher preferences:

the effects of specific measurement and evaluation procedures," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[25] P. Mirkin and S. Deno, "Formative evaluation in the classroom: an approach to improving instruction," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1979.

[26] P. Mirkin, L. Fuchs, G. Tindal, S. Christenson, and S. Deno, "The effect of IEP monitoring strategies on teacher behavior," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[27] B. Sevcik, R. Skiba, G. Tindal et al., "Curriculum-based measurement: effects on instruction, teacher estimates of student progress and student knowledge of performance," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[28] C. Wesson, S. Deno, P. Mirkin et al., "Teaching structure and student achievement effects of curriculum-based measurement: a casual (structural) analysis," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[29] C. Wesson, P. Mirkin, and S. Deno, "Teachers' use of self-instructional materials for learning procedures for developing and monitoring progress on IEP goals," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[30] C. Wesson, R. Skiba, B. Sevcik et al., "The impact of the structure of instruction and the use of technically adequate instructional data on reading improvement," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[31] M. Shinn, R. Good, and S. Stein, "Summarizing trend in student achievement: a comparison of methods," *School Psychology Review*, vol. 18, no. 3, pp. 356–370, 1989.

[32] R. Skiba and S. Deno, "A correlational analysis of the statistical properties of time-series data and their relationship to student achievement in resource classrooms," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[33] R. Skiba, D. Marston, C. Wesson, B. Sevcik, and S. Deno, "Characteristics of the time-series data collected through curriculum-based reading measurement," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[34] G. Tindal, S. Deno, and J. Ysseldyke, "Visual analysis of time series data: factors of influence and level reliability," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[35] D. Marston, L. Lowry, S. Deno, and P. Mirkin, "An analysis of learning trends in simple measures of reading, spelling, and written expression: a longitudinal study," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[36] R. King, S. Deno, P. Mirkin, and C. Wesson, "The effects of training in the use of formative evaluation in reading: an experimental control comparison," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[37] R. Skiba, C. Wesson, and S. Deno, "The effects of training teachers in the use of formative evaluation in reading: an experimental control comparison," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[38] L. Fuchs, C. Wesson, G. Tindal, P. Mirkin, and S. Deno, "Teacher efficiency in continuous evaluation of IEP goals," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[39] R. King, C. Wesson, and S. Deno, "Direct and frequent measurement of student performance: does it take too much time?" Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[40] D. Marston and S. Deno, "Implementation of direct and repeated measurement in the school setting," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[41] D. Marston, G. Tindal, and S. Deno, "Predictive efficiency of direct, repeated measurement: an analysis of cost and accuracy in classification," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[42] P. Mirkin, D. Marston, and S. Deno, "Direct and repeated measurement of academic skills: an alternative to traditional screening referral, and identification of learning disabled students," Research Reports, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[43] G. Tindal, G. Germann, and S. Deno, "Descriptive research on the Pine County norms: a compilation of findings," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[44] G. Tindal, G. Germann, D. Marston, and S. Deno, "The effectiveness of special education," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[45] L. Fuchs, G. Tindal, and S. Deno, "Effects of varying item domain and sample duration on technical characteristics of daily measures in reading," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[46] M. Shinn, M. Gleason, and G. Tindal, "Varying the difficulty of testing materials: implications for curriculum-based measurement," *The Journal of Special Education*, vol. 23, pp. 223–233, 1989.

[47] G. Tindal and S. Deno, "Daily measurement of reading: effects of varying the size of the item pool," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[48] G. Tindal, D. Marston, S. Deno, and G. Germann, "Curriculum differences in direct repeated measures of reading," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[49] S. Deno, D. Marston, P. Mirkin, L. Lowry, P. Sindelar, and J. R. Jenkins, *The Use of Standard Tasks to Measure Achievement in Reading, Spelling, and Written Expression: A Normative and Developmental Study*, Institute for Research on Learning Disabilities, Minneapolis, Minn, USA, 1982.

[50] S. Deno, P. Mirkin, B. Chiang, and L. Lowry, "Relationships among simple measures of reading and performance on standardized achievement tests," Research Report, Institute for

Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1980.

[51] S. Deno, P. Mirkin, L. Lowry, and K. Kuehnle, "Relationships among simple measures of spelling and performance on standardized achievement tests," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1980.

[52] S. Deno, P. Mirkin, and D. Marston, "Relationships among simple measures of written expression and performance on standardized achievement tests," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1980.

[53] L. Fuchs, S. Deno, and D. Marston, "Use of aggregation to improve the reliability of simple direct measures of academic performance," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[54] D. Marston and S. Deno, "The reliability of simple, direct measures of written expression," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1981.

[55] G. Tindal, D. Marston, and S. Deno, "The reliability of direct and repeated measurement," Research Report, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1983.

[56] J. Videen, S. Deno, and D. Marston, "Correct word sequences: a valid indicator of proficiency in written expression," Research Reports, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[57] C. Wesson, S. Deno, and P. Mirkin, *Research on Developing and Monitoring Progress on IEP Goals: Current Findings and Implications for Practice*, Monograph, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[58] G. Tindal, G. Germann, S. Deno, and P. Mirkin, *The Pine County Model for Special Education Delivery: A Data-Based System*, Monograph, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[59] P. K. Mirkin, L. S. Fuchs, and S. L. Deno, *Consideration for Designing a Continuous Evaluation : An Interpretive Review*, Monograph, Institute for Research on Learning Disabilities (IRLD), University of Minnesota, Minneapolis, Minn, USA, 1982.

[60] M. R. Shinn, *Curriculum-Based Measurement: Assessing Special Children*, The Guilford Press, New York, NY, USA, 1989.

[61] M. R. Shinn, *Advanced Applications of Curriculum-Based Measurement*, Guilford Press, New York, NY, USA, 1998.

[62] G. Tindal and D. Marston, *Classroom-Based Assessment: Evaluating Instructional Outcomes*, Merrill, Columbus, Ohio, USA, 1990.

[63] E. E. Gickling and V. P. Thompson, "A personal view of curriculum-based assessment," *Exceptional Children*, vol. 52, no. 3, pp. 205–218, 1985.

[64] K. Howell, *Curriculum-Based Evaluation for Special and Remedial Education: A Handbook for Deciding What to Teach*, Merrill Publishing, Columbus, Ohio, USA, 1987.

[65] C. S. Blankenship, "Using curriculum-based assessment data to make instructional decisions," *Exceptional Children*, vol. 52, no. 3, pp. 233–238, 1985.

[66] M. R. Shinn, S. Rosenfield, and N. Knutson, "Curriculum-based assessment: a comparison of models," *School Psychology Review*, vol. 18, no. 3, pp. 299–316, 1989.

[67] C. A. Espin, K. L. McMaster, S. Rose, and M. M. Wayman, *A Measure of Success: How Curriculum-Based Measurement has Influenced Education and Learning*, University of Minnesota Press, Minneapolis, Minn, USA, 2012.

[68] G. Tindal and J. F. T. Nese, "Within year achievement growth using curriculum based measurement," in *Proceedings of the National Council on Measurement in Education*, Vancouver, Canada, April 2012.

[69] L. Fuchs, D. Fuchs, C. Hamlett, L. Walz, and G. Germann, "Formative evaluation of academic progress: how much growth can we expect?" *School Psychology Review*, vol. 22, pp. 27–48, 1993.

[70] S. L. Deno, L. S. Fuchs, D. Marston, and J. Shin, "Using curriculum-based measurement to establish growth standards for students with learning disabilities," *School Psychology Review*, vol. 30, no. 4, pp. 507–524, 2001.

[71] S. P. Ardoin and T. J. Christ, "Evaluating curriculum-based measurement slope estimates using data from triannual universal screenings," *School Psychology Review*, vol. 37, no. 1, pp. 109–125, 2008.

[72] T. J. Christ, "Short-term estimates of growth using curriculum-based measurement of oral reading fluency: estimating standard error of the slope to construct confidence intervals," *School Psychology Review*, vol. 35, no. 1, pp. 128–133, 2006.

[73] T. J. Christ, B. Silberglitt, S. Yeo, and D. Cormier, "Curriculum-based measurement of oral reading: an evaluation of growth rates and seasonal effects among students served in general and special education," *School Psychology Review*, vol. 39, no. 3, pp. 447–462, 2010.

[74] S. B. Graney, K. N. Missall, R. S. Martínez, and M. Bergstrom, "A preliminary investigation of within-year growth patterns in reading and mathematics curriculum-based measures," *Journal of School Psychology*, vol. 47, no. 2, pp. 121–142, 2009.

[75] J. R. Jenkins, J. J. Graff, and D. L. Miglioretti, "Estimating reading growth using intermittent CBM progress monitoring," *Exceptional Children*, vol. 75, no. 2, pp. 151–163, 2009.

[76] J. Jenkins and K. Terjeson, "Monitoring reading growth: goal setting, measurement frequency, and methods of evaluation," *Learning Disabilities Research & Practice*, vol. 26, no. 1, pp. 28–35, 2011.

[77] J. F. T. Nese, G. Biancarosa, D. Anderson, C. F. Lai, J. Alonzo, and G. Tindal, "Within-year oral reading fluency with CBM: a comparison of models," *Reading and Writing*, vol. 25, no. 4, pp. 887–915, 2012.

[78] S. Raudenbush, A. Bryk, Y. Cheong, and R. Congdon, *HLM 6: Hierarchical Linear and NonLinear Modeling*, Scientific Software International, Lincolnwood, Ill, USA, 2004.

[79] S. Raudenbush and A. Bryk, *Hierarchical Linear Models: Applications and Data Analysis Methods*, Sage Publications, Thousand Oaks, Calif, USA, 2nd edition, 2002.

[80] R. Brennan, *Generalizability Theory*, Springer, New York, NY, USA, 2001.

[81] R. Linn and E. Burton, "Performance-based assessment: implications of task specificity," *Educational Measurement*, vol. 13, pp. 5–8, 1994.

[82] J. M. Hintze, E. J. Daly, and E. S. Shapiro, "An investigation of the effects of passage difficulty level on outcomes of oral reading fluency progress monitoring," *School Psychology Review*, vol. 27, no. 3, pp. 433–445, 1998.

[83] J. M. Hintze and T. J. Christ, "An examination of variability as a function of passage variance in CBM progress monitoring," *School Psychology Review*, vol. 33, no. 2, pp. 204–217, 2004.

[84] J. M. Hintze, S. V. Owen, E. S. Shapiro, and E. J. Daly, "Research design and methodology section—generalizability of oral reading fluency measures: application of g theory to curriculum-based measurement," *School Psychology Quarterly*, vol. 15, no. 1, pp. 52–68, 2000.

[85] B. C. Poncy, C. H. Skinner, and P. K. Axtell, "An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement," *Journal of Psychoeducational Assessment*, vol. 23, no. 4, pp. 326–338, 2005.

[86] T. J. Christ and S. P. Ardoin, "Curriculum-based measurement of oral reading: passage equivalence and probe-set development," *Journal of School Psychology*, vol. 47, no. 1, pp. 55–75, 2009.

[87] S. P. Ardoin and T. J. Christ, "Curriculum-based measurement of oral reading: standard errors associated with progress monitoring outcomes from DIBELS, AIMSweb, and an experimental passage set," *School Psychology Review*, vol. 38, no. 2, pp. 266–283, 2009.

[88] D. J. Francis, K. L. Santi, C. Barr, J. M. Fletcher, A. Varisco, and B. R. Foorman, "Form effects on the estimation of students' oral reading fluency using DIBELS," *Journal of School Psychology*, vol. 46, no. 3, pp. 315–342, 2008.

[89] T. Christ, "Curriculum-based measurement of oral reading: multi-study evaluation of schedule, duration and dataset quality on progress monitoring outcomes," *Journal of School Psychology*, vol. 78, no. 3, 2012.

[90] D. C. Briggs, "Synthesizing causal inferences," *Educational Researcher*, vol. 37, no. 1, pp. 15–22, 2008.

[91] National Institute for Literacy, *Developing Early Literacy: Report of the National Early Literacy Panel (A Scientific Synthesis of Early Literacy Development and Implications for Intervention)*, Washington, DC, USA, 2008.

[92] R. C. Anderson, E. H. Hiebert, J. A. Scott, and I. A. G. Wilkinson, *Becoming a Nation of Readers: The Report of the Commission on Reading*, National Institute of Education, Washington, DC, USA, 1985.

[93] National Institutes of Child Health and Human Development, "Report of national reading panel: teaching children to read: an evidence-based assessment of the scientific literature on reading and its implications for reading instruction," Report of the Subgroups, Washington, DC, USA, 2000.

[94] No Child Left Behind, *Committee on Education and Labor*, Government Printing Office, Washington, DC, USA, 1st edition, 2001.

[95] M. J. Adams, *Beginning to Read: Thinking and Learning about Print*, MIT Press, Cambridge, Mass, USA, 1990.

[96] L. S. Fuchs, D. Fuchs, and D. L. Compton, "Monitoring early reading development in first grade: word identification fluency versus nonsense word fluency," *Exceptional Children*, vol. 71, no. 1, pp. 7–21, 2004.

[97] K. D. Ritchey and D. L. Speece, "From letter names to word reading: the nascent role of sublexical fluency," *Contemporary Educational Psychology*, vol. 31, no. 3, pp. 301–327, 2006.

[98] K. D. Ritchey, "Assessing letter sound knowledge: a comparison of letter sound fluency and nonsense word fluency," *Exceptional Children*, vol. 74, no. 4, pp. 487–506, 2008.

[99] P. G. Aaron, R. Malatesha Joshi, R. Gooden, and K. E. Bentum, "Diagnosis and treatment of reading disabilities based on the component model of reading: an alternative to the discrepancy model of LD," *Journal of Learning Disabilities*, vol. 41, no. 1, pp. 67–84, 2008.

[100] D. Starch, "The measurement of efficiency in reading," *The Journal of Educational Psychology*, vol. 6, no. 1, pp. 1–24, 1915.

[101] V. L. Anderson and M. A. Tinker, "The speed factor in reading performance," *Journal of Educational Psychology*, vol. 27, no. 8, pp. 621–624, 1936.

[102] D. LaBerge and S. J. Samuels, "Toward a theory of automatic information processing in reading," *Cognitive Psychology*, vol. 6, no. 2, pp. 293–323, 1974.

[103] R. Good, D. Simmons, and E. Kame'enui, "The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes," *Scientific Studies of Reading*, vol. 5, no. 3, pp. 257–288, 2001.

[104] R. H. Good, J. Gruba, and R. A. Kaminski, "Best practices in using Dynamic Indicators of Basic Early Literacy Skills (DIBELS) in an outcomes-driven model," in *Best Practices in School Psychology IV*, A. Thomas and J. Grimes, Eds., pp. 679–700, National Association of School Psychologists, Washington, DC, USA, 2001.

[105] R. H. Good and R. A. Kaminski, "Assessment for instructional decisions: toward a proactive/prevention model of decision-making for early literacy skills," *School Psychology Quarterly*, vol. 11, no. 4, pp. 326–336, 1996.

[106] H. L. Rouse and J. W. Fantuzzo, "Validity of the dynamic indicators for basic early literacy skills as an indicator of early literacy for urban kindergarten children," *School Psychology Review*, vol. 35, no. 3, pp. 341–355, 2006.

[107] N. Clemens, E. Shapiro, and F. Thoemmes, "Improving the efficacy of first grade reading screening: an investigation of word identification fluency with other early literacy indicators," *School Psychology Quarterly*, vol. 26, no. 3, pp. 231–244, 2011.

[108] S. Hagan-Burke, M. Burke, and C. Crowder, "The convergent validity of the dynamic indicators of basic and early literacy skills and the test of word reading efficiency for the beginning of first grade," *Assessment for Effective Intervention*, vol. 31, no. 4, pp. 1–15, 2006.

[109] Y. Y. Lo, C. Wang, and S. Haskell, "Examining the impacts of early reading intervention on the growth rates in basic literacy skills of at-risk urban kindergarteners," *The Journal of Special Education*, vol. 43, no. 1, pp. 12–28, 2009.

[110] U. Yesil-Dagli Ummuhan, "Predicting ELL students' beginning first grade English oral reading fluency from initial kindergarten vocabulary, letter naming, and phonological awareness skills," *Early Childhood Research Quarterly*, vol. 26, no. 1, pp. 15–29, 2011.

[111] L. C. Ehri, S. R. Nunes, D. M. Willows, B. V. Schuster, Z. Yaghoub-Zadeh, and T. Shanahan, "Phonemic awareness instruction helps children learn to read: evidence from the National Reading Panels meta-analysis," *Reading Research Quarterly*, vol. 36, no. 3, pp. 250–287, 2001.

[112] S. Stage, J. Sheppard, M. Davidson, and M. Browning, "Prediction of first-graders' growth in oral reading fluency using kindergarten letter fluency," *Journal of School Psychology*, vol. 39, pp. 225–237, 2001.

[113] H. Yopp, "A test for assessing phonemic awareness in young children," *The Reading Teacher*, vol. 49, no. 1, pp. 20–29, 1995.

[114] D. L. Linklater, R. E. O'Connor, and G. J. Palardy, "Kindergarten literacy assessment of English Only and English language learner students: an examination of the predictive validity of three phonemic awareness measures," *Journal of School Psychology*, vol. 47, no. 6, pp. 369–394, 2009.

[115] D. L. Speece, K. D. Ritchey, D. H. Cooper, F. P. Roth, and C. Schatschneider, "Growth in early reading skills from kindergarten to third grade," *Contemporary Educational Psychology*, vol. 29, no. 3, pp. 312–332, 2004.

[116] C. Juel, "Learning to read and write: a longitudinal study of 54 children from first through fourth grades," *Journal of Educational Psychology*, vol. 80, no. 4, pp. 437–447, 1988.

[117] J. M. T. Vloedgraven and L. Verhoeven, "Screening of phonological awareness in the early elementary grades: an IRT approach," *Annals of Dyslexia*, vol. 57, no. 1, pp. 33–50, 2007.

[118] M. Twain, "Simplified spelling," in *Letters from the Earth: Uncensored Writings*, B. DeVoto, Ed., HarperCollins, New York, NY, USA, 1942.

[119] L. Fuchs and D. Fuchs, "Determining adequate yearly progress from kindergarten through grade 6 with curriculum-based measurement," *Assessment for Effective Intervention*, vol. 29, no. 4, pp. 25–37, 2004.

[120] L. S. Fuchs, D. Fuchs, and D. L. Compton, "Monitoring early reading development in first grade: word identification fluency versus nonsense word fluency," *Exceptional Children*, vol. 71, no. 1, pp. 7–21, 2004.

[121] R. Parker, G. Tindal, and J. Hasbrouck, "Countable indices of writing quality: their suitability for screening-eligibility decisions," *Exceptionality*, vol. 2, pp. 1–17, 1991.

[122] R. I. Parker, G. Tindal, and J. Hasbrouck, "Progress monitoring with objective measures of writing performance for students with mild disabilities," *Exceptional Children*, vol. 58, no. 1, pp. 61–73, 1991.

[123] G. Tindal and R. Parker, "Assessment of written expression for students in compensatory and special education programs," *The Journal of Special Education*, vol. 23, no. 2, pp. 169–183, 1989.

[124] G. Tindal and R. Parker, "Identifying measures for evaluating written expression," *Learning Disabilities Research and Practice*, vol. 6, pp. 211–218, 1991.

[125] K. A. Gansle, A. M. VanDerHeyden, G. H. Noell, J. L. Resetar, and K. L. Williams, "The technical adequacy of curriculum-based and rating-based measures of written expression for elementary school students," *School Psychology Review*, vol. 35, no. 3, pp. 435–450, 2006.

[126] K. L. McMaster and H. Campbell, "New and existing curriculum-based writing measures: technical features within and across grades," *School Psychology Review*, vol. 37, no. 4, pp. 550–566, 2008.

[127] K. A. Gansle, G. H. Noell, A. M. VanDerHeyden et al., "An examination of the criterion validity and sensitivity to brief intervention of alternate curriculum-based measures of writing skill," *Psychology in the Schools*, vol. 41, no. 3, pp. 291–300, 2004.

[128] C. Espin, B. Scierka, and S. Skare, "Criterion-related validity of curriculum-based measures in writing for secondary school students," *Reading & Writing Quarterly*, vol. 15, no. 1, pp. 5–27, 1999.

[129] C. Espin, J. Shin, S. L. Deno, S. Skare, S. Robinson, and B. Benner, "Identifying indicators of written expression proficiency for middle school students," *The Journal of Special Education*, vol. 34, no. 3, pp. 140–153, 2000.

[130] J. W. Weissenburger and C. A. Espin, "Curriculum-based measures of writing across grade levels," *Journal of School Psychology*, vol. 43, no. 2, pp. 153–169, 2005.

[131] C. Espin, T. Wallace, H. Campbell, E. S. Lembke, J. D. Long, and R. Ticha, "Curriculum-based measurement in writing: predicting the success of high-school students on state standards tests," *Exceptional Children*, vol. 74, no. 2, pp. 174–193, 2008.

[132] B. Diercks-Gransee, J. W. Weissenburger, C. L. Johnson, and P. Christensen, "Curriculum-based measures of writing for high school students," *Remedial and Special Education*, vol. 30, no. 6, pp. 360–371, 2009.

[133] C. A. Espin, S. De La Paz, B. J. Scierka, and L. Roelofs, "The relationship between curriculum-based measures in written expression and quality and completeness of expository writing for middle school students," *The Journal of Special Education*, vol. 38, no. 4, pp. 208–217, 2005.

[134] S. Fewster and P. D. Macmillan, "School-based evidence for the validity of curriculum-based measurement of reading and writing," *Remedial and Special Education*, vol. 23, no. 3, pp. 149–156, 2002.

[135] J. M. Amato and M. W. Watkins, "The predictive validity of CBM writing indices for eighth-grade students," *The Journal of Special Education*, vol. 44, no. 4, pp. 195–204, 2011.

[136] K. McMaster and C. Espin, "Technical features of curriculum-based measurement in writing: a literature review," *The Journal of Special Education*, vol. 41, no. 2, pp. 68–84, 2007.

[137] A. Foegen and S. L. Deno, "Identifying growth indicators for low-achieving students in middle school mathematics," *The Journal of Special Education*, vol. 35, no. 1, pp. 4–16, 2001.

[138] M. Calhoon, "Curriculum-based measurement for mathematics at the high school level: what we do not know and what we need to know," *Assessment for Effective Intervention*, vol. 33, pp. 234–239, 2008.

[139] L. Fuchs, D. Fuch, and S. Courey, "Curriculum-based measurement of mathematics competence: from competence to concepts and applications to real life problem solving," *Assessment for Effective Intervention*, vol. 30, no. 2, pp. 33–46, 2005.

[140] T. Christ, S. Sculin, A. Tolbize, and C. Jiban, "Implication of recent research: curriculum-based measurement of math computation," *Assessment for Effective Intervention*, vol. 33, pp. 198–205, 2008.

[141] A. Foegen, C. Jiban, and S. Deno, "Progress monitoring measures in mathematics. A review of the literature," *The Journal of Special Education*, vol. 41, no. 2, pp. 121–139, 2007.

[142] M. K. Burns, A. M. VanDerHeyden, and C. L. Jiban, "Assessing the instructional level for mathematics: a comparison of methods," *School Psychology Review*, vol. 35, no. 3, pp. 401–418, 2006.

[143] T. J. Christ and O. Vining, "Curriculum-based measurement procedures to develop multiple-skill mathematics computation probes: evaluation of random and stratified stimulus-set arrangements," *School Psychology Review*, vol. 35, no. 3, pp. 387–400, 2006.

[144] L. S. Fuchs, D. Fuchs, D. L. Compton, J. D. Bryant, C. L. Hamlett, and P. M. Seethaler, "Mathematics screening and progress monitoring at first grade: implications for responsiveness to intervention," *Exceptional Children*, vol. 73, no. 3, pp. 311–330, 2007.

[145] C. Jiban and S. Deno, "Using math and reading curriculum-based measurements to predict state mathematics test performance: are simple one-minute measures technically adequate?" *Assessment for Effective Intervention*, vol. 32, pp. 78–89, 2007.

[146] P. Seethaler and L. Fuchs, "Using curriculum-based measurement to monitor kindergarteners' mathematics development," *Assessment for Effective Intervention*, vol. 36, no. 4, pp. 219–229, 2011.

[147] E. Shapiro, L. Edwards, and N. Zigmond, "Progress monitoring of mathematics among students with learning disabilities," *Assessment for Effective Intervention*, vol. 30, no. 2, pp. 15–32, 2005.

[148] B. Clarke, S. Baker, K. Smolkowski, and D. J. Chard, "An analysis of early numeracy curriculum-based measurement: examining the role of growth in student outcomes," *Remedial and Special Education*, vol. 29, no. 1, pp. 46–57, 2008.

[149] D. Chard, B. Clarke, S. Baker, J. Otterstedt, D. Braun, and R. Katz, "Using measures of number sense to screen for difficulties in mathematics: preliminary findings," *Assessment for Effective Intervention*, vol. 30, no. 2, pp. 3–14, 2005.

[150] E. Lembke, A. Foegen, T. Whittaker, and D. Hampton, "Establishing technically adequate measures of progress in early numeracy," *Assessment for Effective Intervention*, vol. 33, no. 4, pp. 206–214, 2008.

[151] R. Martinez, K. Missal, S. Graney, O. Aricak, and B. Clarke, "Technical adequacy of early numeracy curriculum-based measurement in kindergarten," *Assessment for Effective Intervention*, vol. 34, pp. 116–125, 2009.

[152] A. M. VanDerHeyden, C. Broussard, and A. Cooley, "Further development of measures of early math performance for preschoolers," *Journal of School Psychology*, vol. 44, no. 6, pp. 533–553, 2006.

[153] J. Leh, A. Jitendra, G. Caskie, and C. Griffin, "An evaluation of curriculum-based measurement of mathematics word problem solving measures monitoring third-grade students' mathematics competence," *Assessment for Effective Intervention*, vol. 32, pp. 90–99, 2007.

[154] L. Fuchs, D. Fuchs, and D. Compton, "The early prevention of mathematics difficulty: its power and limitations," *Journal of Learning Disabilities*, vol. 45, no. 3, Article ID 269, p. 257, 2012.

[155] A. Foegen, "Technical adequacy of general outcome measures for middle school mathematics," *Assessment for Effective Intervention*, vol. 25, pp. 175–203, 2000.

[156] A. Foegen, "Progress monitoring in middle school mathematics: options and issues," *Remedial and Special Education*, vol. 29, no. 4, pp. 195–207, 2008.

[157] L. S. Fuchs, C. L. Hamlett, and D. Fuchs, *Monitoring Basic Skills Progress: Basic Math Computation*, 2nd edition, 1998.

[158] L. S. Fuchs, C. L. Hamlett, and D. Fuchs, *Monitoring Basic Skills Progress: Basic Math Concepts and Applications*, 1999.

[159] E. Lembke and P. Stecker, *Curriculum-Based Measurement in Mathematics: An Evidence-Based Formative Assessment Procedure*, RMC Research Corporation, Center on Instruction, Portsmouth, UK, 2007.

[160] C. Espin and G. Tindal, "Curriculum-based measurement for secondary students," in *Advanced Applications of Curriculum-Based Measurement*, M. R. Shinn, Ed., Guilford Press, New York, NY, USA, 1998.

[161] R. Quirk, *The Linguist and the English Language*, Arnold, London, UK, 1974.

[162] D. Corson, "The learning and use of academic english words," *Language Learning*, vol. 47, no. 4, pp. 671–718, 1997.

[163] C. Espin and S. Deno, "Content-specific and general reading disabilities of secondary-level students: identification and educational relevance," *The Journal of Special Education*, vol. 27, pp. 321–337, 1993.

[164] C. Espin and S. Deno, "Performance in reading from content area text as an indicator of achievement," *Remedial and Special Education*, vol. 14, no. 6, pp. 47–59, 1993.

[165] C. Espin and S. Deno, "Curriculum-based measures for secondary students: utility and task specificty of text-based reading and vocabulary measures for predicting performance on content-area tasks," *Diagnostique*, vol. 20, no. 1–4, pp. 121–142, 1994.

[166] C. A. Espin and A. Foegen, "Validity of general outcome measures for predicting secondary students performance on content-area tasks," *Exceptional Children*, vol. 62, no. 6, pp. 497–514, 1996.

[167] C. Espin, T. Busch, J. Shin, and R. Kruschwitz, "Curriculum-based measurement in the content areas: validity of vocabulary matching as an indicator of performance in social studies," *Learning Disabilities*, vol. 16, pp. 142–151, 2001.

[168] C. A. Espin, J. Shin, and T. W. Busch, "Curriculum-based measurement in the content areas: vocabulary matching as an indicator of progress in social studies learning," *Journal of Learning Disabilities*, vol. 38, no. 4, pp. 353–363, 2005.

[169] C. Espin, T. Wallace, E. Lembke, H. Campbell, and J. Long, "Creating a progress-monitoring system in reading for middle school students: tracking progress toward meeting high-stakes standards," *Learning Disabilities Research and Practice*, vol. 25, pp. 60–75, 2010.

[170] V. Nolet and G. Tindal, "Special education in content area classes: development of a model and practical procedures," *Remedial and Special Education*, vol. 14, no. 1, pp. 36–48, 1993.

[171] G. Roid and T. M. Haladyna, *A Technology for Test-Item Writing*, Academic Press, Orlando, Fla, USA, 1982.

[172] G. Tindal and V. Nolet, "Curriculum-based measurement in middle and high schools: critical thinking skills in content areas," *Focus on Exceptional Children*, vol. 27, no. 7, pp. 1–22, 1995.

[173] S. Engelmann and D. Carnine, *Theory of Instruction: Principles and Applications*, Irvington Publishers, New York, NY, USA, 1982.

[174] T. Twyman, J. McCleery, and G. Tindal, "Using concepts to frame history content," *Journal of Experimental Education*, vol. 74, no. 4, pp. 331–349, 2006.

[175] T. Twyman and G. Tindal, "Reaching all of your students in social studies," *Teaching Exceptional Children*, vol. 1, no. 5, article 1, 2005.

[176] T. Twyman, L. R. Ketterlin-Geller, J. D. McCoy, and G. Tindal, "Effects of concept-based instruction on an English language learner in a rural school: a descriptive case study," *Bilingual Research Journal*, vol. 27, no. 2, pp. 259–274, 2003.

[177] S. Embretson and S. Reise, *Item Response Theory for Psychologists*, Lawrence Erlbaum Associates, Mahwah, NJ, USA, 2000.

[178] T. Twyman and G. Tindal, "Extending curriculum-based measurement into middle/secondary schools: the technical adequacy of the concept maze," *Journal of Applied School Psychology*, vol. 24, no. 1, pp. 49–67, 2007.

[179] J. E. Heimlich and S. D. Pittelman, *Semantic Mapping: Classroom Applications*, International Reading Association, Newark, Del, USA, 1986.

[180] F. P. Hunkins, *Teaching Thinking Through Effective Questioning*, Christopher-Gordon Publisher, Boston, Mass, USA, 1989.

[181] G. Tindal and V. Nolet, "Serving students in middle school content classes: a heuristic study of critical variables linking instruction and assessment," *The Journal of Special Education*, vol. 29, no. 4, pp. 414–432, 1996.

[182] V. Nolet and G. Tindal, "Curriculum-based collaboration," *Focus on Exceptional Children*, vol. 27, no. 3, pp. 1–12, 1994.

[183] V. Nolet and G. Tindal, "Instruction and learning in middle school science classes: implications for students with disabilities," *The Journal of Special Education*, vol. 28, no. 2, pp. 166–187, 1994.

[184] G. Tindal, V. Nolet, and G. Blake, *Focus on Teaching and Learning in Content Classes*, University of Oregon Behavioral Research and Teaching, Eugene, Ore, USA, 1992.

[185] V. Nolet, G. Tindal, and G. Blake, *Focus on Assessment Learning in Content Classes*, University of Oregon Behavioral Research and Teaching, Eugene, Ore, USA, 1993.

[186] L. Ketterlin-Geller and G. Tindal, *Concept-Based Instruction: Science*, University of Oregon Behavioral Research and Teaching, Eugene, Ore, USA, 2002.

[187] T. Twyman, L. Ketterlin-Geller, and G. Tindal, *Concept-Based Instruction: Social Science*, University of Oregon Behavioral Research and Teaching, Eugene, Ore, USA, 2002.

[188] M. McDonald, L. Ketterlin-Geller, and G. Tindal, *Concept-Based Instruction: Mathematics*, University of Oregon Behavioral Research and Teaching, Eugene, Ore, USA, 2002.

[189] G. Tindal, J. Alonzo, and L. Ketterlin-Geller, *Concept-Based Instruction: Language Arts*, University of Oregon Behavioral Research and Teaching, Eugene, Ore, USA, 2002.

[190] S. Berkeley, W. N. Bender, L. Gregg Peaster, and L. Saunders, "Implementation of response to intervention: a snapshot of progress," *Journal of Learning Disabilities*, vol. 42, no. 1, pp. 85–95, 2009.

[191] J. Wanzek and C. Cavanaugh, "Characteristics of general education reading interventions implemented in elementary schools with reading difficulties," *Remedial and Special Education*, vol. 33, no. 3, pp. 192–202, 2012.

[192] D. Barnett, N. Elliott, J. Graden et al., "Technical adequacy for response to intervention practices," *Assessment for Effective Intervention*, vol. 32, no. 1, pp. 20–31, 2006.

[193] S. Messick, "Validity," in *Educational Measurement*, R. Linn, Ed., pp. 13–103, Macmillan Publishing Company, New York, NY, USA, 3rd edition, 1989.

[194] A. VanDerHeyden, "Technical adequacy of response to intervention decisions," *Council for Exceptional Children*, vol. 77, no. 3, pp. 335–350, 2011.

[195] R. Gersten, T. Keating, and L. K. Irvin, "The burden of proof: validity as improvement of instructional practice," *Exceptional Children*, vol. 61, no. 5, pp. 510–519, 1995.

[196] R. Allinder, "An examination of the relationship between teacher efficacy and curriculum-based measurement and student achievement," *Remedial and Special Education*, vol. 16, pp. 247–254, 1995.

[197] R. M. Allinder, "When some is not better than none: effects of differential implementation of curriculum-based measurement," *Exceptional Children*, vol. 62, no. 6, pp. 525–535, 1996.

[198] R. Allinder and M. BeckBest, "Differential effects of two approaches to supporting teachers' use of curriculum-based measurement," *School Psychology Review*, vol. 24, pp. 287–298, 1995.

[199] R. M. Allinder, R. M. Bolling, R. G. Oats, and W. A. Gagnon, "Effects of teacher self-monitoring on implementation of curriculum-based measurement and mathematics computation achievement of students with disabilities," *Remedial and Special Education*, vol. 21, no. 4, pp. 219–226, 2000.

[200] P. M. Stecker, L. S. Fuchs, and D. Fuchs, "Using curriculum-based measurement to improve student achievement: review of research," *Psychology in the Schools*, vol. 42, no. 8, pp. 795–819, 2005.

[201] M. K. Burns and B. V. Senesac, "Comparison of dual discrepancy criteria to assess response to intervention," *Journal of School Psychology*, vol. 43, no. 5, pp. 393–406, 2005.

[202] D. Mellard, M. McKnight, and K. Woods, "Response to intervention screening and progress-monitoring practices in 41 local schools," *Learning Disabilities Research & Practice*, vol. 24, no. 4, pp. 186–195, 2009.

[203] C. H. Hofstetter, "Contextual and mathematics accommodation test effects for English-language learners," *Applied Measurement in Education*, vol. 16, no. 2, pp. 159–188, 2003.

[204] L. Fuchs and S. Deno, "Must instructionally useful performance assessment be based in the curriculum?" *Exceptional Children*, vol. 61, no. 1, pp. 15–24, 1994.

[205] P. Stecker and E. Lembke, *Advanced Applications of CBM in Reading (K-6): Instructional Decision-making Strategies Manual*, National Center on Student Progress Monitoring, Washington, DC, USA, 2011.

[206] G. Tindal, J. Alonzo, J. F. T. Nese, and L. Saez, "Validating progress monitoring in the context of RTI," in *Pacific Coast Research Conference (PCRC)*, Coronado, Calif, USA, February 2012.

[207] R. Gersten and J. A. Dimino, "RTI (Response to Intervention): rethinking special education for students with reading difficulties (yet again)," *Reading Research Quarterly*, vol. 41, no. 1, pp. 99–108, 2006.

[208] J. M. Hintze and E. S. Shapiro, "Curriculum-based measurement and literature-based reading: is curriculum-based measurement meeting the needs of changing reading curricula?" *Journal of School Psychology*, vol. 35, no. 4, pp. 351–375, 1997.

[209] J. Hintze, E. Shapiro, and J. Lutz, "The effects of curriculum on the sensitivity of curriculum-based measurement in reading," *The Journal of Special Education*, vol. 28, pp. 188–202, 1994.

[210] L. Fuchs, G. Tindal, and S. Deno, *Effects of Varying Item Domains and Sample Duration on Technical Characteristics of Daily Measures in Reading*, University of Minnesota Institute for Research on Learning Disabilities, Minneapolis, Minn, USA, 1982.

[211] Department of Education, *Assisting Students Struggling With Reading: Response to Intervention and Multi-Tier Intervention in the Primary Grades*, Institute of Education Sciences, Washington, DC, USA, 2009.

[212] Department of Education, *Improving Reading Comprehension in Kindergarten Through 3rd Grade*, Institute of Education Sciences, Washington, DC, USA, 2010.

[213] Department of Education, *WWC Evidence Review Protocol for K-12 Students with Learning Disabilities*, Institute of Education Sciences, Washington, DC, USA.

[214] E. R. O'Connor and P. Vadasy, *The Handbook of Reading Interventions*, Guilford Press, New York, NY, USA, 2011.

[215] R. M. Schwartz, M. C. Schmitt, and M. K. Lose, "Effects of teacher-student ratio in response to intervention approaches," *The Elementary School Journal*, vol. 112, no. 4, pp. 547–567, 2012.

[216] L. Saez, "Instructional responsiveness: what are teachers doing?" in *Proceedings of the Pacific Coast Research Conference*, Coronado, Calif, USA, Februrary 2012.

[217] B. Elbaum, S. Vaughn, M. T. Hughes, and S. W. Moody, "How effective are one-to-one tutoring programs in reading for elementary students at risk for reading failure? A meta-analysis of the intervention research," *Journal of Educational Psychology*, vol. 92, no. 4, pp. 605–619, 2000.

[218] R. E. O'Connor, "Phoneme awareness and the alphabetic principle," in *The Handbook of Reading Interventions*, R. E. O'Connor and P. Vadasy, Eds., Guilford Press, New York, NY, USA, 2011.

[219] R. Spicuzza, J. Ysseldyke, A. Lemkuil, S. Kosciolek, C. Boys, and E. Teelucksingh, "Effects of curriculum-based monitoring on classroom instruction and math achievement," *Journal of School Psychology*, vol. 39, no. 6, pp. 521–542, 2001.

[220] S. Deno, "Individual differences and individual difference: the essential difference of special education," *The Journal of Special Education*, vol. 24, pp. 160–173, 1990.

[221] D. L. Gast, *Single-Subject Research Methodology in Behavioral Science*, Routledge, New York, NY, USA, 2010.

[222] R. H. Horner, E. G. Carr, J. Halle, G. Mcgee, S. Odom, and M. Wolery, "The use of single-subject research to identify evidence-based practice in special education," *Exceptional Children*, vol. 71, no. 2, pp. 165–179, 2005.

[223] L. Fuchs, "Assessing intervention responsiveness: conceptual and technical issues," *Learning Disabilities Research & Practice*, vol. 18, no. 3, pp. 172–186, 2003.

[224] M. K. Burns, S. E. Scholin, S. Kosciolek, and J. Livingston, "Reliability of decision-making frameworks for response to intervention for reading," *Journal of Psychoeducational Assessment*, vol. 28, no. 2, pp. 102–114, 2010.

[225] S. P. Ardoin, "The response in response to intervention: evaluating the utility of assessing maintenance of intervention effects," *Psychology in the Schools*, vol. 43, no. 6, pp. 713–725, 2006.

[226] A. M. VanDerHeyden, J. C. Witt, and D. W. Barnett, "The emergence and possible futures of response to intervention," *Journal of Psychoeducational Assessment*, vol. 23, no. 4, pp. 339–361, 2005.

[227] J. B. Willet, "Measuring change more effectively by modeling individual change over time," in *The International Encyclopedia of Education*, T. Husen and T. N. Postlethwaite, Eds., Pergamon Press, Elmsford, NY, USA, 2nd edition, 1994.