Educational Assessment, 16:15–34, 2011

ISSN: 1062-7197 print/1532-6977 online
DOI: 10.1080/10627197.2011.551084

Routledge
Taylor & Francis Group

Examining a Grade-Level Math CBM Designed for Persistently Low-Performing Students

Daniel Anderson, Cheng-Fei Lai, Julie Alonzo, and Gerald Tindal
University of Oregon

Students with disabilities participate in two major measurement systems. The Individuals with Disabilities Education Act emphasizes working within a Response to Intervention (RTI) framework to identify and monitor the progress of low-performing students. Persistent low-performing students also may be eligible for some form of an alternate assessment for accountability purposes. Working within these two systems, educators need technically sound measures to inform decision making. This study presents scaling results from a Curriculum Based Measurement tool designed within an RTI framework and specifically for persistently low-performing students. We use the phrase "persistently low-performing students" to refer to a specific group of students who have been identified with a nonsevere learning disability and who perform well below grade-level expectations. Key findings indicate that items appear to function well in the lower tail of the distribution of students' estimated ability level. Further, the distribution of items is positively skewed, resulting in many accessible items that are most informative for low-performing students. Results provide initial validity evidence for the measurements as one source of data for progress monitoring within an RTI framework and the identification of persistent low-performing students who may be eligible for a large-scale assessment option other than the general grade-level assessment.

Response to Intervention (RTI) is a multitiered instructional paradigm that involves educators identifying and monitoring the progress of students at risk for low achievement through the application of screeners and progress monitoring. Often Curriculum Based Measurement assessments (CBMs; Busch & Reschly, 2007) serve both purposes. In this approach, CBMs are used to screen all students at a school and then to monitor the progress of those identified as being at risk for low achievement. CBM, as promulgated by Deno (1985), originated from the Data Based Program Modification Model in the late 1970s (Deno & Mirkin, 1977). Educators using CBMs repeatedly measure a student on equivalent alternate forms of the same test over time (Deno, 2003). Observed changes in the student's scores can then be attributed to a change in the student's knowledge. This approach is emphasized in the most recent reauthorization of the Individuals with Disabilities Education Act (2004) and provides a means both of identifying

Correspondence should be sent to Daniel Anderson, Behavioral Research and Teaching, University of Oregon, 275F Education, Eugene, OR 97403. E-mail: daniela@uoregon.edu

students for special education services and of targeting instruction and resources to meet the most pressing academic needs in a school system.

In this article, we focus on the appropriateness of using CBMs not only for screening students for additional intervention as part of an RTI system but also for identifying students for the most appropriate large-scale testing option. We use the phrase "persistent low performers" throughout this article to describe students who (a) have an identified learning disability not classified as "severe," and (b) perform far below grade-level expectations. This group of students has also been referred to as the "2% population" (Perie, 2009; Quenemoen, 2009), but we opt for referring to these students as persistent low performers in an attempt to describe the characteristics of the students rather than the policies surrounding them.

From a large-scale assessment-embedded accountability perspective, persistent low performers pose a substantial challenge, as the grade-level statewide assessment proves too difficult. Previous practice allowed for out-of-level testing, where persistent low performers would take the general education assessment from a grade level below the grade in which they were enrolled (Thurlow, Elliott, & Ysseldyke, 1999). This out-of-level testing was controversial, however, because some believed it held students to an inappropriately low standard. Federal policies thus required the option for out-of-level testing be removed (Lazarus & Thurlow, 2009). Current regulations allow for states to adopt an alternate assessment judged against modified achievement standards (AA-MAS), which must be aligned to grade-level content standards. Yet the AA-MAS has also been controversial (Quenemoen, 2009). With the reauthorization of the No Child Left Behind Act of 2002 (NCLB) pending, it remains unknown how testing policies and regulations will effectively include this group of students in accountability assessments. However, in this article we are less concerned with testing regulations than with a measurement system sensitive to student progress with implications for participation in the large-scale testing program. With RTI becoming an integral part of many schools' practice, there is added impetus for the development of an efficient system where teachers can use the same CBM measures for dual purposes: RTI progress monitoring and identification of persistent low performers.

The purpose of this study is to analyze the appropriateness of using a CBM developed for use in an RTI system as an instrument to identify persistent low performers. The RTI lens requires an assessment that can both reliably identify students' proficiency in relevant skills for screening and provide sensitive measures of growth on comparable alternate forms to monitor instructional progress. The measures need to be sensitive to students with persistent learning problems yet be aligned with state (or national) grade-level content standards. Given these different needs, we evaluate an easyCBM® fifth-grade mathematics RTI measure to provide validity evidence for both of these uses, progress monitoring and participating in a large-scale test.

We examine the appropriateness of the assessment with the Rasch model, using person-item maps showing students and items on the same scale to ascertain sensitivity. In addition, we map the functioning of each item through tracelines showing the percentage of students responding correctly at various estimated ability levels to provide evidence that the items are appropriate for the full range of low-achieving students. We present this evidence to document the degree to which measures designed specifically for persistent low performers can serve concurrently as an accurate source of data for informing decisions about both RTI and identification of a large-scale testing option.

In the RTI model, both screening and progress monitoring are used to appropriately target instructional programs for specific student needs. Typically, students who score below a normative cut point (e.g., below the 20th percentile) on the screening assessment are assigned to Tier 2, where they are given an intervention intended to address their specific academic needs. CBM probes are then administered intermittently (i.e., weekly, biweekly, or monthly) to monitor the intervention's effectiveness. Students failing to respond adequately at this level after a district-specified amount of time or a series of interventions are then assigned to Tier 3, which provides the student with the most intensive interventions, including possible special education placement (Ardoin, Witt, Connell, & Koenig, 2005).

A number of studies have shown the technical features of CBM to be quite strong (Christ, Sculin, Tolbize, & Jiban, 2008; Deno, 2003; Fuchs, 2004). As RTI becomes more widely adopted, it is likely to become increasingly clear that the most appropriate CBMs for use in an RTI framework are those that enable educators to monitor the progress of students who perform—at least initially—significantly below grade-level expectations. We believe the information from the screener may also serve to identify a large-scale testing option that is different than the standard grade-level test but is not as different as the alternate assessment judged against alternate achievement standards. This latter option is reserved for students with significant cognitive disabilities who perform significantly below grade-level expectations. By applying principles of universal design for assessment—a test development process aimed at maximizing accessibility of items—this assessment can serve in an RTI mode as well as provide an appropriate option for large-scale testing.

PERSISTENT LOW PERFORMERS

NCLB (2002) requires that all students be tested annually in Grades 3 to 8 in reading and math. Personnel within schools must ensure an increasing percentage of students reach proficiency benchmarks each year until the 2013–2014 school year, when 100% of students must reach the state's proficiency benchmark to avoid potential schoolwide sanctions. Under the current 2001 version of NCLB, up to 1% of students with the most significant cognitive disabilities, often referred to as *the 1% population*, are allowed to be given an alternate assessment judged against alternate achievement standards (AA-AAS) for reporting purposes (U.S. Department of Education, 2007). Until 2007, there were three tests available to educators to meet the NCLB reporting requirements: (a) general grade-level assessment, with or without accommodations; (b) alternate assessment based on grade-level achievement standards; and (c) alternate assessments based on alternate achievement standards (Perie, 2009). Persistent low performers, however, were not appropriately assigned to any of these tests. The general grade-level assessment displayed a "floor effect," as students were not able to effectively participate in the measurement process and demonstrate their knowledge, yet their cognitive functioning did not qualify them for the AA-AAS. Out of this critique came a fourth assessment type, AA-MAS, which under NCLB (2002) allows up to an additional 2% of the total number of students assessed to take tests developed with an expected student outcome differing from the general grade-level tests (Perie, 2009).

With the reauthorization of NCLB on the horizon, the future of these specific assessment options is unknown. Although there has been some debate as to which students are eligible for the AA-AAS, educators and policymakers generally agree that these students are distinct

from the general education population. Much more confusion and contention exists, however, regarding which students are persistent low performers and which assessment option is most appropriate (Quenemoen, 2009). The current test for persistent low performers, the AA-MAS, requires three basic eligibility criteria. First, the student must be identified with at least one of 13 disability classifications defined by Individuals with Disabilities Education Act. Second, the student must have an individual education plan (IEP). Third, the student's IEP team must deem the assessment appropriate given the student's disability, using multiple sources of objective measurement.

At present, the most ambiguous requirement is the judgment of the IEP team. The U.S. Department of Education (2007) provided guidelines for deciding who should be assessed with the current version of the test for persistent low performers: Students must be included because of a disability that precludes them from taking the general assessment, not because of the student's racial or economic background or because of the lack of quality instruction. Given that students identified as persistent low performers are held to a different achievement standard for statewide accountability purposes than students not identified, it is essential that the IEP teams have as much information as possible to ensure accurate decisions are made.

According to Quenemoen (2009), many proponents of students with disabilities (SWD) are concerned that the current regulations for persistent low performers reduce educators' expectations for these student and compromise instruction. The U.S. Department of Education (2006) justified moving to different achievement standards for this group of students, citing research indicating that "there are about 1.8% to 2.5% of children who are not able to reach grade level standards, even with the best instruction" (U.S. Department of Education, 2006, para. 12). Therefore, providing a 2% cap on statewide assessments specifically targeted to this student population represents a compromise between trying to hold students to a high standard but present an option more closely aligned with the general education grade level test than is typical of most alternate assessments.

The controversy, however, has not been resolved. The numbers reported by the U.S. Department of Education were obtained through a retrospective analysis, which weakens the argument, as the studies were not originally designed to identify students who unable to reach grade-level standards with the best instruction. Furthermore, a plethora of definitions exist for persistent low performers, and whether they are indeed a distinct group from both the 1% population of students with severe cognitive disabilities and the general education population is still hotly debated (Filbin, 2008; Quenemoen, 2009).

A recent report authored by Filbin (2008) on behalf of the U.S. Department of Education was released to clarify prominent areas of confusion for states attempting to design a statewide test for persistent low performers. The report predominately focused on two areas: state policies guiding IEP teams and test development methods used by vendors. The report acts as a call for research on clearly articulating the population, stating,

> To date, little empirical research exists to define who the most appropriate students are for the AA-MAS or how their learning and performance abilities differ from students performing successfully on the general assessment. Yet understanding how the cognitive processing of learners with disabilities within the different domains assessed through the AA-MAS differ from those students tested through the general or AA-AAS is essential to inform test development decisions. (Filbin, 2008, p. 3)

Both educators and test developers thus need multiple sources of evidence and technically adequate tools to aid in the identification of persistent low performers. A CBM that can be used to monitor the progress of students over time and function well with low-performing students may provide valuable information for this kind of decision making. However, validity evidence is needed to document measurement sensitivity within a distribution of students and items/responses.

Using RTI to Identify Persistent Low Performers

Educators have had difficulty identifying which students should be classified as persistent low performers, as they exist in a "gap" between the general education students and the 1% of students with severe cognitive disabilities. We suggest that the use of CBM in an RTI framework may be a valid method of acquiring and supplementing existing data for informing identification decisions, particularly if the CBMs are designed to measure low performing students. In the RTI approach, the student is first screened and then monitored over time to observe how he or she is responding to instruction. Given that the current regulations for participation in the statewide test for persistent low performers require multiple sources of objective measurement, it is logical that one such source examines *learning* rather than the static achievement of the student at a single point in time. The evidence of rate of learning is more justifiable if measured through technically adequate CBMs linked to grade-level content standards than if the evidence is primarily composed of anecdotal teacher notes. Moreover, if CBMs are to be used to inform identification decisions and teachers' instructional decision making within RTI, they should be accessible to persistent low performers while not sacrificing alignment with the grade level content standards. If the screener were a parallel form of the measures used for progress monitoring, then the RTI model would have great implications for use in identifying students for an appropriate large-scale test option.

Critical Features in the Development of Tests for Persistent Low Performers

A well-designed test for persistent low performers should provide evidence in four critical areas: reliability, validity, accessibility, and alignment with grade-level content standards. As with any assessment, reliability and validity are paramount to ensuring the target construct is being measured accurately. Without these features, the test is, in essence, meaningless. Perhaps just as critical, however, is the accessibility of items—without which validity and reliability are compromised. Finally, given that current versions of statewide tests for persistent low performers must adhere to grade-level content standards, it is important that any tool used for identification similarly reflects these standards. Maintaining a clear connection to grade-level content standards also helps ensure high expectations of student performance, a feature particularly relevant for RTI decision making.

Reliability. An important requirement for tests is that the results can be reproduced, or be reliable. Generally, reproducibility is a function of the number of items and can be enhanced by including more items. The more times one measures an individual on a specific task (i.e., number of items or test forms), the closer the measurement comes to the *true* ability level

of the individual. When ability is examined through any single measure, error is likely to be large. For instance, the individual may perform above or below his or her true ability level by simply guessing the correct option or making errors due to extraneous circumstances such as distractions in the testing environment. But as the number of measurement occasions increases, these factors begin to average out, creating a more accurate measurement of true ability (Feldt & Brennan, 1989).

When designing a test for persistent low performers, some test developers have attempted to follow the same blueprint used to develop general education assessments, with the primary modification being a reduction in the number of items rather than the type of items. However, this approach substantially weakens the overall reliability of the test (Filbin, 2008), as test length and reliability are inextricably linked. Test developers might also justify this approach as an attempt to decrease the amount of reading required of the student. However, this goal can be accomplished by another, more psychometrically sound practice: reducing the item answer options from five or four, to three. In a meta-analytic review of more than 80 years of item option research, Rodriguez (2005) found moving from a five- or four-option item to a three-option item reduced the item difficulty but did not affect the reliability. This change in item construction would achieve the goal of reducing the amount of reading required of the student while maintaining the reliability of the score. A test for low-performing students could thus maintain an appropriate number of items and be simplified in psychometrically sound ways, such as limiting answer option choices to three.

Validity. When designing a test for persistent low performers, it is difficult to adhere to a specific definition of the target population. However, it is important that a general definition of the intended users be kept in mind from the inception of the test so that each developmental aspect can be tailored to their characteristics. According to Messick (1989), "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores" (p. 13 emphasis in the original). Therefore, it is the use, interpretations, and decisions made from the test that determine validity, not the test itself. A valid use of a test could be thought of as using the tool to provide accurate information from which decisions and inferences can be made with reasonable confidence. This confidence, and the validity of the use of the test, may be considerably downgraded if a measure was originally designed for a different purpose than its actual use.

Often, tests used for persistent low performers are retrofitted general education assessments (Kettler & Elliott, 2009). General education items on these tests are modified with the intent of increasing accessibility. Although these modifications are often made with the new target population in mind (Kettler, Elliott, & Beddow, 2009), the item as a whole was still originally designed for a different population, thereby weakening the validity of its use with persistent low performers.

Given the challenges of validating a test for use with persistent low performers, a national technical advisory committee on behalf of the U.S. Department of Education released a report in November 2008 to guide test developers in this effort (U.S. Department of Education, 2008). This document contains a section entitled Possible Probing Questions that highlights areas for further investigation of alternate assessments. One probing question pertains to how persistent low performers have been identified. This focus reiterates Filbin's (2008) point about

the importance of developing items with the target population in mind. A test for persistent low performers should thus be tailored to the characteristics of this population from the beginning stages of development.

Accessible items. Designing items that are accessible to as many students as possible is a substantial challenge. Universal design (UD) has shown promise as a developmental tool. Key elements of UD include: (a) a consideration of *all* characteristics of test takers; (b) precisely defined constructs; (c) accessible, nonbiased items; (d) items amendable to accommodations; (e) simple, clear, and intuitive instructions and procedures; (f) maximum readability and comprehensibility; and (g) maximum legibility of text, tables, figures, and illustrations (Thompson, Johnstone, & Thurlow, 2002).

Accessibility can be increased by simplifying the language, reducing the number of answer options, adding graphics, and reducing the total number of items on the test (Filbin, 2008). Shaftel, Yang, Glasnapp, and Poggio (2005) used many of these strategies to modify a fourth-grade mathematics test. Fifteen modified fourth-grade math items were administered to students in both special and general education programs for comparison. A one-parameter Rasch model was used to estimate item difficulty and student mathematics ability on the 15 common items. Items were scaled for both populations accordingly. Results indicated that the modified test had more items at a lower difficulty, which SWD could better access. Item characteristic curves indicated that the modified tests were more accurate at measuring students of lower estimated abilities. The test modified for persistent low performers provided a more targeted measurement of expected outcomes for lower achieving students.

Although the findings of Shaftel et al.'s (2005) study supported a separate modified test, the authors noted that a general test with an adequate number of items at the appropriate difficulty level would be able to accurately measure some of the SWDs with higher ability who took the modified tests. Similarly, a modified test can also have items with greater difficulty level to better measure these students. Further, the principles used to *modify* items can be used in the development of *new* items. Applying the principles of UD increases the accessibility of the items for use with persistent low performers by beginning the developmental process with the target population in mind, rather than retrofitting items previously designed for a different population.

Two other empirical studies have examined the effect of a UD test on student outcomes (Dolan, Hall, Benerjee, Chun, & Strangman, 2005; Johnstone, 2003). These studies provide some initial evidence supporting the use of UD; however, they are limited in their findings. The internal validity of the Johnstone study is questionable because random assignment of students into traditional and universally designed test groups was not explicitly stated. The generalizability of the Dolan et al. study is limited because of the small sample ($n = 9$). Despite these limitations, the studies do provide some evidence of the utility of UD.

Alignment with grade-level standards. One of the primary difficulties in designing a test for persistent low performers is reducing the complexity of the test to a degree that students can access items and demonstrate ability yet simultaneously adhering to and maintaining alignment with the grade-level content standards. This challenge has left test developers in a "gray zone" where items must be simplified but not so much that they align better with the content standards of a lower grade. This problem has been addressed in a variety of ways with varying levels of

success, but to date, most test development has begun by modifying existing items instead of starting anew. Modifying items saves money over developing new items, and it should result in a high degree of alignment with the grade-level standards, given that the modified items are based directly on the aligned grade-level items. However, the compromise may be reduction in technical adequacy.

When test items are changed to increase accessibility, the change should not compromise alignment to the content standard they were originally built to measure. The first step in modifying a general education item into an item for persistent low performers is to examine the original item for accessibility (Kettler et al., 2009), a largely judgmental process. One general approach has been to first select items from the general assessments that appear easiest for SWD and then to reduce the number of options and/or language complexity for these items. Exploratory studies have examined the effect of these modifications, yet none have provided strong evidence for specific guidelines that should be used to change existing test items into appropriate items for persistent low performers.

Johnstone, Bottsford-Miller, and Thompson (2006) conducted a cognitive lab study with 231 sixth-grade students from traditionally underperforming schools. Their purpose was to better understand the cognitive processes of persistent low performers. The results provide some promising implications for the methods used in item modification and development. For instance, the think-aloud strategy may aid in the understanding of problem-solving processes (Johnstone et al., 2006) and identifying specific linguistic challenges unique to SWDs (Johnstone, Liu, Altman, & Thurlow, 2007). Therefore, items can potentially be changed or developed in alignment with the content standard but maintain accessibility by accounting for these unique characteristics. However, although the think-aloud strategy may show promise, it is limited to students who are concrete in their language and thought processes. It does not work well if the item is difficult (particularly in math) or if students have difficulty verbalizing their thought processes, nor does it work well with students with severe cognitive disabilities (Johnstone et al., 2006). However, further research in this area may increase test developers' ability to create accessible items for persistent low performers that maintain alignment with the grade level content standard.

When IEP teams identify students as persistent low performers, the most appropriate testing option may be different from the general grade-level assessment, Previous practices assessed these students with out-of-level tests, whereas current regulations allow for the use of an assessment judged against modified achievement standards. Future regulations may provide still other options, yet it is likely that any such option continues to require multiple sources of objective measurement.

One of these measurement sources should be a test designed specifically for persistent low performers. However, to ensure the student is not just identified but also continues to learn at an acceptable rate, the progress of the student should be monitored over time using CBMs. If the school or district is working within an RTI framework, identification and progress monitoring should be part of that system. Both identification procedures and RTI require data. One well-designed measurement tool may fill the needs of both. To provide validity evidence, this tool should be aligned with grade-level content standards and be specifically designed for low performing students. In addition, validity evidence needs to highlight the fit of the items with the students. In the following sections, we present Rasch scaling results from a tool designed in such a manner.

METHODS

Setting and Subjects

During the fall of 2009, we collected data from two midsized school districts located in the Pacific Northwest serving approximately 41.5% and 60.1% of students from economically disadvantaged backgrounds, respectively. Each participating district assessed all students who were in attendance when the fall easyCBM® benchmark assessment was administered. As part of their district-mandated RTI programs, both districts require that all students, including students with special needs and English language learners, participate in district benchmark assessments. The demographics of both districts were largely homogenous, with 71.4% and 72.2% of students identifying as White, 12.7% and 7.8% Hispanic, 1.9% and 5.0% Asian, 3.0% and 1.5% Native American, respectively. Of the total sample, 8.3% of students had missing data or had declined to identify ethnic background. District 1 had 8.1% English language learners (ELL), 2.2% identified as gifted or talented and 19.9% in special education programs, whereas District 2 had 3.9% ELL, 5.3% identified as gifted or talented and 16.7% in special education. Each district had roughly an even number of male and female students. Given that the current study examines mathematics CBMs, it is important to note that 83% and 76% of students in Grade 5 in each district had passed the previous year's state math proficiency test; however, state test data were not accounted for in any of the analyses.

Data from both districts were combined to obtain a larger sample. Although we found similar results at each grade level, in this study, we focus exclusively on data from students in Grade 5 to limit the length of our article. We chose this grade because fifth grade acts as a "gateway" grade where decisions are made concerning statewide testing options for persistent low performers. Decisions related to statewide testing options occur for the first time at fifth grade because the current policy requires that eligibility data be based on more than one year of performance on the state test and most states do not begin state testing until Grade 3, when it is mandated under NCLB. Our sample size ranged from 2,085 to 2,099 students per item, with 135 to 149 missing cases per subtest. We detected no pattern in missing cases; they appeared to be missing at random.

Measurement/Instrument Development

The easyCBM® online benchmark and progress monitoring system is available at http://www.easycbm.com. The mathematics measures on easyCBM® are based on the three National Council of Teachers of Mathematics (NCTM) Focal Point Standards per grade level. There are three 16-item fall benchmark tests within each grade, each corresponding with a single Focal Point standard, and 10 additional progress monitoring probes per Focal Point at each grade, developed concurrently with the benchmark tests. The three main steps in developing the math measures were: (a) item writing; (b) item piloting to determine difficulty, reliability, and appropriateness for use with the target grade level; and (c) assembling equivalent forms for benchmarking and progress monitoring. The NCTM Focal Points for Grade 5 are: *number and operations*; *geometry, measurement, and algebra*; and *number and operations and algebra*.

Reliability. The reliability of easyCBM® math measures has been shown to be strong. Anderson, Tindal, and Alonzo (2009) examined the internal consistency of test forms in

Grades 1 to 8. The sample for Grade 5 ranged from 1,269 to 1,270. Items had a mean of .71 (with dichotomously coded data: 0 = incorrect, 1 = correct), mean variance of .17, and a mean number of items correct of 34.16 of the 48 total items across the three tests with a standard deviation of 7.04. The interitem correlations had a mean of .11. A Cronbach's alpha of .85 indicated strong internal consistency.

Validity. The validity evidence supporting easyCBM® as a source of data for RTI and the identification of persistent low performers stems primarily from two sources: (a) the development of the measures specifically for use within an RTI framework with persistent low performers; and (b) the adherence of items to national grade-level standards. In addition, as we present in the Results section, the distribution of items and students along with the functioning of items should provide validity evidence on making inferences on progress within RTI and participation in large-scale testing programs The inferences and actions taken from the results can thus be deemed appropriate if used for benchmark screening or progress monitoring within an RTI framework. Further, having items linked to grade-level standards increases the validity of the measures' use as a supplemental data source for the identification of persistent low performers because decisions around the current test for persistent low performers *must* be linked with the grade-level standard. If easyCBM® items were not developed to grade-level standards, it would likely not be valid as a supplemental source of data because the grade level of student performance would be unknown, thus leading to incorrect educational decisions. For example, students may show high progress on CBMs, providing educators with evidence that those students should not be included in an assessment other than the general grade-level assessment. However, if the observed progress is being measured at a level below the students' grade (e.g., items simplified to an extent that they align better with a lower grade-level standard) this may not be the correct decision. If students were measured with grade-level items, a different conclusion may be reached. Because easyCBM® was designed for use nationally, rather than for a single state, a general rather than specific definition of persistent low performers was used in item development and review.

Accessibility through universal design. All easyCBM® mathematics measures were developed adhering to the principles of UD. For instance, to minimize construct irrelevant variance, the item reviewers verified that each item targeted a single construct with few extraneous stimuli. During the item review process, developers eliminated culturally biased language and any culturally specific background knowledge. In item design and development, developers reduced language complexity and designed the computer interface to be as simple and accessible as possible for students with visual or fine-motor impairments. Some of the computer interface adaptations include the ability to magnify text and graphics on the screen and to select answer responses by clicking anywhere on the answer choice box (rather than only on a single bullet). After students respond to an item, they click a "next" button to see the next item or a "back" button to go to the preceding item and change a response. This design feature eliminated the need for students to scroll up or down. If students attempted to proceed to the next question without responding to the item on the screen, they were prompted with the message "Please pick an answer first" and were not allowed to continue until they had selected a response. An algorithm randomly rotates all item options each time the item appears

to prevent "copying" by students sitting next to each other in a computer lab. In addition, only one test item appears on the screen at a time.

Adhering to grade-level standards with reduced complexity. To make the items as accessible as possible for persistent low performers, items targeted at-grade-level NCTM Focal Point Standards while keeping the language, numbers, and graphics as simple as possible. For example, where the Focal Point called for multidigit multiplication, the numbers in most items were confined to two or three digits. As another example, if the Focal Point called for reading and analyzing charts, a simple chart was presented with only the information necessary to answer the question. These considerations allowed students to demonstrate mastery of the target constructs while reducing the chance of errors related to simple computation, reading, or skills other than the target construct. Finally, all items had only three answer choices. Therefore the number of steps, working memory demand, and extraneous information were minimized in most items. Relevant to our focus here, the most important features of the test development process were the alignment with grade-level content, adherence to UD, and the reduction of complexity to increase item accessibility for persistent low-performers.

Data Preparation and Analysis

Data from both districts were transferred into a single file with three fields: (a) the option selected; (b) whether the item was correct or not (coded dichotomously: 0 = incorrect and 1 = correct); and (c) the focal point domain. Data were analyzed using a one-parameter logistic Rasch model with Winsteps 3.68. Subtests within each grade were analyzed concurrently. To examine how items functioned with different populations (i.e., low-performing students and average or high-performing students), we graphed all items by the percentage of students responding correctly within different estimated ability percentile ranges. We created four new variables: (a) a total score for the test being analyzed for each student; (b) the total number of students scoring in that range, which accumulates with each student; (c) the student's percentile, based on the accumulated *n* variable and the total number of valid cases; and (d) the quintile of estimated ability for each student from the raw percentile rank (0–20th percentile = 1, 20.01–40th percentile = 2, etc.). Data were then organized in ascending order by the total score on the subtest being analyzed, and line charts were created with each item appearing as a new line on the chart, estimated ability by quintile on the *x*-axis and the percentage of students responding correctly within each quintile range on the *y*-axis. This procedure was repeated for each subtest. Finally, percentile was recoded into a new variable labeled 'lowperformers,' which divided students into the 0–5th, 5.01–10th, and 10.01–15th percentile ranges of estimated ability. The charts were thus redrawn with this group of students to produce a second figure.

RESULTS

Results of the Rasch scaling analysis are presented by each NCTM Focal Point Standard in Table 1, Table 2, and Table 3. Items ranged in difficulty from −2.56 (*geometry, measurement, and algebra*; Item 1) to 2.58 (*geometry, measurement, and algebra*; Item 5). Overall, 24 items had a difficulty greater than 0, and 24 items had a difficulty less than 0. Each NCTM focal

TABLE 1
Rasch Analysis for Grade 5 Focal Point—Number and Operations

Item No.	*Difficulty*	*Outfit*	*Point Measure*	*Observed Match*	*Expected Match*	*SE*
1	−1.97	0.99	0.25	93.9	93.9	.09
2	−1.99	0.61	0.27	94.0	94.0	.09
3	−1.40	0.86	0.28	90.0	90.0	.08
4	−1.58	0.73	0.31	91.3	91.3	.08
5	2.45	1.40	0.29	79.0	78.4	.06
6	−0.45	0.77	0.42	79.7	79.4	.06
7	0.07	0.83	0.43	74.4	72.8	.05
8	−0.37	0.85	0.39	79.5	78.4	.06
9	−0.36	0.86	0.39	79.5	78.2	.06
10	0.70	0.95	0.43	68.9	67.5	.05
11	0.56	0.86	0.46	70.9	68.2	.05
12	1.34	1.05	0.37	66.7	68.1	.05
13	0.99	1.11	0.30	60.0	67.0	.05
14	2.26	1.29	0.30	75.1	76.4	.05
15	1.80	1.04	0.40	71.1	71.7	.05
16	−0.19	0.85	0.39	77.3	75.9	.05

point was reasonably well distributed in difficulty. *Geometry, measurement, and algebra* was the easiest focal point, with nine items below 0 and 7 above. *Number and operations* was more difficult, with eight items below 0 and 8 above. *Number and operations and algebra* was the most difficult focal point, with seven items below 0 and the remaining nine above. Data fit the model well, with outfit statistics ranging from .61 to 1.65 and centering relatively close

TABLE 2
Rasch Analysis for Grade 5 Focal Point—Geometry, Measurement, and Algebra

Item No.	*Difficulty*	*Outfit*	*Point Measure*	*Observed Match*	*Expected Match*	*SE*
1	−2.56	1.40	0.10	96.4	96.4	.12
2	−1.70	0.87	0.25	92.2	92.2	.08
3	−1.24	1.65	0.14	88.4	88.5	.07
4	−1.62	0.84	0.24	91.7	91.7	.08
5	2.58	1.56	0.14	74.4	79.9	.06
6	−1.79	1.08	0.18	92.8	92.8	.09
7	−0.33	1.20	0.22	75.9	77.8	.06
8	0.17	0.96	0.35	70.5	71.7	.05
9	−0.07	0.92	0.36	74.7	74.5	.05
10	0.36	1.11	0.32	69.5	69.7	.05
11	0.90	1.07	0.35	64.6	67.0	.05
12	−0.76	1.05	0.25	83.5	83.3	.06
13	0.53	1.16	0.28	65.2	68.4	.05
14	0.93	1.05	0.36	63.4	67.0	.05
15	1.03	1.05	0.36	65.0	67.0	.05
16	−1.01	1.04	0.25	86.2	86.2	.07

TABLE 3
Rasch Analysis for Grade 5 Focal Point—Number and Operations and Algebra

Item No.	*Difficulty*	*Outfit*	*Point Measure*	*Observed Match*	*Expected Match*	*SE*
1	−0.11	1.34	0.21	73.5	75.0	.05
2	−1.34	0.77	0.33	89.4	89.4	.07
3	−0.50	0.72	0.43	80.7	80.0	.06
4	0.52	0.93	0.40	69.1	68.5	.05
5	1.01	0.99	0.42	68.2	67.0	.05
6	−0.04	0.84	0.42	75.0	74.1	.05
7	−0.21	0.92	0.40	78.3	76.3	.05
8	−0.95	0.95	0.32	85.7	85.5	.07
9	−0.03	0.92	0.38	74.8	74.0	.05
10	0.30	0.94	0.41	72.7	70.4	.05
11	0.34	0.86	0.46	73.1	69.9	.05
12	0.72	0.95	0.43	69.2	67.4	.05
13	1.05	0.94	0.45	70.2	67.1	.05
14	0.67	1.00	0.38	67.2	67.6	.05
15	0.93	1.09	0.35	64.8	67.0	.05
16	0.36	0.98	0.36	67.1	69.8	.05

to 1. Of the 48 items, 28 had an outfit value of less than 1, and the remaining 20 cases had a value greater than 1. The point measure correlations had low-moderate values, ranging from .10 (*geometry, measurement, and algebra*; Item 1) to .46 (*number and operations*; Item 11/*number and operations and algebra*; Item 11). The observed match of the items was generally close to the expected match. Measurement error was generally quite low, ranging from .05 to .12.

Figure 1 shows the difficulty of each item mapped against the estimated ability of students in the sample. A vertical line separates the two, with students appearing on the left and items appearing on the right. The right portion of the figure, items, shows items with a more frequent correct response rate (i.e., easier items) near the bottom of the figure and items with a less frequent correct response rate (i.e., difficult items) near the top. The left portion, persons, shows students of higher estimated ability near the top of the figure and students of lower estimated ability near the bottom. The "M" on each graph represents the mean of each distribution, the "S" represents 1 standard deviation from the mean, and the "T" represents the upper and lower 10% of the distribution (i.e., the 10th and 90th percentiles). Each pound sign (#) represents 16 students and each @ symbol represents a single student. The mean of item difficulties is slightly more than one standard deviation below the mean of the estimated student abilities.

Figure 2 and Figure 3 graphically represent the functioning of all 16 items from one of the fifth-grade Focal Points Standards: *Number and operations*. Although the figures were produced for all Grades 1 to 8 and all subtests, we highlight only one here as an example. The representation of items is typical of the other tests analyzed. In each figure, the *x*-axis plots the estimated ability of students in the sample by their total score on the test, grouped into percentile ranges. The *y*-axis plots the average percent correct by students within that range for the given items (lines). The figures are essentially the same, with each representing a different group of students. Figure 2 represents all students in the sample, grouped by estimated ability into quintiles, whereas Figure 3 represents only low achievers—defined here as those below

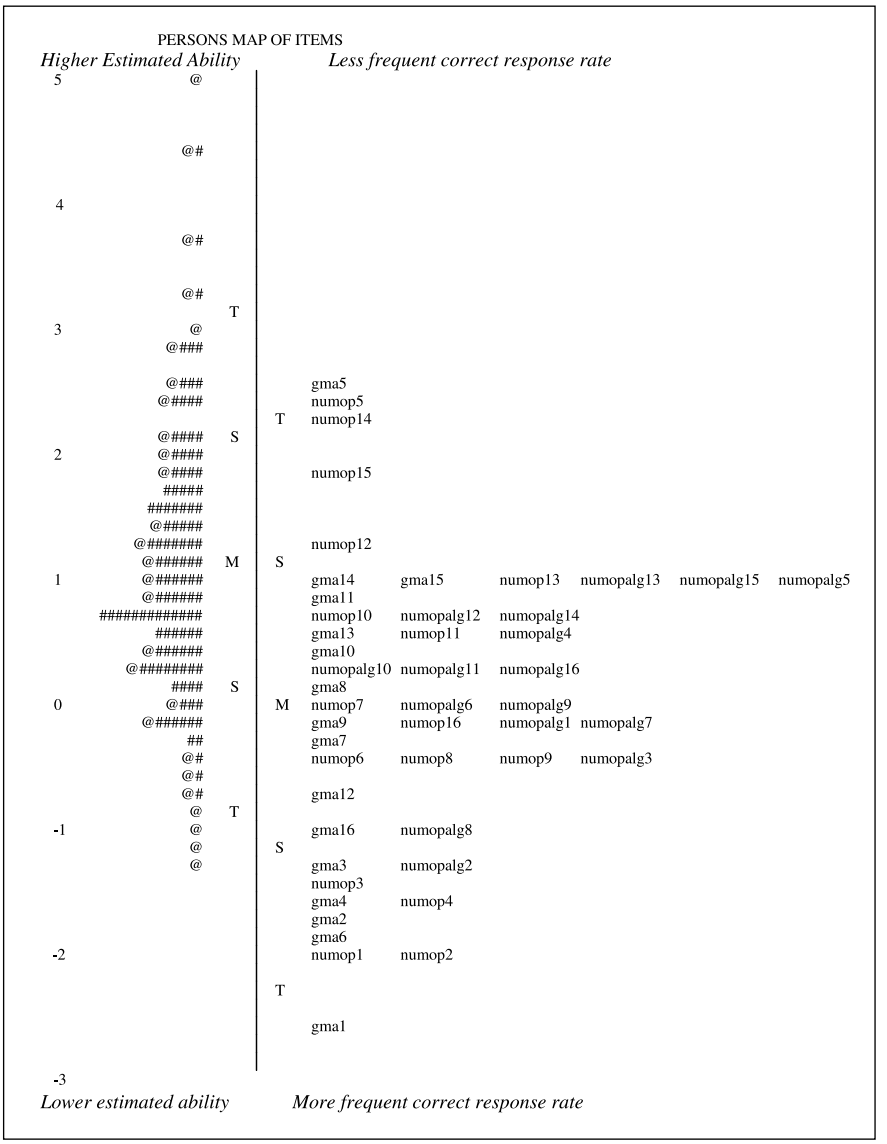

FIGURE 1 Person-item maps. *Note.* Each # represents 16 students, each @ represents one student.

the 15th percentile and grouped into three percentile ranges: 0–5, 5.1–10, and 10.1–15. Each line on the graph represents a different item from the test. Items are represented by the same type of line in each figure (i.e., Item 1 is solid black line in Figure 2 and Figure 3) and are flagged with the item number on each. Well-functioning items should have a monotonically increasing percentage of students responding correctly as ability increases. Given that Figure 3 represents a much smaller portion of the sample, the *n* is considerably lower within each

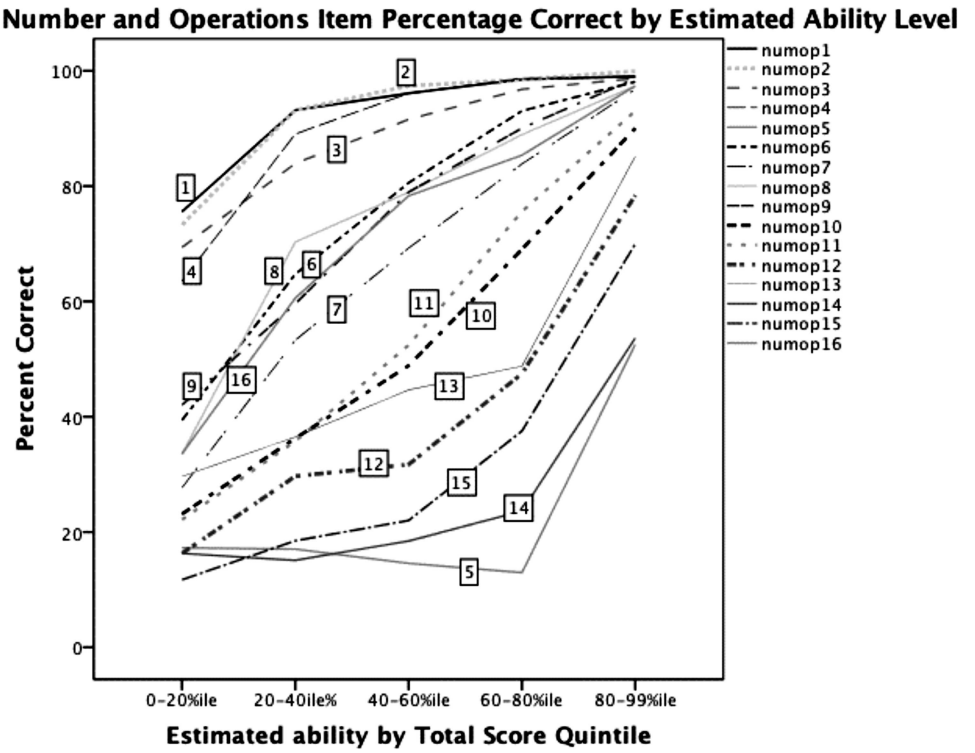

FIGURE 2 Item functioning by quintile group—All students. *Note.* Total valid $n = 2099$: 0–20%ile = 312, 20.1–40%ile = 411, 40.1–60%ile = 309, 60.1–80%ile = 533, 80.1–99%ile = 534. 135 cases were missing.

percentile range. The items are thus more variable, and the observed item functioning may be sample specific. Of the 2,085 total valid cases for the *Number and operations* subtest, only 92 fell within the 0–5 percentile range, 84 within the 5.1–10 percentile range, and 129 within the 10.1–15 percentile range. Thus, despite the relatively large original sample size, only 312 students, or 13.9% of the total sample (including missing cases), are represented in Figure 3.

Figure 2 shows that the majority of items function well and have high potential to sort students in the general education population by ability. Only four items appear to exhibit signs of a "ceiling effect." Items 5 and 14 appear to be very difficult items through approximately the 60th percentile of estimated ability, at which point each rises steeply. This finding may suggest that a particular option choice distracter is enticing to students of lower ability, but the enticement is diminished past a certain ability level. Overall, however, the majority of items show a roughly monotonically increasing slope. No items appear to show a floor effect, where students are not able to access items and access the scale. Within the 0–5 percentile range of estimated ability, the lowest measured ability range, a large portion of items still appear accessible and no items had a 0% correct response rate. A number of items show erratic functioning (i.e., Items 10, 11, and 15), and it is unknown whether this is an issue of sample size or with the items themselves.

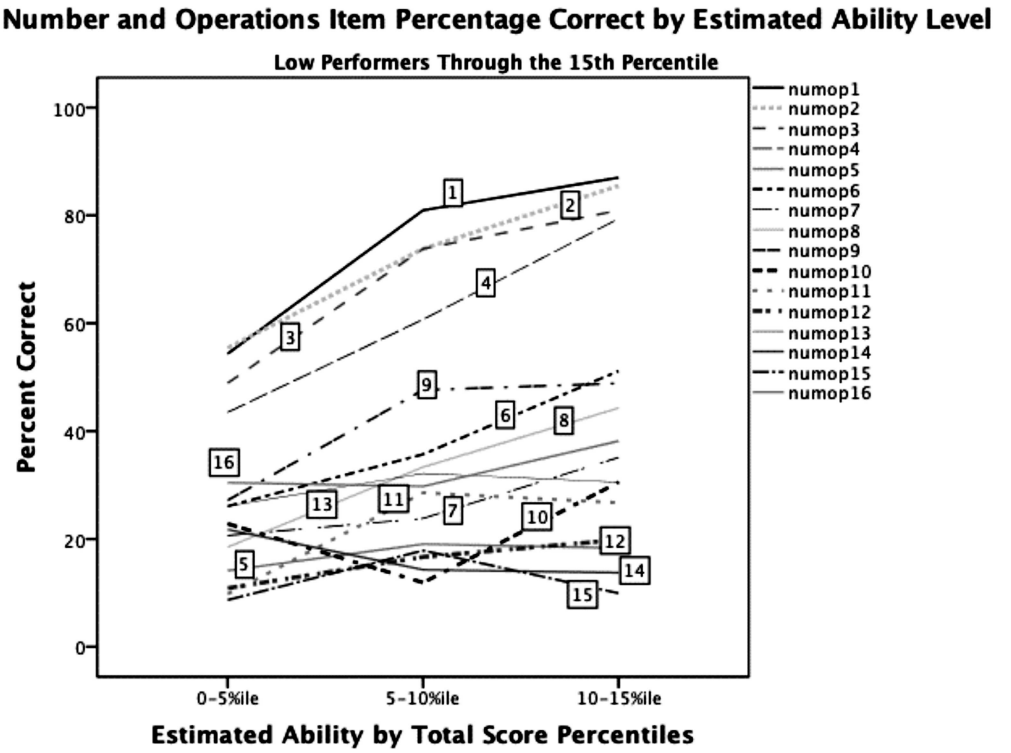

FIGURE 3 Item functioning by percentile group—Low performers. *Note.* Total $n = 305$: 0–5%ile = 93, 5.1–10%ile = 86, 10.1–15%ile = 133. Figure 3 represents 13.9% of the total sample, including missing cases.

DISCUSSION

This study examined the accessibility of the easyCBM® benchmark and progress monitoring assessments for very low-performing students. Our goal was to provide validity evidence for using the test as a source of data for both RTI and as a multiple measure for the identification of persistent low performers. The measures on easyCBM® were of particular interest because of the process used in their development—adhering to grade-level content standards while focusing on creating accessible items for low-performing students.

Although the current study examined accessibility at a single point in time rather than a longitudinal model to track student progress (i.e., Do items function differently over time?), the results indicate a high potential to sort students by ability at the lower tail of estimated ability on one form (benchmark assessment). This finding does not suggest that only one form should be utilized for educational decisions. Rather, it suggests that the form we analyzed appears to be one potentially useful tool to inform educational decisions. When the difficulty of each item was mapped against the estimated ability of students in the sample, the mean of the items fell one standard deviation below the mean of students' estimated ability level. Thus, students with an ability level far below the grade-level expectation would still be able to access many items,

producing more reliable results. However, caution should be advised when viewing easyCBM® results obtained from high ability students (i.e., one standard deviation above the mean), as fewer items fell within this range, and measurement error is likely greater.

One of the major statistical limitations of this study was that of the 2,083 total valid cases, only 104 fell within the 0–5 percentile range, 102 within the 5.1–10 percentile range, and 174 within the 10.1–15 percentile range. The observed functioning of items is thus more variable and may be sample specific. Despite the low number of students in each estimated ability range, the results provide some preliminary evidence of the item functioning for low performers. Many items show a roughly monotonically increasing slope. The small sample size could, however, be the reason for the erratic functioning of other items, as each student substantially affects the total percentage of students responding correctly/incorrectly (e.g., Items 10, 11, and 15). Yet it remains unknown whether the functioning displayed for items in Figure 3 accurately represents the true functioning of the items or the characteristics of the sample.

Further, the results presented in Figures 2 and 3 are only from one subtest in Grade 5 aligned with the NCTM Focal Point standard *Number and operations*. Similar figures were produced for all easyCBM® math tests in Grades 1 to 8 and the results presented are fairly typical of tests in other Focal Points and/or grade levels. However, space consideration prevents their inclusion in this article.

In sum, the Rasch analyses and figures provide preliminary evidence that the easyCBM® mathematics measures are accessible for low-performing students while being linked to grade-level content standards. Further, the easyCBM® measures provide useful information for potential use within RTI and large-scale testing programs; it is important to note that they were designed with low-performing students in mind from the inception of development. Within an RTI framework, easyCBM® may provide an efficient means for teachers to identify and monitor the progress of low-performing students while providing supplemental data for the identification of students who persistently perform far below grade-level expectations but who are not diagnosed with a severe cognitive disability. Such identified students may be eligible for a large-scale assessment option differing from the general grade-level assessment.

As noted by Perie (2009), the alternate assessment judged against modified achievement standards poses a unique challenge to both educators and test developers. With each testing option for including students with disabilities, both groups are faced with additional complexities. What information is most useful? How does information from various sources converge or diverge? What are the technical characteristics of the tests and measures used to document student performance and identify a testing option? Although some educators and vendors may view more options as an advantage, we assert that a consistent group of students with disabilities needs to be identified through an RTI screener or identified to participate in any of the available large-scale test option. Otherwise, the field becomes "Balkanized" and confusion is likely to exist. At this point, more options are likely to result in more confusion.

We also assert that policy cannot identify this population (U.S. Department of Education, 2006), Rather, it is through empirical studies and the collection of validity evidence such as that which we present in this study, that a group of students can be appropriately defined and sensitive measures developed to not only screeen their participation in an RTI model but also provide an appropriate large-scale test for accountability purposes. Obviously, more research needs to be completed with this group of students as well as across other measures. However, in viewing how students and items perform in a distribution and how items are responded to

according to student ability level, the most basic functioning of the test is being displayed. Other, more traditional metrics of test functions and student performance could easily be used to complement this basic view.

Limitations

All analyses presented came from one benchmark screening administration across two districts in the same state during the fall of 2009. However, numerous equivalent forms of the measures are available for progress monitoring. Given that, as of this writing, easyCBM® registers an average of approximately 3,500 new teacher users nationally every month, future research should examine the item functioning of these equivalent forms. It may be that the functioning changes over time as learning occurs. The sample for the current study came from only two districts in the Pacific Northwest. However, with the exception of the results presented in Figure 3, findings should be generalizable to other populations based on the sample independence in Rasch analyses (Bond & Fox, 2007). Although this analysis examined the functioning of items based on estimated ability level, future research should examine the functioning of items for students with and without *specific* learning disabilities and/or students who have and have not been previously assigned to an alternate assessment for persistent low performers.

CONCLUSION

Despite the limitations of the study, the overall results suggest that CBMs designed with a target population in mind (i.e., persistent low performers) can be reliable and accessible; can adhere to grade-level content standards; and, in the end, provide sufficient validity evidence to guide teacher decision making on participation in RTI and large-scale testing programs. These CBMs could provide educators a powerful measurement tool, which when used in conjunction with RTI, could inform accurate decision making. Further, CBMs developed for persistent low performers may provide additional information for the identification of students in this ambiguous group.

Utilizing RTI and CBMs with sufficient evidence of technical adequacy, educators can document not only students' current level of achievement (benchmark screener) but also their rate of learning (progress monitoring). This is an important point because to be eligible for current, and likely future versions, of a statewide test for persistent low performers the student must be predicted to not pass the state's general assessment "even with the best instruction" (U.S. Department of Education, 2006, para. 12). If the student is receiving quality instruction, but the observed learning (measured through progress monitoring CBMs) is insufficient, he or she may be a prime candidate for an assessment other than the general grade-level assessment. CBMs can thus provide important information to educators in this decision-making process.

To support valid decision making, it is critical that the information used to guide assessment and instruction decisions for persistent low performers include data from reliable and accessible measures linked to grade-level content standards and designed specifically for use with low-performing students. The results of this study provide some preliminary evidence that a measurement tool designed with the target population (i.e., persistent low performers) in mind

from the inception of development may be able to provide data for RTI, persistent low performer identification, and/or alternate state test eligibility decisions.

ACKNOWLEDGMENTS

Funds for the data set used to generate this report come from a federal grant awarded to the UO from the Institute for Education Sciences, U.S. Department of Education: *Assessments Aligned with Grade Level Content Standards and Scaled to Reflect Growth for Students with Disabilities* (Award # R324A070188 funded 2007–2010).

REFERENCES

Anderson, D., Tindal, G., & Alonzo, J. (2009). *Internal consistency of general outcome measures in Grades 1–8* (Tech. Rep. No. 0915). Eugene: University of Oregon, Behavioral Research and Teaching.

Ardoin, S. P., Witt, J. C., Connell, J. E., & Koenig, J. L. (2005). Application of a three-tiered response to intervention model for instructional planning, decision making, and the identification of children in need of services. *Journal of Psychoeducational Assessment, 23*, 362–380.

Bond, T. G., & Fox, C. M. (2007). *Applying the rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Busch, T. W., & Reschly, A. L. (2007). Progress monitoring in reading: Using curriculum based measurement in a response to intervention model. *Assessment for Effective Intervention, 32*, 223–230.

Christ, T. J., Sculin, S., Tolbize, A., & Jiban, C. L. (2008). Implication of recent research: CBM measurement of math computation. *Assessment for Effective Intervention, 33*, 198–205.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219–232.

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*, 184–192.

Deno, S. L., & Mirkin, P. K. (1977). *Data based program modification: A manual*. Minneapolis, MN: Leadership Training Institute for Special Education.

Dolan, R. P., Hall, T. E., Benerjee, M., Chun, E., & Strangman, N. (2005). Applying principles of universal design to test delivery: The effect of computer-based read aloud on test performance of high school students with learning disabilities. *Journal of Technology, Learning, and Assessment, 3*(7). Available from http://escholarship.bc.edu/jtla/vol3/7/

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 105–146). New York, NY: Macmillan.

Filbin, J. (2008). *Lessons from the initial peer review of alternate assessments based on modified achievement standards.* Washington, DC: U.S. Department of Education: Office of Elementary and Secondardary Education—Student Achievement and School Accountability.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review, 33*, 188–192.

Individuals with Disabilities Education Act, 20 U.S.C., Pub. L. No. 108-446 § 1400 *et seq.* (2004).

Johnstone, C. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Tech. Rep. No. 37). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C., Bottsford-Miller, N., & Thompson, S. (2006). *Using the think aloud method (cognitive labs) to evaulate test design for students with disabilities and English language learners* (Tech. Rep. No. 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Johnstone, C., Liu, K., Altman, J., & Thurlow, M. (2007). *Student think aloud reflections on comprehensible and readable assessment items: Perspectives on what does and does not make an item readable* (Tech. Rep. No. 48). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Kettler, R., & Elliott, S. (2009). Introduction to the special issue on alternate assessments based on modified academic achievement standards: New policy, new practices, and persistent challenges. *Peabody Journal of Education, 84*, 467–477.

Kettler, R., Elliott, S., & Beddow, P. (2009). Modifying achievement test items: A theory-guided and data-based approach for better measurement of what students with disabilities know. *Peabody Journal of Education, 84*, 529–551.

Lazarus, S., & Thurlow, M. (2009). The changing landscape of alternate assessments based on modified academic achievement standards: An analysis of early adopters of AA-MASs'. *Peabody Journal of Education, 84*, 496–510.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

No Child Left Behind Act, 20 U.S.C., Pub. L. No. 107-110 § 1424 *et seq.*, 6301 Stat. (2002).

Perie, M. (2009). Introduction. In M. Perie (Ed.), *Considerations for the alternate assessment based on modified achievement standards (AA-MAS): Understanding the eligible population and applying that knowledge to their instruction and assessment* (AA-MAS White Paper; pp. 1–14).

Quenemoen, R. (2009). Identifying students and considering why and whether to assess them with an alternate assessment based on modified achievement standards. In M. Perie (Ed.), *Considerations for the alternate assessment based on modified achievement standards (AA-MAS): Understanding the eligible population and applying that knowledge to their instruction and assessment: A white paper commissioned by the New York Comprehensive Center in collaboration with the New York State Education Department.*

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Shaftel, J., Yang, X., Glasnapp, D., & Poggio, J. (2005). Improving assessment validity for students with disabilities in large-scale assessment programs. *Educational Assessment, 10*, 357–375.

Thompson, S., Johnstone, C., & Thurlow, M. (2002). *Universal design applied to large scale assessments* (Synthesis Rep. No. 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M., Elliott, J., & Ysseldyke, J. (1999). *Out-of-level testing: Pros and cons* (Policy Directions No. 9; Vol. 2010). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

U.S. Department of Education. (2006). *Raising achievement: Alternate assessments for students with learning disabilities.* Retrieved from http://www.ed.gov/policy/elsec/guid/raising/alt-assess-long.html

U.S. Department of Education. (2007). *Final regulations on modified academic achievement standards.* Retrieved from http://www.ed.gov/policy/speced/guid/modachieve-summary.html

U.S. Department of Education. (2008). *Validity evidence for alternate assessments based on modified achievement standards*. Washington, DC: National Technical Advisory Council.