

Vol. 69, No. 4, pp. 481-494.
 ©2003 Council for Exceptional Children.

Alternate Assessments in Reading and Math: Development and Validation for Students With Significant Disabilities

GERALD TINDAL
 MARILEE MCDONALD
 MARICK TEDESCO
 AARON GLASGOW
University of Oregon

PAT ALMOND
Oregon Department of Education

LINDY CRAWFORD
University of Colorado

KEITH HOLLENBECK
Springfield School District, Oregon

ABSTRACT: *Given the mandates of IDEA to include students with disabilities in large-scale assessments, most states have either adopted alternate standards or developed alternate assessments. In either case, it is difficult to understand the students' performance relative to the primary assessment program. And in both cases, the technical adequacy is generally assumed rather than specifically documented. In this study, we developed a series of standardized tasks that can be considered as part of the same construct as operationalized in the primary large-scale assessment program. We then analyzed student performance to ascertain reliability and initial validity. In reading and math, teachers were trained to administer the tasks and judge performance, providing a system with instructional and evaluative uses. The results support the technical adequacy of the alternate assessment.*

The statutory provisions on general state- and districtwide assessments from the IDEA Amendments of 1997 included the provision that "(1) Children with disabilities are included in general State and districtwide assessment programs, with appropri-

ate accommodations and modifications in administration, if necessary; and (2) As appropriate, the State or I.E.A.—(A) develops guidelines for the participation of children with disabilities in alternate assessments for those children who cannot participate in State and districtwide assessment programs; (B) develops alternate assessments; and

(C) beginning not later than July 1, 2000, conducts the alternate assessments" (§300.138; Part B) utilizing statistically sound regular and alternate assessments (§300.139).

While IDEA '97 stipulates participation, the regulations do not provide "specific direction to states about what an alternate assessment is, what it should look like, or how it should be scored or reported, nor does it specify the type or number of alternate assessment participants" (Thompson & Thurlow, 2000, p. 1). According to Thompson and Thurlow, a majority of states' alternate assessments are tied to their academic content standards and are designed as substitutes to the states' standard large-scale tests. Thompson and Thurlow also report that the most common alternate assessment approach uses portfolios reflecting the collection of evidence as functional indicators on the student's progress toward state standards. However, not all states have chosen to use a portfolio-base as their alternate assessment, and this study describes the results from a non-portfolio approach.

A majority of states' alternate assessments are tied to their academic content standards and are designed as substitutes to the states' standard large-scale tests.

First, we describe how one state developed its statewide alternate assessment system and then present results from an initial validation study. This state's alternate statewide assessment system in reading and math is considered an extension of the standard assessment rather than an alternate assessment, primarily because of the tight relationship between it and the standard assessment. The measures are designed to reflect a developmental progression of skills within a continuum of the same construct. The purpose of the system is to produce clear and meaningful data that can be prescriptively used within the individualized education program (IEP) rather than simply descriptively applied to reporting performance as part of an accountability system. It also is to be used in conjunction with the general education assessment system so that performance on one is related to performance on the other.

Utilizing Deno's (1985) propositions, the measures were designed to evaluate student progress, interventions, and programs in the same manner that the statewide assessment is used. Most important, the measures were designed to be systemically valid (see Frederiksen & Collins, 1989) as a direct measure of the constructs of interest and sensitive to the effects of instruction. Students' skills have been expressed in terms of performance on tasks that can provide instructional feedback. Ideally, systemic change in teaching would positively affect performance.

ALTERNATE ASSESSMENT SYSTEMS AND THE CANONS OF VALIDITY

It is important that *all* large-scale assessments, including alternative assessments, reflect the traditional canons of measurement standards, like content validity, concurrent criterion-related validity, and predictive criterion-related validity. Also, measurement systems should reflect Messick's (1988, 1994) four facets of validity: (a) construct validity, (b) values implications, (c) relevance and utility, and (d) social consequences, which refer to both the justification for and the function of the test. Messick's validity criteria are used to characterize the quality of an assessment—whether the test scores accurately reflect the knowledge, skills, and/or abilities the test is intended to measure. "Special validity dispensations" (Messick, 1995, p. 5-8) are unwarranted to legitimize all the various alternate assessment approaches, and there should be no differentiation of validity standards between newer types of assessments and other types of older, traditional tests.

Accordingly, evaluations of all tests need to use similar evidential and consequential validity criteria because validity, reliability, comparability, and fairness are in effect "social values that have meaning and force outside of assessment wherever evaluative judgments and decisions are made" (Messick, 1994, p. 13). Basic to good assessment is the notion that results represent important knowledge and/or capabilities, broader than the specific tasks that happen to be chosen for assessment. Test performance must generalize to a

larger domain of knowledge and/or skills and thus enable us to make accurate inferences about students' capabilities and accomplishments.

An important component of Messick's conception of validity is construct underrepresentation and construct irrelevant variance. The former term refers to assessments that poorly reflect the depth and breadth of knowledge and skills, and the latter term reflects assessments that include the confounding of skills and knowledge other than the construct of interest. In this assessment, we attempt to adequately reflect functional skills of importance while also avoiding the influence of other skills that are not part of the construct. We, therefore, present a series of measures to capture performance on component skills instead of trying to assess complex knowledge or skills that would use a single, complex task. Many students with significant disabilities need carefully constructed tasks that can capture a range of behaviors (Fuchs et al., 1994). In our system, these component skills are assessed using tasks that elicit specific behaviors. We believe that complex performances often reflect construct underrepresentation because few tasks are used to make statements about a student's proficiency in an academic area. We also reason that complex performances might entail construct irrelevant variance, given the confluence of many separate skills needed for any single performance.

In these extended measures, replicability and generalizability also must be considered, not only in determining the boundaries of score meaning as it pertains to the relevance, utility, and social consequences of interpretations, but also as it relates to the consistency of score meaning. Low scores on the extended measures must occur because of the lack of student's competence, not because part of the focal construct is absent or because the assessment "contains something irrelevant that interferes with the affected students' demonstration of competence" (Messick, 1994, p. 21). Finally, the extended measures must be construct-driven rather than task-driven. We define tasks within a universe in which a systematic sampling plan can be used to create alternate forms. In this way, we adopt an assessment system to document student performance and progress with both dimensions oriented toward the same construct.

RESEARCH ON PORTFOLIOS AS AN ALTERNATE ASSESSMENT

As stated earlier, many states have implemented alternate assessment systems that do not directly measure constructs of the standards for which their general education assessment system has been created. And in this process, many states also have relied on portfolio assessments, which have been difficult to implement with any level of technical adequacy. Any assertions proclaimed by portfolio assessment have been tantamount to a claim of construct validity and need to be supported by empirical evidence of construct validity" (Messick, 1994, p. 21). The major problem may be that while portfolios contain relevant products, they typically have ignored the process and therefore have remained ill defined as measurement constructs.

Many states have implemented alternate assessment systems that do not directly measure constructs of the standards for which their general education assessment system has been created.

Research shows that portfolios as part of a large-scale assessment program have posed significant validity problems. Vermont's early experiments with portfolios and Arizona's recent foray into portfolios both ended disastrously. Both states suffered from poor rater reliability problems that prevented the public release of the assessment results (Koretz, McCaffrey, Klein, Bell, & Stecher, 1993).

ONE STATE'S EXTENDED ASSESSMENTS IN READING AND MATHEMATICS

As noted earlier, in this large-scale assessment program, alternate academic assessments are considered an extension of the standard assessment. The genesis of this system is curriculum-based measurement (CBM). The measures require multiple alternate forms using CBM technology drawn on a foundation of time series data. The technical ad-

equacy and psychometric properties of CBM have been extensively represented in the professional literature of experimental investigations (e.g., Bradley-Klug, Shapiro, Lutz, & DuPaul, 1998; Fuchs, Fuchs, Hamlett, & Stecher, 1990; Fuchs et al., 1994; Hintze & Shapiro, 1997; Howell, Fox, & Morehead, 1993; Marston, 1989; Nolet & McLaughlin, 1997; Tindal, 1998; Wesson & King, 1992; Yell, Deno, & Marston, 1992).

According to Deno (1985), CBM is an assemblage of procedures for devising assessments by sampling the domain in question, administering and scoring those assessments, and using the data to help make and evaluate instructional decisions. CBM quantifies student performance in the basic academic skill areas and can be used as an index of student progress over time because it is sensitive to a student's academic growth (Fuchs, Fuchs, Bishop, & Hamlett, 1992; Hartman & Fuller, 1997; Swain & Allinder, 1996). It also can capture both the range and depth of student achievement (Nolet, 1992).

An important component of this study was the development of generalized academic skill indicators that would not be tied to a specific curriculum as reported by many CBM researchers (Bradley-Klug et al., 1998; Hartman & Fuller, 1997; Hintze & Shapiro, 1997; Tindal, Flick, & Cole, 1993). In vindication of general outcome measures using CBM technology, Mehrens and Clarizio (1993) declared that education suffers from a faulty assumption that all assessments tasks need to be sampled directly from some instructional curriculum to be thought of as valid. In the end, we wanted to avoid using measures tied to a specific curriculum, which would limit generalizability and, thus, limit score inferences to the larger domain (Tindal et al., 1993).

EXTENDED MEASURES: SUBTESTS

Based upon CBM research, the extended measures were designed to assess skills that are more unitary in the construct they purport to measure instead of measuring complex, multiskill performances. We developed the tasks to reflect more singular dimensions using the logic of Kane, Crooks, and Cohen (1999): First links are estab-

lished between observations and observed scores (requiring adequate control of administration and scoring procedures); then inferences can be made between the observed score and the universe of scores (providing generalizability); finally, inferences can be made to the range of possible tasks within the same construct (allowing extrapolation).

For us, this specificity avoided construct irrelevant variance. For example, instead of math multiple-choice tests that required reading as well as math skill to perform successfully, the extended measures sampled math skills alone. Or, instead of math open-ended measures that required skill in mathematics *and* writing, the measures assessed math skills only. In both computation and open-ended problems, the measures utilized multiple forms of access (pointing, speaking, and writing). Similar requirements were made of reading and writing measures.

We also wanted to avoid construct underrepresentation by having many different tasks within each of our constructs that sampled different components of each academic skill. Specifically, the reading subtests ranged from naming or pointing to pictures, to blending sounds, to reading sentences, to retelling stories as listening comprehension. The math subtest extended from copying numbers, to telling time, to math concepts like discriminating differences to mixed computations.

METHOD

PARTICIPANTS

The study took place in one state in the Pacific Northwest. Thirty-six school districts within 11 Educational Service Districts from all regions of the state were included.

Teachers. A sample of 131 teachers participated in this study. Teachers were elected to participate by district administrators and regional representatives who were chosen to supervise training, test administration, and data collection throughout the study.

Students. Each teacher assessed 1 to 5 students. A total of 437 students in Grades Kindergarten through postsecondary school were

TABLE 1
Disabilities of Participants

| <i>Primary Disability</i> | n | % |
|-----------------------------------|-----|----|
| Mental retardation | 249 | 57 |
| Autism | 73 | 17 |
| Specific learning disabilities | 16 | 4 |
| Orthopedic impairment | 25 | 6 |
| Other health impairments | 21 | 5 |
| Speech or language impairments | 12 | 3 |
| Visual impairments (or blindness) | 5 | 1 |
| Traumatic brain injury | 3 | 1 |
| Emotional disturbance | 2 | .5 |
| Hearing impairments (or deafness) | 1 | .2 |
| Unknown | 30 | 7 |

selected using the following three criteria: (a) student was exempt from the standard statewide assessment; (b) student had been diagnosed with a moderate to severe disability; and (c) student participated in a functional daily living skills curriculum. Students' average age was 15 years with a minimum of 5 and a maximum of 21.

All 437 students received special education services. Seventy-five of the 437 students were unable to be assessed due to extremely low academic skills, resulting in 112 students in K–3 (26% of the sample), 77 students in Grades 4–6 (18% of the sample), 83 students in Grades 7–8 (19% of the sample), 54 students in Grades 9–10 (12% of the sample), and 41 in Grades 11–12 (9% of the sample); 32 students were postsecondary, and 38 students had no grade reported. Forty-one percent of the students were female and 56% were male. Ethnic make-up included: 80% White, 2% African American, 1% Asian, 1% American Indian, 1% Multi-Racial, and 8% Unknown. See Table 1 for disability information of participating students.

MEASURES

This report includes only data from reading and math, the two areas with multiple-choice tests in the standard assessment program and most dissimilar to the extended measures. From each content domain, we specified tasks that measured knowledge of “basic skills” and applications. We defined a range of tasks that were sensitive to sub-

tle differences in skill levels but also allowed a variety of response options, particularly given the range of disabilities within the population. Hence, items were included that could be administered or responded to using a variety of formats (see Scoring section that follows). We developed tasks using the logic of CBM and as logical prerequisites for success in meeting the standards at the first benchmark. During the process of implementation, new state standards were adopted for students in Grades K–2 and have been used to back-map these measures.

For example, in reading, phonemic awareness is now a standard in Kindergarten with outcomes such as listening to words spoken and identifying the beginning and ending sounds, segmenting single-syllable spoken words, orally blending two to three spoken sounds into recognizable words, and so forth. Other standards address recognition of sight words, and learning, using and understanding new vocabulary. In math, Kindergarten students are now expected to read, write, order, and identify whole numbers less than 100. We have several tasks that assess this standard. Other standards also can be applied (dealing with concepts, money, counting, adding, etc.) to the tasks in this alternate assessment.

We used two pilot studies to guide framework specification and test development, each conducted with students who had moderate to severe cognitive and physical disabilities. The

intent of the first pilot study was to determine an appropriate range of test content, to detail the types and formats of test items, and to identify the number of tasks to be included in each subdomain. The second pilot study was conducted to refine the procedures used for administering and scoring the test. Particular consideration was given to developing a consolidated, yet comprehensive, system of test administration and scoring that facilitated a controlled and minimally distracting test administration environment. Ultimately, these pilot studies provided us with valuable evidence regarding item sets that could maintain their integrity even if alternate administration and response forms were needed due to individual student requirements (needs). The result of the pilot studies was a refinement of a pre-assessment interview and a basic skills assessment consisting of reading, writing, and math components. A discussion of these components follows.

In reading, students progressed through the following tasks: (a) Naming Pictures, (b) Naming or Pointing to Letters, (c) Blending Sounds, (d) Reading Words, (e) Reading Names, (f) Reading Sentences, (g) Reading Text Orally, and (h) Retelling Stories as Reading or Listening Comprehension. In mathematics, students progressed through the following tasks: (a) Naming Numbers, (b) Pointing to Numbers, (c) Copying Numbers, (d) Counting on Dictation, (e) Writing Numerals, (f) Writing Numerals, (g) Naming Shapes, (h) Telling Time, (i) Naming and Counting Money, (j) Manipulation of Objects using Math Concepts, and (k) Computation. In both content areas, the skills are anchored to newly developed state standards established for students in Grades K–2.

PROCEDURES

All of the administration and scoring procedures were included in a testing kit that was developed by the authors. The tasks were administered so that the most basic skill was presented first, with each successive task presented on the basis of increasing difficulty or complexity.

Administration. The test materials included all of the items required for both the administration and scoring of the test and were compiled into a three-ring binder. In addition, the folders were arranged by domain, with each

area containing a document with administration directions, administration scripts, a representation of each student task, scoring instructions, and student record sheets. The folder also contained testing materials as well as student response forms. In addition, the folders contained standardized procedures that provided the test administrator with rules for determining basal and ceiling levels as well as options for documenting format changes.

Scoring. We provided flexibility in scoring to avoid irrelevant variance from two sources. First, teachers could avoid administering a task if they believed the student could successfully complete all items in the task: They simply marked the task NA-P for Not Administered-Proficient. This feature allowed teachers to focus their assessment on the key level of difficulty most relevant to the student, a particularly important aspect to consider when testing students who often exhibit other interfering behaviors. Second, teachers could avoid administering tasks that were inappropriate, such as a blending task to a student who is deaf. In our sequence of tasks, students could be administered those that appeared later and reflected more complex performances revealing advanced skill and receive a higher score without exhibiting previous component performance reflecting individual skills. In this example, a student who is deaf could read passages and retell stories to receive a very high score without ever having taken previous component skill tasks.

Each task was scored both quantitatively and qualitatively. The quantitative scoring allowed for small units of progress to be documented (i.e., letters and sounds correct in reading or correct digits in math). For each subdomain, the quantitative scores were then transformed into qualitative scores to reflect a global estimate of performance. Students in a graduate special education program were trained to review the protocols and make a global judgment about proficiency (see Table 2).

TRAINING

Our goal for each training session was to ensure that participants were extremely familiar with test materials, proficient in their administration, and competent in scoring the academic behav-

TABLE 2
Example of the Extended Reading and Math Scoring Sheet

Each task is located on a 1-6 rubric with the number of points possible in parentheses following the task name. Place the student on the rubric where the most points exist.

| Task # | Skill | Rubric | Task # | Skill | Rubric |
|--|----------------------|----------|--|--|-----------------|
| Task 1. Naming Pictures Pointing To Pictures (16) | Picture Naming | 0.5 1 | Task 1. Naming Numbers (16) Pointing To Numbers (8) Task 2. Copying Numbers (24) Task 5. Naming Shapes Pointing To Shapes (8) Task 4. Writing Numerals (18) | Rote numbers with a stimulus | 0.5 1 1.5 |
| Task 2. Naming Letters Pointing To Letters (16) | Letter Names | 1.5 | Task 3. Counting On Dictation (16) Task 8a-b. Manipulating Math Concepts (6) | Sequence numbers in a pattern | 2 2.5 |
| Task 3. Reading Words Pointing To Words (16) | High Frequency Words | 2 | | | |
| Task 4. Phonemic Segmentation (22) | Letter Sounds | 2.5 | Task 6. Telling Time (12) Task 7a. Naming Coins and Bills (8) Pointing To Coins and Bills (8) Task 7b. Counting Money (12) | Numbers in context | 3 3.5 |
| Task 5. Blending Sounds (22) | Blending Sounds | 3 | | | |
| Task 6. Reading Names (9) | Names | 3.5 | Task 8c-d. Manipulating with Math Concepts (6) | Concrete operations with manipulatives | 4 4.5 |
| Task 7 Reading Sentences (14) | Sentence Reading | 4 | | | |
| Task 8. Reading Text Orally (open with CWPM) | Oral Reading Fluency | 4.5 5 | Task 9. Timed Computation (open) a. Addition Facts b. Subtraction Facts c. Multiplication Facts | Symbolic operations (timed math facts) | 5 5.5 |
| Task 10. Retelling Stories: Listening Comprehension Level 1 = 22, Level 2 = 44, Level 3 = 66 | Listening Comprehend | 5.5 | Task 10. Mixed Computation a. Adding b. Subtracting (18) c. Multiplying d. Dividing (18) | Symbolic operations: adding, subtracting, multiplying and dividing | 6 |
| Task 9. Retelling Stories: Reading Comprehension Level 1 = 22, Level 2 = 44, Level 3 = 66 | Reading Comprehend | 6 | | | |

iors assessed. The agenda for each training session was arranged according to academic skill areas, with the Reading, Writing, and Math tests introduced and modeled separately. We first provided a thorough description of test materials, then showed a videotape to model test administration and scoring, and finally allocated time for teachers to practice administering the test to each other while trainers observed. A sample notebook of testing materials was provided to teachers. This hands-on approach ensured teachers' familiarity with all of the materials.

We believe teachers left the training well prepared. Total time for each training session was approximately 6 hr with some variations in time depending on the size of each group, the questions asked during training, and participants' prior knowledge related to administration of basic skills tests. The videotape provided a clear model of how test conditions or test format might be altered to capture as much academic behavior as possible from each participating student. Four different students, each with varying types of disability and levels of proficiency were shown on

the videotape, as well as four different test administrators, each with his or her own testing style. Finally, practicing test administration with a peer afforded teachers the opportunity to become comfortable with the flashcards and manipulatives included in the student packets, as well as learn how to manipulate test materials with one hand while scoring student responses with the other hand.

We conducted four training sessions encompassing geographically diverse regions. All training was conducted in March, with sessions being attended by groups ranging from 12 to 42 teachers. Regional representatives who had been trained prior to our workshops were available to answer any questions that teachers had after their training, when they were actually administering and scoring tests in their buildings. They were also responsible for ensuring that test materials were completed and returned to the research site.

RESULTS

Performance was analyzed separately for reading and math academic skill areas in two ways: (a) using quantitative indices of performance for subtasks, and (b) using summary judgments of overall proficiency (see Table 2 for reading and math rating sheets respectively). The total group of students in reading was 362 and 359 students in math.

READING

In reading, the same groups of students were analyzed to ascertain initial validity. The largest group were students with mental retardation ($n = 249$), followed by 73 students with autism, 25 students with orthopedic impairments, 21 students with other health impairments, and 16 students with specific learning disabilities. The lowest performance was for students with autism, then students with mental retardation, and finally, the highest reading performance was for students with specific learning disabilities. The differences between students with mental retardation and with autism were less distinguishable in the early tasks than they were in the later ones. With naming letters, blending sounds, and reading words, both groups of stu-

dents appeared to have at least some minimal proficiency. However, in reading contiguous text, more group differences appeared in their performance. In the most developed tasks, which involved reading text and relating the meaning of the story, students from all three groups were relatively poor performers. In many of the individual subtasks, considerable variation can be seen, reflecting the fact that some students perform very well and others very poorly. The standard deviation frequently is greater than the mean (see Table 3).

Trained students with master's degrees (8 total students with two pairs each in reading and math) provided individual judgments of each student's overall performance, taking into account the range of tasks completed and the levels of performance on each one. All students had been part of one of our regional training. The reliability of ratings was quite high: Fully half of all the judgments were in complete agreement and the other half were within one-half point. Only 17 judgments of the 362 exceeded one point, with 161 being exact matches, 162 being off by one-half point, and 21 off by one point. For each of the ratings, the number of students scoring at that level is reported. The actual distribution of scores (for both judges) was bimodal and platykurtic with a large number and percentage of students rated 2–5. Pairs of judges rated a large number of students low on the scale as well as high. The highest level of performance reflecting the most complex construct (comprehension) was notably low with few students scoring near the top end of the scale (see Table 4).

MATH

We tested a total of 359 students in math, with the majority of the students having mental retardation ($n = 249$), students with autism ($n = 73$), and students with learning disabilities ($n = 16$). The other two prominent groups were students with orthopedic impairment and other health impairments. While students with other disabilities were tested, we report only on these three groups in our initial validation.

We found students distributed themselves differently according to their primary disability. Students with autism performed the lowest, followed by students with mental retardation, and

TABLE 3
Reading for Students With Mental Retardation (MR), Autism, and Learning Disabilities (LD)

| <i>Task and Number of Items</i> | M | SD | M | SD | M | SD |
|--|-------|-------|-------|-------|--------|-------|
| Task 1: Naming Letters (16) | 11.30 | 6.49 | 9.89 | 7.14 | 15.88 | 0.50 |
| Task 1: Pointing to Letters (8) | 1.31 | 2.66 | 1.63 | 3.08 | 1.00 | 2.73 |
| Task 2: Naming Pictures (16) | 11.31 | 5.39 | 8.18 | 6.42 | 14.00 | 2.13 |
| Task 2: Pointing to Pictures (8) | 1.27 | 2.60 | 1.88 | 3.00 | 0.00 | 0.00 |
| Task 3: Blending Sounds (22) | 10.60 | 9.18 | 7.66 | 9.17 | 20.00 | 3.14 |
| Task 4: Words (16) | 6.82 | 6.97 | 5.80 | 6.80 | 14.88 | 2.53 |
| Task 4: Pointing to Words (8) | 1.02 | 2.20 | 1.43 | 2.49 | 0.00 | 0.00 |
| Task 5: Names (min. of 9) | 6.07 | 6.90 | 5.48 | 6.92 | 10.50 | 5.72 |
| Task 5: Names (min. of 9) | 14.11 | 6.51 | 12.61 | 7.25 | 13.19 | 4.37 |
| Task 6: Sentences (14) | 5.97 | 5.98 | 5.04 | 6.10 | 12.94 | 3.00 |
| Task 7: Text Orally, Kevin's Story (cw) | 9.55 | 23.57 | 6.27 | 16.21 | 32.50 | 26.69 |
| Task 7: Read Text, Kevin's Story (iw) | 1.91 | 5.09 | 1.18 | 3.37 | 2.81 | 4.94 |
| Task 7: Read Text, Sue & Peg (cw) | 10.00 | 29.31 | 8.16 | 22.54 | 15.69 | 38.44 |
| Task 7: Read Text, Sue & Peg (iw) | 1.22 | 11.17 | 1.16 | 8.02 | 0.38 | 1.02 |
| Task 7: Read Text, TV (cw) | 11.89 | 36.01 | 5.32 | 21.84 | 11.06 | 30.26 |
| Task 7: Read Text, TV (iw) | 0.37 | 1.60 | 0.21 | 0.99 | 0.00 | 0.00 |
| Task 8: Read Comprehend, Kevin's story | 1.44 | 3.49 | 1.00 | 3.26 | 6.25 | 5.59 |
| Task 8: Read Comprehend, Sue & Peg | 1.70 | 4.85 | 1.59 | 4.84 | 4.19 | 8.29 |
| Task 8: Read Comprehension, TV | 0.91 | 3.32 | 0.27 | 1.55 | 1.25 | 2.98 |
| Task 9: Listen Comprehend, Kevin's story | 2.11 | 3.62 | 1.23 | 2.79 | 0.00 | 0.00 |
| Task 9: Listen Comprehend, Sue & Peg | 1.25 | 4.13 | 0.45 | 2.04 | 3.94 | 8.71 |
| Task 9: Listen Comprehend, TV | 0.36 | 1.65 | 0.48 | 2.90 | 0.00 | 0.00 |
| Test total points student earned | 63.44 | 39.93 | 52.13 | 41.25 | 104.81 | 24.01 |
| Average of Qualitative Score—Judge 1 | 2.67 | 1.66 | 2.28 | 1.72 | 4.63 | 0.77 |
| Average of Qualitative Score—Judge 2 | 2.86 | 1.74 | 2.40 | 1.79 | 4.75 | 0.88 |

Note: cw=correctly read words, iw=incorrectly read words
MR ($n=249$), Autism ($n=73$), LD ($n=16$)

then those with specific learning disabilities performed the highest. Students with mental retardation completed about half the items correctly, whereas those with specific learning disabilities were completely correct in their responses in all tasks except the computation problems, an area in which students from all disabilities performed quite poorly. Students with autism performed not that much differently than those with mental retardation in the early tasks (naming, copying, counting, and writing numerals) but did differ in most of the later tasks. Again, a large amount of variation is present, reflecting considerable diversity among the students on many of the tasks (see Table 5). As in reading, for many of the tasks the standard deviation is greater than the mean.

The data on the qualitative judgments reflected consistency in overall value as well as in the distribution of scores. Of the 359 ratings, 95 were exact matches, another 135 were within one-half point, 89 were within one point, and 31 were within one and a half points. Only 9 of the total sample were 2 or more points apart. For each of the ratings, the number of students scoring at that level also is reported. For both judges, the distributions were slightly negatively skewed with a large group of students achieving values of nearly 5 points and exceeding 5 points. With the last task involving symbolic problems, however, a large drop-off occurred with few students performing at that level (see Table 6).

TABLE 4
Score Distributions for Two Judges in Evaluating Reading

| <i>Judge 1 Rating</i> | <i>Count</i> | <i>Judge 2 Rating</i> | <i>Count</i> |
|-----------------------|--------------|-----------------------|--------------|
| 0.0 | 27 | 0.0 | 18 |
| 0.5 | 35 | 0.5 | 26 |
| 1.0 | 23 | 1.0 | 35 |
| 1.5 | 40 | 1.5 | 45 |
| 2.0 | 40 | 2.0 | 35 |
| 2.5 | 29 | 2.5 | 27 |
| 3.0 | 23 | 3.0 | 23 |
| 3.5 | 23 | 3.5 | 22 |
| 4.0 | 25 | 4.0 | 17 |
| 4.5 | 32 | 4.5 | 35 |
| 5.0 | 40 | 5.0 | 26 |
| 5.5 | 18 | 5.5 | 40 |
| 6.0 | 1 | 6.0 | 5 |
| Missing | 6 | Missing | 8 |
| Total | 362 | Total | 362 |

TABLE 5
Math for Students with Mental Retardation (MR), Autism, and Learning Disabilities (LD)

| <i>Task and Number of Items</i> | <i>M</i> | <i>SD</i> | <i>M</i> | <i>SD</i> | <i>M</i> | <i>SD</i> |
|---|----------|-----------|----------|-----------|----------|-----------|
| Task 1: Naming Numbers (16) | 12.35 | 6.29 | 10.68 | 7.27 | 16.00 | 0.00 |
| Task 1: Pointing to Numbers (8) | 1.13 | 2.59 | 1.57 | 3.03 | 0.00 | 0.00 |
| Task 2: Copying Numbers (24) | 16.24 | 10.40 | 12.54 | 11.12 | 23.56 | 0.73 |
| Task 3: Counting on Dictation (16) | 11.34 | 6.11 | 9.55 | 7.08 | 16.00 | 0.00 |
| Task 4: Writing Numerals (18) | 11.07 | 8.03 | 9.09 | 8.38 | 17.69 | 0.70 |
| Task 5: Naming Shapes (8) | 4.69 | 2.21 | 4.21 | 2.72 | 6.31 | 1.20 |
| Task 6: Telling Time (12) | 4.16 | 4.35 | 3.23 | 3.95 | 8.81 | 3.58 |
| Task 7a: Naming Coins (8) | 4.80 | 3.14 | 3.75 | 3.42 | 7.94 | 0.25 |
| Task 7a: Pointing to Coins (4) | 0.58 | 1.25 | 0.64 | 1.33 | 0.25 | 1.00 |
| Task 7b: Counting Money (12) | 3.33 | 4.56 | 2.02 | 3.80 | 9.88 | 2.55 |
| Tasks 8ah: Manipulate Math Concepts (6) | 4.13 | 2.22 | 3.21 | 2.51 | 6.00 | 0.00 |
| Tasks 8cd: Manipulate Math Concepts (6) | 2.40 | 1.80 | 1.84 | 1.95 | 4.13 | 1.45 |
| Task 9a: Computation (22) | 8.79 | 9.25 | 5.59 | 8.07 | 20.44 | 2.71 |
| Task 9b: Computation (16) | 2.89 | 5.21 | 1.32 | 3.63 | 10.56 | 5.39 |
| Test total points student earned | 87.90 | 51.18 | 69.25 | 51.13 | 147.56 | 13.42 |
| Test total points possible | 155.06 | 38.73 | 149.50 | 47.62 | 164.25 | 1.00 |
| Test percent correct | 0.53 | 0.31 | 0.42 | 0.31 | 0.90 | 0.08 |
| Average of Qualitative Score—Judge 5 | 3.02 | 1.58 | 2.50 | 1.61 | 5.19 | 0.68 |
| Average of Qualitative Score—Judge 6 | 3.57 | 1.67 | 2.98 | 1.89 | 5.53 | 0.39 |

Note: MR (n=249), Autism (n=73), LD (n=16)

DISCUSSION

The main findings from this study focus on scaling performance on reading and math tasks that is consistent with disability and establishing reliability of rating performance on a qualitative scale. We found students distributed themselves on our performance tasks in roughly the manner one would expect given their disability. For example, students with mental retardation performed less well than students with learning disabilities. This finding may reflect the fact that students with mental retardation spend much less time in academically focused classroom settings than students with learning disabilities. Probably a more important lesson learned is that many students with significant disabilities are learning the symbol systems of our language. Therefore, their instructional programs need to provide them more direct instruction of these skills and our measurement systems need to reflect performance and progress in their learning.

We have argued for the use of well-standardized tasks that (a) fit the constructs of basic skills (reading and math in this article), (b) are aligned with the standards used to form the large-scale assessments, and (c) have documented technical adequacy. The initial data we have reported focus primarily upon reliability and construct validity: Students performed on a set of progressively difficult tasks to the point that it was clear their performance could generalize to a larger domain and eventually be connected to the standards themselves. We focused on two dimensions of construct validity in particular: construct underrepresentation and construct irrelevant variance (Messick, 1994).

Using a progression of tasks, we attempted to sample complex behaviors that did not underrepresent the construct; and by using well-designed sampling plans and task formats, we attempted to avoid many other nuisance influences on performance. Finally, using the logic of Kane et al. (1999) in constructing tasks, we wanted to generalize from observations and observed scores to a universe of scores. With partial scoring and trained testers, we arranged the tasks into a progression, allowing us to infer the level of a student's skill within a construct like reading or math, providing us a means to extrapolate. All of

these components are likely important in making inferences from observed scores to generalizations and inferences in targeted measurement domains used to document performance proficiencies.

Our approach contrasts with the traditional alternate assessments based on portfolios in structuring the domain and the format of the assessment administration and scoring. Given the reliability problems reported by Koretz et al. (1993), we focused on products, trying to minimize the problems with process. Furthermore, in the development of the tasks, as well as their administration and scoring, we emphasized explicitness, resulting in relatively consistent judgments in performance across raters. It is likely that the relatively high reliability of our outcomes is because of this explicitness. At the same time, a range of administration formats was available while controlling the content that was sampled. For example, in reading, students unable to pronounce the letter or word were allowed to point to it (the letter or word) after it had been read to them. For a student with a speech-language impediment, this flexibility allowed us to document whether or not the student "understood" the grapheme-phoneme relationship.

In the end, we argue for the assessment of students with significant disabilities so that performance *and* progress can be documented on tasks reflecting construct validity. However, the assessments must be related to the standards, not underrepresent the skill, and reflect minimal irrelevant variance. Although it makes little sense to administer the standard test, the standards themselves are important for them. Students with significant disabilities are learning to read and compute. They, therefore, should be given appropriate assessments that can document this performance. Furthermore, this assessment should be sensitive to their instructional level, so that a range of tasks are present allowing them to perform now as well as "grow" into more complex performances later, as they receive more instruction. This range of tasks should allow teachers to sample their performance formatively so they can see them progress. We believe the assessments we have described in this article provide the means for accomplishing this outcome and can be used to complement large-scale assessment systems.

TABLE 6
Score Distributions for Two Judges in Evaluating Math

| <i>Judge 5 Rating</i> | <i>Count</i> | <i>Judge 6 Rating</i> | <i>Count</i> |
|-----------------------|--------------|-----------------------|--------------|
| 0.0 | 25 | 0.0 | 25 |
| 0.5 | 16 | 0.5 | 16 |
| 1.0 | 28 | 1.0 | 12 |
| 1.5 | 18 | 1.5 | 9 |
| 2.0 | 28 | 2.0 | 21 |
| 2.5 | 33 | 2.5 | 20 |
| 3.0 | 42 | 3.0 | 39 |
| 3.5 | 23 | 3.5 | 26 |
| 4.0 | 38 | 4.0 | 36 |
| 4.5 | 45 | 4.5 | 30 |
| 5.0 | 33 | 5.0 | 59 |
| 5.5 | 18 | 5.5 | 45 |
| 6.0 | 9 | 6.0 | 18 |
| Blank | 1 | Blank | 1 |
| Missing | 2 | Missing | 2 |
| Total | 359 | Total | 359 |

In the end, we argue for the assessment of students with significant disabilities so that performance and progress can be documented on tasks reflecting construct validity.

IMPLICATIONS FOR PRACTICE

Three distinct issues become prominent in the use of this assessment system and converge. First, the measurement system combines a number of brief measures that allow teachers to sample a range of behaviors, which should allow teachers to document student performance in many of the component skills in that academic area. Ideally, they would be able to use this information to focus their IEPs, providing them with useful information in documenting present levels of performance as well as in establishing long-range goals and short-term objectives. At the same time, with a holistic rating, teachers can maintain the assessment system for a long enough time to provide a developmental trend. When students perform near the top of this scale, it is clear that they

should begin participating in the standard assessment program, with possible accommodations or modifications.

Second, this system allows teachers to maintain a standardized focus on relevant skills and ensures that their students participate in the large-scale assessment programs. It also requires them, however, to devise their own measurement systems for formatively evaluating instructional effects. Hopefully, teachers would use similar formats for devising their own measures. Nevertheless, whether using these tasks at benchmark levels or using alternate forms of their own creation, teachers are working with a uniform set of constructs that are highly related to performance on state tests and other published achievement measures. This allows teachers to collect information that is predictive of performance, from their own measures to these alternate measures or from these alternate measures to the state standard measures.

Finally, while we have focused on documenting student performance, the most critical issue is training teachers in how to administer the measures and use the outcomes. We have not addressed this issue in this study but are currently

working on the development of systems for professional development. Whether we incorporate this training into our preservice programs or embed it in staff development for practicing teachers, any successful use of the system depends on qualified teachers. Specific strategies need to be available for helping teachers understand how to ensure student's performance is well documented, to use the information diagnostically in building instructional programs, and to evaluate and report on student outcomes.

These three issues (a part-whole focus on behavior, multiple measures that establish relationships, and training for decision making) all converge to determine success at the systems level. The practical implications are that large-scale assessment programs need to explicitly address all three issues to successfully support teachers. Although we have reported on one component, it is clear that in this state, much more development needs to take place for the system to effectively serve teachers and students. A long-term focus is needed, therefore, in which state departments allocate their resources for such supports. Then the field can more appropriately devote attention not just to simple participation of students with significant disabilities in large-scale testing programs, but to the improvement of their proficiencies and development of effective programs.

REFERENCES

- Bradley-Klug, K. L., Shapiro, E. S., Lutz, J. G., & DuPaul, G. J. (1998). Evaluation of oral reading rate as a curriculum-based measure within literature-based curriculum. *Journal of School Psychology, 36*, 183-197.
- Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 52*, 219-232.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.
- Fuchs, L. S., Fuchs, D., Bishop, N., & Hamlett, C. L. (1992). Classwide decision-making strategies with curriculum-based measurement. *Diagnostique, 18*(1), 39-52.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecher, P. M. (1990). The role of skills analysis in curriculum-based measurement in math. *School Psychology Review, 19*(1), 6-22.
- Fuchs, L. S., Fuchs, D., Hamlett, C. L., Thompson, A., Roberts, P. H., Kubek, P., & Stecher, P. M. (1994). Technical features of a mathematics concepts and applications curriculum-based measurement system. *Diagnostique, 19*(4), 23-49.
- Hartman, J. M., & Fuller, M. L. (1997). The development of curriculum-based measurement norms in literature-based classrooms. *Journal of School Psychology, 35*, 377-389.
- Hintze, J. M., & Shapiro, E. S. (1997). Curriculum-based measurement and literature-based reading: Is curriculum-based measurement meeting the needs of changing reading curricula? *Journal of School Psychology, 35*, 351-375.
- Howell, K. W., Fox, S. L., & Morehead, M. K. (1993). *Curriculum-based assessment: Teaching and decision making* (2nd ed.). Pacific Grove, CA: Brooks/Cole.
- Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C. §1400 et. seq. (ERIC Document Reproduction Service No. ED 412 721)
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*(2), 5-17.
- Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). *The reliability of scores from the 1992 Vermont portfolio assessment program*. (CSE Technical Report 355.) Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing.
- Marston, D. B. (1989). A curriculum-based measurement approach to assessment: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford.
- Mehrens, W. A., & Clarizio, H. F. (1993). Curriculum-based measurement: Consequential and psychometric considerations. *Psychology in the Schools, 30*, 241-254.
- Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 89-103). Hillsdale, NJ: Erlbaum.
- Messick, S. (1994). The interplay of evidence and consequences in the validity of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice*, 14(4), 5-8.

Noler, V. (1992). Classroom-based measurement and portfolio assessment. *Diagnostique*, 18(1), 5-26.

Noler, V., & McLaughlin, M. (1997). Using CBM to explore a consequential basis for the validity of a statewide performance assessment. *Diagnostique*, 22(3), 146-163.

Swain, K. D., & Allinder, R. M. (1996). The effects of repeated reading on two types of CBM: Computer maze and oral reading with second-grade students with learning disabilities. *Diagnostique*, 21(2), 51-66.

Thompson, S., & Thurlow, M. (2000). State alternate assessments: Status of IDEA alternate assessment requirements take effect (Synthesis Report 35). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Tindal, G. (1998). *Models for understanding the task comparability in accommodated testing*. Washington, DC: Council of Chief State School Officers.

Tindal, G., Flick, D., & Cole, C. (1993). The effect of curriculum on inferences of reading performance and improvement. *Diagnostique*, 18(1), 69-84.

Wesson, C. L., & King, R. P. (1992). The role of curriculum-based measurement in portfolio assessment. *Diagnostique*, 18(1), 27-37.

Yell, M. L., Deno, S. L., & Marston, D. B. (1992). Curriculum-based measures as formative evaluation. *Diagnostique*, 18(1), 99-112.

Address all correspondence to Gerald Tindal, Behavioral Research and Teaching, 241 College of Education, University of Oregon, Eugene, OR 97403-5262.

E-mail: geraldtr@darkwing.uoregon.edu

Preparation of this document was supported in part by the Office of Special Education and Rehabilitative Services (OSEP), grant award number H324D000063-01. Opinions expressed herein do not necessarily reflect the position or policy of OSEP, and no official endorsement by OSEP should be inferred.

Manuscript received May 2002; accepted October 2002.

BooksNow

* To order books referenced in this journal, please call 24 hrs/365 days: 1-800-BOOKS-NOW (266-5766); or 1-732-728-1040; or visit them on the Web at <http://www.BooksNow.com/ExceptionalChildren.htm>. Use Visa, M/C, AMEX, Discover, or send check or money order + \$4.95 S&H (\$2.50 each add'l item) to: Clicksmart, 400 Morris Avenue, Long Branch, NJ 07740; 1-732-728-1040 or FAX 1-732-728-7080.

ABOUT THE AUTHORS

GERALD TINDAL (CEC #20), Professor, College of Education, University of Oregon, Eugene. **PAT ALMOND** (CEC #20), Evaluation Specialist, Oregon Department of Education, Salem. **MARILEE MCDONALD**, Research Assistant, Behavioral Research and Teaching, University of Oregon, Eugene. **LINDY CRAWFORD** (CEC #1111), Assistant Professor, University of Colorado at Colorado Springs. **MARICK TEDESCO** (CEC Oregon Federation), Research Associate; **AARON GLASGOW**, Technology Consultant, University of Oregon, Eugene. **KEITH HOLLENBECK** (CEC #1111), Director of Programs, Springfield School District, Oregon.
