

Scaling Early Reading Alternate Assessments With Statewide Measures

PAUL YOYANOFF

GERALD TINDAL

University of Oregon

ABSTRACT: *Alternatives to the standard statewide assessment often are necessary for valid measurement of students with significant disabilities. These alternate assessments must be carefully developed and evaluated with respect to generally accepted psychometric standards. Ideally, these measures should be sensitive to growth and scaled to state tests that are aligned with grade level content standards. In this study, we investigated the technical adequacy of early reading performance tasks for a Grade 3 statewide alternate reading test. Our research hypotheses pertain to (a) equitable measurement function across populations of students, (b) measurement dimensionality, and (c) item difficulty. Using test results obtained from annual testing in one state, our findings suggest that performance tasks are adequate for use in the alternate assessment, and they can be scaled in terms of the statewide testing metric.*

Development of alternate assessments took a sharp turn in direction with the new regulations that announced the possibility of states having alternate achievement standards (U.S. Department of Education, 2003). With that announcement came further clarity on the alignment of assessments with either grade level standards or the new alternate achievement standards, either of which could be used for accountability purposes. It is now possible to deem as proficient up to 1% of students with significant disabilities for purposes of reporting Adequate Yearly Progress (U.S. Department of Education). In either case, the federal government is clear that alternate

assessments must be aligned with grade level standards and that they must be technically adequate.

Our research tests hypotheses pertaining to the technical adequacy and vertical *scale alignment* of Oregon's early reading alternate assessment with the general education grade level assessments. Vertical scale alignment of the alternate assessment with the general education assessment allows educators to make valid inferences from obtained scores, specifically score interpretations related to growth and comparison to benchmark standards. Our research hypotheses underlying scaling alternate assessments pertain to (a) measurement dimensionality, (b) measurement-psychometric model fit, (c) performance task difficulty relative to statewide assessment items, and (d) performance task bias with respect to examinee popula-

tions. With this research we argue that carefully selected and scaled performance tasks will function as alternatives to the statewide assessment for students with significant disabilities.

BRIEF HISTORY OF ALTERNATE ASSESSMENTS

The use of alternate assessments has a short, compelling history with respect to their use in the context of large-scale statewide testing. Passage of the 1997 Individuals With Disabilities Education Act first introduced the need for alternate assessments so that students with significant disabilities could participate in large-scale testing programs. Alternate assessments at that time were quite ill defined and diverse in both focus and format; needless to say, the empirical support for them was and continues to be debatable (Ysseldyke & Olsen, 1997).

Browder et al. (2004) tracked the shift in assessment focus for students with significant disabilities over the past 30 years. They noted four major phases of intervention research and its impact on assessment. Initially (in the late 1960s), programs were aligned with infant and early childhood developmental theories using age-based norms. The second phase in the late 1970s focused on functional curricula with four major domains: vocational, community, recreational, and school. A third phase appeared in the 1990s, combining this functional approach (and complementing community and school access) with more school-based tasks to address self-determination. Finally, in the most recent and sweeping change ushered in with the Individuals With Disabilities Education Act of 1997 (IDEA), the fourth phase was clearly oriented toward academic standards.

Since implementation of IDEA, three academic assessment response formats have been used for alternate assessment: portfolios, observations, and performance assessments. These constructed response formats have been under psychometric evaluation (Bennett, 1993; Messick, 1996; Thissen, Wainer, & Wang, 1994; Traub & Fisher, 1977), and if used carefully, they appear to have promise as part of the statewide alternate assessment program (Bennett & Ward, 1993; Linn, Baker, & Dunbar, 1991; Robinson, 1993;

Thissen & Wainer, 2001). With *portfolios*, teachers collected student work samples during the year, usually selected according to some general criteria. Often, teachers are provided with both a book of state standards from which to choose relevant benchmarks and a book of directions on how to create a portfolio (what to include and how to present the information). *Observations* also are used in several states. Teachers are asked to select a behavior that represents a student need and then observe it in a functional environment. *Performance* assessments are used in only a few states and reflect the kind of measurement most similar to the traditional testing program. Basically, a series of tasks are administered to the student and scored in terms of correctness. Tindal (2005) provides technical perspective on these procedures as they relate to measurement of students with disability.

The critical issue to resolve is the process for translating standards and tests designed for use in general education to something that is appropriate for students with significant disabilities. This translation requires a perspective on function, along with measurement and procedural considerations. However, as Ford, Davern, and Schnorr (2001) have aptly described,

the options selected by states in their efforts to develop alternative performance indicators, reflect one of two approaches: (a) simplify the regular standard until we can find something (anything!) that the student can do, or (b) redefine the regular curriculum standard so that it represents some type of functional skill. (pp. 213–214)

NEED FOR TECHNICAL ADEQUACY AND EMPIRICAL SUPPORT

With the focus of alternate assessments on grade level standards as part of No Child Left Behind (NCLB), neither option of simplifying or redefining is likely to provide a technically adequate alternate assessment. NCLB legislation is essentially about high standards for ALL students, and to comply, state education agencies (SEAs) must align alternate assessments to state grade level content and performance (achievement) standards. Furthermore, they must be technically adequate

although few, if any, models exist for investigating the technical adequacy of these assessments. The Federal Register, in the Rules and Regulations, clarifies the requirement as follows:

The requirements for high technical quality set forth in §§200.2(b) and 200.3(a)(1), including validity, reliability, accessibility, objectivity, and consistency with nationally recognized professional and technical standards apply to alternate assessments as well as to regular State assessments. (Rules and Regulations, 68, p. 68609. Fed. Reg. 236, December 9, 2003)

Yet, very few research studies have been published on the technical adequacy of alternate assessments; the following represents nearly the entirety of recent work.

Most of the alternate assessment validation research has been completed through survey research procedures. For example, three leading studies report on participant perceptions. Kleinhert and Kearns (1999) reported validation of the Kentucky state alternate assessment on the basis of survey response from 44 content experts (of 80 surveys sent). Among their obtained findings, respondents perceived most of the state's alternate assessment indicators and outcomes as important to students, although they expressed concerns about the narrowness of a domain-based approach reflecting a curricular model. Kleinhert, Kennedy, and Kearns (1999) received 331 surveys (of 508 mailed out) asking teachers to respond to seven issues about Kentucky's alternate assessment (importance, benefit, use, application, schedules, progress, and participation). On most of these issues, nearly half of the teachers agreed or strongly agreed about the impact of the assessment, though considerable concern was raised about the time spent in completing portfolios and the focus on teacher evaluation. Finally, Kampfer, Horvath, Kleinhert, and Kearns (2001) reported on 206 teachers' (of 400 sent out) perceptions of time and effort associated with a number of variables relevant to the operation of the alternate assessment: eligibility, materials sent out, schedules, entries, progress, social relationships, access to multiple settings, and development of natural supports. They reported that teachers spend 25 to 35 hr outside of instructional time to complete the port-

folios. Furthermore, only modest relationships were obtained between outcome performance and a number of these operational variables.

Research on technical adequacy issues and score reliability of alternate assessments has not progressed beyond an initial documentation of conventional indices and criterion validity. For example, Kleinhert, Kearns, and Kennedy (1997) reported on the reliability (at that time) of their "alternate portfolio" approach over 3 years, with agreement indices of 57%, 49%, and 49% achieved on a 4-point scale. They also reported a correlation of .45 between a measure of program quality and performance on the alternate portfolio (a finding also reported by Turner, Baldwin, Kleinhert, & Kearns; 2000). Likewise, Tindal et al. (2003) reported on the reliability of scores and learning performance levels in reading and mathematics for approximately 350 students with significant disabilities participating in a state alternate assessment. Both quantitative scores and qualitative judgments were reported with about 67% within 1 point on a 5-point scale. Tindal, Glasgow, Gall, VanLoo, and Chow (2002) submitted a technical adequacy analysis summary to the Oregon Department of Education in which data were reported on the reliability of reading, writing, and mathematics performance assessments for students with significant disabilities participating in the 2001 state testing program. Both format and content changes were reported as well as test-retest reliabilities (typically reported in ranges that exceeded .90).

In summary, as states (re)define and align their assessment systems in relation to grade level content and performance standards, the research to date inadequately addresses emerging psychometric considerations. Relevant measurement research requires rigorous examination of measurement alignment along with conventional psychometric characteristics. Two types of measurement alignment must be addressed as part of the validation process. First, assessment alignment to grade level content is imperative, whether using systems like those promulgated by Webb (1999) and adapted by Tindal (2005) or through vertical alignment and scaling as proposed by Wise (Townsend et al., 2004). A second form of alignment pertains to the measurement scale. Validation of score interpretation rests on assumptions

linking the alternate assessment scores to the general assessment score metric. In this study, we approach measurement scale alignment using item response theory, testing measurement hypotheses about the alternate assessment and the general state assessment as a unified testing program.

SCALING PERFORMANCE TASKS AND ITEM RESPONSE MODELING

Problems with scaling items from two independent tests onto a common metric are not new (Kolen, 1981; Kolen & Brennan, 1995). Historically, using classical test equating methods, examinee performance has been estimated in terms of total scores from multiple measures, which are then equated. More recently, in the context of item response theory (IRT), item responses rather than total scores have been identified as more informative for ability level inferences (Lord, 1980; Stocking & Lord, 1983; Thissen & Wainer, 2001). IRT procedures are more flexible with respect to measurement construction, equating, and scoring. For instance, using item-based procedures enables an expanded array of measurement equating/linking designs.

LINKING RESEARCH DESIGN

Often discussed in the context of test equating (Kolen & Brennan, 1995), linking is a procedure for placing two or more measures on a common metric. Once linked, persons taking one or the other of the assessments can be compared with respect to proficiency on the measurement dimension. Linking procedures require a research design (Kim & Cohen, 1998; Kolen & Brennan; Muraki, Hombo, & Yong-Won Lee, 2000; Vale, 1986). There are numerous considerations when designing a linking study. An appropriate measurement model describing the response process must be identified, for example the one parameter logistic Rasch model (see Equation 1). The dimensionality of items composing the assessments must be specified in the model. Usually, assessments are assumed to be constructed of items sampled from one dimension only, that is, unidimensional. Also, students are sampled such that their responses enable generalization of item characteristics across important comparison groups. Frequently, re-

searchers will use a linking design that requires examinees to complete multiple measurement forms that are being linked (common person), or they may use a common item design in which an overlapping set of items are found in both measurement forms. Using the correct design, it is possible to estimate the measurement scale alignment of measurements from two or more different assessments. Based on data obtained from our research design (described in the following text), we hypothesize that (a) the early reading alternative performance tasks and the statewide test items are unidimensional, (b) responses to the performance assessment tasks and the statewide test items are adequately fitted by the Rasch model, and (c) the alternative performance tasks are unbiased with respect to disability.

UNIDIMENSIONALITY AND THE BI-FACTOR MODEL

Thissen et al. (1994) asked a question that is very relevant to our research. Are tests comprised of both multiple-choice and free-response items less unidimensional than multiple-choice tests? In other words, does the measurement construct and what is being assessed depend on item format, or does content sufficiently characterize the construct? The specific hypothesis being tested pertains to the underlying factor structure associated with the free-response items (the performance tasks) and the multiple-choice response items (the state assessment items). They found that one dominant trait (essential unidimensionality; Nandakumar, 1993) rarely, if ever, characterizes educational measurements. This has an important implication in our research because the IRT model we use assumes a unidimensional data structure. If there is more than one measurement dimension, and we ignore it in our measurement model, then the estimates of item characteristics and person abilities may be unreliable and seriously misleading. In this case, our observations may be due to an ability that we are neglecting to consider.

The confirmatory bi-factor model (Gibbons & Hedeker, 1992) is useful for testing the unidimensionality of an item set. It specifically hypothesizes that each item loads on two factors, a *general* factor, and a *specific* factor associated with

the item's response format. For the data obtained in our research, we considered a general factor and the two specific factors—constructed response and selected response. The fundamental idea here is to compare two models and see which one best describes these data. Under the assumption of unidimensionality, a one-factor model is compared to the bi-factor model in terms of log-likelihood fit statistics. The difference between the models is used to test if the bi-factor is an improved description of the observed responses. If the item responses are unidimensional, then there will be little improvement by using a bi-factor.

THE RASCH MEASUREMENT MODEL

In this section we introduce the IRT model we think describes performance on the alternate and the statewide assessments. Expressing an essential IRT principle, Equation 1 is the IRT one parameter logistic (1PL) model estimating the probability of a correct response of student s to a given item i , where P is governed by the student ability θ_s and item difficulty β_i .

$$P(X_{is} = 1 | \theta_s, \beta_i) = \frac{e^{(\theta_s - \beta_i)}}{1 + e^{(\theta_s - \beta_i)}} \quad \text{Equation 1} \quad (\text{where } i = 1, 2, \dots, n)$$

The simple response process implied by Equation 1 can be described in terms of (a) examinee ability and (b) item difficulty. Given the student's ability, as an item becomes more difficult, the probability of a correct response diminishes. From another perspective, given an item's difficulty, as the student's ability increases, the probability of a correct response increases. Estimating the probability of a specific response is based on a comparison of the person ability and the item characteristic(s).

Equation 1 is often identified as the Rasch measurement model, named after George Rasch (Masters & Wright, 1984; Rasch, 1960). The Rasch model is perhaps the most simplistic of all IRT models. Its simplicity is due in large part to how restrictive it is. According to the Rasch model, each item has a difficulty parameter, β_i , only. Items are constrained to be equally discriminating and with equal probability of correct guessing. It is worth noting here that there are many other IRT models in current use (Boomsma, van Duijn, & Snijders, 2001; van der Linden & Hambleton, 1997). For instance, there

are models that include item discrimination, item guessing, and time to respond, among numerous other item characteristics. Also, there are models for polytomously scored items (items scored in more than two categories), such as the popular rating scale, partial credit, and graded response models (Embretson & Reise, 2000). For our research, we hypothesize that the Rasch model for dichotomously scored responses (Equation 1) accurately describes the generating process underlying responses to both the performance tasks and the statewide items.

ITEM DIFFICULTY

The IRT approach to measurement explicitly identifies and estimates item characteristics. The Rasch model we described earlier contains one item parameter—item difficulty. Briefly, it is important to clarify the meaning of the item difficulty parameter. In IRT, for items scored incorrect/correct, item difficulty refers to that point on the ability scale where a correct item response becomes more likely than an incorrect response. This allows for the location of each item on the ability scale. Some items will be easy (located at the low ability end of the scale) and other items will be difficult (located at the high ability end of the scale). For our research, we hypothesize that the early reading alternative assessment tasks will be relatively easy compared to the general statewide assessment items. In this sense, the alternate assessment “extends” the measurement scale, enabling the accurate measurement of low ability performance.

DIFFERENTIAL ITEM FUNCTIONING

If test users plan to use a measurement system with two distinct populations, special education and general education, then it is imperative that our measurement is unbiased with respect to population membership. Inferences regarding relative ability assume that the measures provide unbiased scores. In the current research we are interested specifically in the performance tasks and whether or not the item response generating processes are invariant across special education and general education populations. Ideally, the measurement system should function identically irrespective of whether a student has a disability. Stated another

way, an individual's scale score should depend on ability, only, and in no way should it be affected by group membership (Holland & Thayer, 1988; Holland & Wainer, 1993). Differential item functioning (DIF) occurs when item scale values depend on group membership. In many contexts, occurrence of DIF is explained in terms of measurement dimensionality (Gierl, 2005).

Using a common item nonequivalent groups research design (Kolen & Brennan, 1995; Vale, 1986), we collected data for the purpose of better understanding the scale alignment of Oregon's early reading alternate assessment with the general statewide first benchmark reading assessment. The design samples data from two hypothetically nonequivalent groups—special education students, and general education students. Scaling the alternative performance tasks onto the statewide measurement scale involved administration of the alternative performance tasks to both groups. Responses from 163 third grade special education and 840 second grade general education students were analyzed to test four hypotheses:

1. The 6 performance tasks and the 25 statewide items measure a unidimensional reading ability construct and that item format (constructed response, selected response) is a trivial factor.
2. Responses to the 6 performance tasks combined with 25 state assessment items are adequately explained by a Rasch measurement model.
3. The 6 alternative performance tasks extend the lower measurement range of the 25 multiple-choice items.
4. The 6 alternative performance tasks function without bias across special and general education populations.

METHOD

First, we tested the hypothesis that the constructed response performance tasks combine with the selected response statewide items to form a unidimensional measure. We compared a one-factor model to the bi-factor model (Gibbons & Hedeker, 1992), hypothesizing that the bi-factor model offers little additional explanation of the item correlations beyond that obtained with the

one-factor model. In effect, this constitutes a test of the assumption that the data are unidimensional. Next, dichotomized responses to the reading performance tasks and the statewide items are hypothesized to be adequately explained by the Rasch model, and furthermore, the alternative performance tasks are hypothesized to be less difficult than the statewide assessment items. This is the essence of the extended alternate assessment.

Finally, we tested for DIF under the hypothesis that our performance assessment tasks function without bias for special education and general education students. Our underlying hypothesis is that the performance on the performance assessment tasks is governed only by ability and is not affected by factors associated with the general/special education classification. First, as part of a criterion validation analysis, we tested the expectation that general education students will have higher performance than students with disabilities on the performance tasks. We expected the tasks to discriminate performance, but once ability is controlled for, we hypothesize no differential functioning. Then, using IRT-based DIF detection procedures provided by the Rasch analysis software, WINSTEPS (Linacre, 2005), the DIF effects were tested. This has implications for dimensionality, which is an important assumption underlying the Rasch one-dimensional model used to describe these data.

SAMPLE

In the spring of 2001, a total of 840 second grade general education students and 163 third grade special education students were sampled for participation in this study. General education students were sampled from the second grade because we wanted low skill when observing response to the alternate assessment performance tasks. The Grade 2 general education students completed a special edition of the Grade 3 statewide test and the early reading performance tasks. Both assessments were aligned with Grade 3 benchmark 1 reading standards. The special education students completed the alternate assessment only. Table 1 describes the two samples and how their data were used for these analyses. The Grade 3 special education student data were essential for studying the differential item function-

TABLE 1*Sampling of Students by Test for Item Analyses*

<i>Student Sample</i>	<i>6 Alternate</i>	
	<i>Assessment Performance Tasks</i>	<i>25 Statewide Assessment Items</i>
163 Grade 3 special education students	<ul style="list-style-type: none"> • Calibrated for DIF analysis • Test success on performance tasks compared to general education participants 	<ul style="list-style-type: none"> • Not Administered
840 Grade 2 general education students	<ul style="list-style-type: none"> • Single group equating to state assessment • Calibrated for DIF Analysis • Test of unidimensionality • Test of 1 PL Model Fit 	<ul style="list-style-type: none"> • Pre-calibrated anchor items for single group equating to performance tasks

ing of the alternate assessment performance tasks. The general education student data were useful for scaling, dimensionality analyses, and differential item function analysis.

The data were obtained as part of an annual testing program. The original naturalistic sample included 1,136 Grade 2 general education students and 650 Grade 3 special education students. A decision was made to use only complete data records, (i.e., student data that included responses to all items on the tests administered). Note that the special education students responded to the alternate assessment only. Although it is unfortunate to eliminate data records because of missing data, the missing responses were not known to be missing at random and were therefore deemed non-ignorable (Little & Schenker, 1995). Data records were analyzed only if responses were provided to all administered items.

The general education student data were obtained from two neighboring school districts. One district was a relatively large, middle socioeconomic status (SES), urban district with approximately 17,000 students overall, of which 7,000 were elementary-school age. The other district was a smaller, lower SES school district with approximately 5,000 students, of which 1,500 were elementary-school age. In the large urban district, teachers were solicited through the central office and those who responded affirmatively were provided schedules and materials for testing. For the smaller school district, all elementary school teachers participated.

Of the 840 second grade general education students, 48% were male, 52% were female; 3.2%

were American Indian, 3.7% were Asian, 5.8% were Black, 9.3% were Latino, and 78.0% were White. All general education students were in second grade and primarily participating in general education programs, although a few students with learning disabilities being served in mainstream classrooms also took part in the testing.

The 163 third grade special education students were tested with the alternate assessment that included the six performance tasks analyzed in this study. The majority of special education students had serious cognitive disability (82%), and the remaining students (18%) had speech/language and learning disabilities. All students were being served in their respective building. Of the special education students, 37% were male and 63% were female; 2.1% were American Indian, 3.7% were Asian, 5.8% were Black, 2.8% were Latino, and 78% were White. Approximately, 7% of the special education students were bilingual and/or used English as a second language.

MEASURES

Our early reading test data were obtained with two measures—(a) tasks being prepared for the Oregon Benchmark 1 alternate assessment and (b) a special edition of the Oregon Benchmark 1 general assessment. Six early reading alternative performance assessment tasks were developed for consideration as part of the Oregon alternate assessment. The tasks sampled behavior in the following early reading skill areas: (a) reading words, (b) blending sounds, (c) reading names, (d) reading sentences, (e) easy oral reading fluency

(ORF), and (f) difficult oral reading fluency. The tasks were carefully constructed based on results from the U.S. National Reading Panel study (2000). The general assessment was a 25-item special edition of the Benchmark 1 (Grade 3) multiple choice reading assessment. The alternate assessment was administered to the Grade 3 special education sample. Also, the alternate assessment was administered to the Grade 2 general education sample concurrent with the special edition of the Grade 3 statewide reading assessment.

Partial credit scoring is usually applied to the alternative performance tasks. To simplify presentation of findings, we dichotomized the scores. Table 2 describes the alternate assessment tasks and the dichotomization scoring rules for each performance task. Dichotomous scoring of the first four tasks was based on the idea that students should have sufficient skill to perform the task. However, recognizing the possibility that a student with acceptable proficiency may miss an item because of a careless error and to compensate for measurement error, the dichotomization decision rule was set 2 points below the total possible score. The frequency distributions suggested that this rule was appropriate. The modal score for each item was the highest score possible, with a high percentage of students obtaining the second and third highest scores. Below this maximum level, the percentages were dispersed uniformly. Dichotomous scoring of the oral reading fluency measures was based on performance standards estimated in oral reading fluency research (Hassbrouck & Tindal, 1992). Also, when modeling performance, several different scoring options were considered, including partial credit models. Irrespective of the scoring rules, similar results were obtained. For purposes of interpretation, use, and presentation, we chose to use the dichotomous scoring presented in Table 2.

The Grade 2 special edition of the statewide multiple-choice test had 25 items sampled from the first benchmark (third grade level). These items are relatively easy compared to items from other Grade 3 tests. The multiple-choice test consisted of a series of extended paragraphs with four or five questions and four response options each. This reading test samples items from six different areas in reading: (a) literary forms, (b) word meaning, (c) literal comprehension, (d) differen-

tial comprehension, (e) evaluative comprehension, and (f) literary device. Approximately three to five items are included in each area. The statewide test is usually scaled using a one-parameter IRT model.

ANALYSIS

Three analyses were completed as part of our hypothesis testing. First, a test of unidimensionality was completed with confirmatory factoring procedures using TESTFACT software (Wilson, Wood, & Gibbons, 2003). Specifically, a single factor model was compared to a bi-factor model (Gibbons & Hedeker, 1992), trying to understand if the format-specific factors explained any additional item variance. The one-factor model estimated the factor loadings and explained variance for a single common factor. The bi-factor model estimated a general factor in addition to two format specific factors. Under the assumption that there are no format-specific factors, we expected the bi-factor model to offer no improvement over the one-factor model.

Our second hypothesis was that the Rasch model for dichotomously scored responses fit the data, and equally important, we hypothesized that the alternate assessment tasks would calibrate to be less difficult than items on the statewide assessment. Using WINSTEPS (Linacre, 2005), we fitted the data with the Rasch model for dichotomously scored responses. As part of these analyses we observed the relative difficulty of the performance tasks and the standardized statewide measurement items.

Finally, we tested whether the performance tasks function differently for special education and general education students. This was considered from two perspectives. First, we expected the performance tasks to discriminate performance. The general education students, although in second grade, were apt to perform higher in reading than the third grade special education students. Second, we did not expect the performance tasks to exhibit differential item functioning (DIF) with respect to the special education and general education samples. We hypothesized that the tasks function equivalently for the two samples. Our two analyses include (a) testing group differences in performance on the performance tasks

TABLE 2*Description of the Six Reading Performance Tasks and Dichotomous Scoring*

<i>Performance Task</i>	<i>Description</i>	<i>Dichotomous Scoring</i>
Task 1. Reading words	Say to the student, "Read each word as I show you the flashcard." Continue presenting words. Prompt the student after 3 s if no response.	Range 0 to 16 0 through 14 = 0, 15 through high = 1
Task 2. Blending sounds	Say to the student, "I will show you a card with a word on it. Say ALL of the sounds in the word. Watch me." Show student the example flashcard. Say to the student, "This word is cuuuuut." Emphasize the process of sounding out. As you read the word, point to each letter with your finger just under the word and slide your finger from left to right.	Range 0 to 22 0 through 19 = 0, 20 through high = 1
Task 3. Reading names	Place the flashcards of proper names in a stack in front of the student in the order shown in the table below. Say to the student, "Read each name as I show you the flashcard." Continue presenting names. Prompt the student after 3 s if no response.	Range 0 to 9 0 through 6 = 0, 7 through high = 1
Task 4. Reading sentences	Place the flashcards of sentences in a stack in front of the student in the order shown in the table below. Say to the student, "Read each sentence as I show you the flashcard." Continue presenting sentences. Prompt the student after 3 s if no response.	Range 0 to 14 0 through 12 = 0, 13 through high = 1
Task 5. Easy oral reading fluency	<i>This is a timed task.</i> Place the story on the table in front of the student. Say to the student, "When I say begin, start reading aloud at the top of the page (point). Read across the page (point). Try to read each word. If you come to a word you don't know, I'll tell it to you. Be sure to do your best reading. Do you have any questions? (pause) Begin."	Range 0 to + ∞ 0 through 52 = 0, 51 through high = 1
Task 6. Difficult oral reading fluency	<i>This is a timed task.</i> Place the story on the table in front of the student. Say to the student, "When I say begin, start reading aloud at the top of the page (point). Read across the page (point). Try to read each word. If you come to a word you don't know, I'll tell it to you. Be sure to do your best reading. Do you have any questions? (pause) Begin."	Range 0 to + ∞ 0 through 52 = 0, 51 through high = 1

and (b) testing differences in Rasch performance task difficulty calibrations for each group.

RESULTS

Our research hypothesized that the constructed-response performance tasks and the selected-response statewide assessment items formed a unidimensional measure generated by a process that could be described by a unidimensional Rasch model. By definition of an extended alternate assessment, we further hypothesized that as part of the alternate assessment, the performance tasks would have Rasch difficulty scale scores

lower than those of the statewide items. Essentially, this implies that the alternate assessment is functional for estimating ability levels below those possible with the standardized statewide assessment. Finally, we wanted to know that the alternate assessment tasks were sensitive to ability only, and not other factors that may differentiate the special education and general education populations, for example, ability to interact with the performance task materials.

Results of scaling the early reading alternate assessment with the statewide general assessment are consistent with our hypotheses. The measures appear to be sufficiently unidimensional, the Rasch model fits the data adequately, the early

reading alternative performance tasks are scaled to be less difficult than the majority of general assessment items. Finally, the alternative performance tasks function comparably for both the sample of students with disability and the general student sample. Overall, these results suggest that performance tasks are technically appropriate for use in the alternate assessment, resulting in a more accurate measurement of lower ability on the general statewide assessment scale. Described in the following, Tables 3, 4, 5, and 6 provide the results of our dimensionality, scaling, and differential item functioning analyses, respectively.

DIMENSIONALITY

The test of dimensionality compared a single-factor to bi-factor model. The bi-factor model contains a general factor along with two item format-specific factors—constructed response and selected response. We tested the difference between these two models. If there is not a format-specific effect, then the two models should explain similar amounts of variance in the observed responses. Most important, the single-factor loadings and the bi-factor general factor loadings should be comparable.

The factor loadings and model fit statistics reported in Table 3 suggest that the alternate assessment constructed response tasks and the statewide assessment selected response items do measure a single, dominant dimension. Both the one-factor model and the general factor from the bi-factor model are comparably strong, explaining 40.64% and 39.62%, respectively, of the item variance. Furthermore, the item loadings are very similar for the one factor (one-factor model) and general factor (bi-factor model). In addition to a comparable strong general single factor, the bi-factor model fit does result in nontrivial unique factors associated with the constructed response and the selected response item formats. Although the unique factors are not strong, the bi-factor model is a statistical improvement over the one-factor model as shown by the chi-square change statistic reported at the bottom of Table 3. However, equally noteworthy here and consistent with our hypothesis of unidimensionality are the factor loadings of the one factor and general factor.

RASCH MODEL FIT AND ESTIMATED ITEM DIFFICULTIES

The Rasch dichotomous model appears to fit the data very well. As indicated in Table 4, the average fit statistic is 1.11. Desirable fit statistics for this model, when estimated with WINSTEPS (Linacre, 2005), center around the value 1 (Smith, 2000). Also, evident in Table 4 and Figure 1, the alternate assessment tasks are less difficult than the statewide assessment items. The difficulty values are locations on the state assessment RIT scale where an item has a probability of 0.50 of being responded to correctly. Starting with the least difficult item (bottom of Table 4) and increasing in item difficulties are the following tasks and items: (a) reading words, (b) reading sentences, (c) blending sounds, (d) easy fluency, (e) statewide item 1, (f) reading sentences, and finally, (g) reading names, the most difficult alternative performance task. Of all responses, performance tasks and statewide items combined, clearly the easy items tend to be the alternate assessment tasks, with the exception of statewide Item 1. These results support the hypothesis that the alternate assessment does extend the lower end of the read ability scale. Figure 1 displays the item characteristic curves for the performance tasks and the least difficult statewide test items, indicating the item locations of the measurement scale. Note, all statewide assessment items not displayed in Figure 1 are more difficult than the alternate assessment tasks.

DIFFERENTIAL ITEM FUNCTIONING

Finally, the DIF analyses indicate that (a) the performance tasks are sensitive to the ability differences of the special education and the general education samples and (b) there is little differential functioning, that is, item functioning with ability held constant. The results of tests of performance differences on the performance tasks, comparing the special education to general education students, are provided in Table 5. Clearly, the Grade 2 general education students perform significantly higher than the Grade 3 special education students. This is expected, though the extent to which this was obtained is surprising given the grade level differences. We include these data with the DIF analyses as criterion validation of the per-

TABLE 3*One-Factor and Bi-Factor Analysis of 6 CBM Tasks and 25 OSA Items*

Item	One-Factor	General	Bi-factor	
			CBM	OSA
Task 1. Reading words	.66	.74	.19	—
Task 2. Blending sounds	.41	.41	.49	—
Task 3. Reading names	.55	.60	.54	—
Task 4. Reading sentences	.26	.27	.32	—
Task 5. Easy oral reading fluency	.84	.90	.23	—
Task 6. Difficult oral reading fluency	.89	.97	.19	—
Statewide test item 1	.43	.42	—	.21
Statewide test item 2	.38	.44	—	.02
Statewide test item 3	.38	.35	—	.24
Statewide test item 4	.56	.54	—	.26
Statewide test item 5	.55	.49	—	.32
Statewide test item 6	.50	.41	—	.42
Statewide test item 7	.50	.46	—	.31
Statewide test item 8	.81	.68	—	.56
Statewide test item 9	.73	.59	—	.58
Statewide test item 10	.79	.69	—	.50
Statewide test item 11	.70	.62	—	.45
Statewide test item 12	.71	.66	—	.35
Statewide test item 13	.76	.68	—	.47
Statewide test item 14	.51	.48	—	.26
Statewide test item 15	.56	.55	—	.24
Statewide test item 16	.64	.58	—	.34
Statewide test item 17	.59	.61	—	.20
Statewide test item 18	.77	.80	—	.16
Statewide test item 19	.69	.74	—	.15
Statewide test item 20	.46	.54	—	.03
Statewide test item 21	.69	.71	—	.23
Statewide test item 22	.78	.79	—	.25
Statewide test item 23	.77	.80	—	.18
Statewide test item 24	.77	.77	—	.26
Statewide test item 25	.51	.56	—	.09
% of variance	40.64	39.62	2.47	8.33
reliability	0.84		0.75	
χ^2	8828.12		8506.49	
<i>df</i>	718		687	
χ^2 change			321.63	
<i>df</i> change			31	
probability			0.00	

Note. CBM = curriculum-based measurement; OSA = Oregon Statewide Assessment Program.

formance tasks and as a way to help clarify the meaning of DIF.

The DIF results are presented in Table 6. Item DIF pertains to how an item functions for students with comparable abilities, but who are

from different groups (e.g., general education and special education). In these analyses the general education sample is the *reference group* and the special education sample is the *focal group*. Though we know that the general education stu-

TABLE 4

Item Calibrations for Dichotomous One-Parameter Logistic Model in Order of Difficulty (items listed in order of Rasch calibration high to low)

<i>Item</i>	<i>Number Correct of 933</i>	<i>RIT Scale Calibration</i>	<i>Mean Square Fit</i>	<i>Item Score to Measure Correlation</i>
Statewide assessment item 25	382	201.85	1.19	.53
Statewide assessment item 19	512	201.66	.85	.66
Statewide assessment item 20	391	201.13	1.23	.52
Statewide assessment item 21	489	200.92	.86	.66
Statewide assessment item 12	560	197.33	.84	.67
Statewide assessment item 11	561	194.98	.85	.68
Statewide assessment item 5	514	194.71	1.12	.60
Statewide assessment item 14	573	193.61	1.11	.59
Statewide assessment item 16	568	193.17	1.01	.63
Statewide assessment item 15	647	191.99	.90	.65
Statewide assessment item 2	508	188.91	1.66	.52
Statewide assessment item 3	544	188.66	1.47	.55
Statewide assessment item 6	613	188.47	1.13	.61
Statewide assessment item 17	671	187.97	.84	.67
Statewide assessment item 13	665	187.07	.63	.74
Statewide assessment item 10	678	187.02	.66	.73
Statewide assessment item 18	665	185.88	.71	.72
Statewide assessment item 4	615	184.42	1.18	.64
Statewide assessment item 24	650	183.58	.83	.72
Statewide assessment item 9	700	182.80	.69	.71
Statewide assessment item 22	680	182.10	.71	.73
Statewide assessment item 23	691	181.47	.69	.73
Statewide assessment item 8	720	181.21	.57	.74
Statewide assessment item 7	651	179.90	1.28	.63
Task 3. Reading names	698	175.98	1.48	.54
Task 6. Difficult oral reading fluency	706	175.05	1.09	.64
Statewide assessment item 1	741	170.32	.99	.65
Task 5. Easy oral reading fluency	752	168.25	1.07	.61
Task 2. Blending sounds	778	163.49	1.18	.56
Task 4. Reading sentences	801	158.73	1.61	.39
Task 1. Reading words	826	152.28	1.72	.33

Note. The average fit statistic equals 1.11; RIT = Rasch unit.

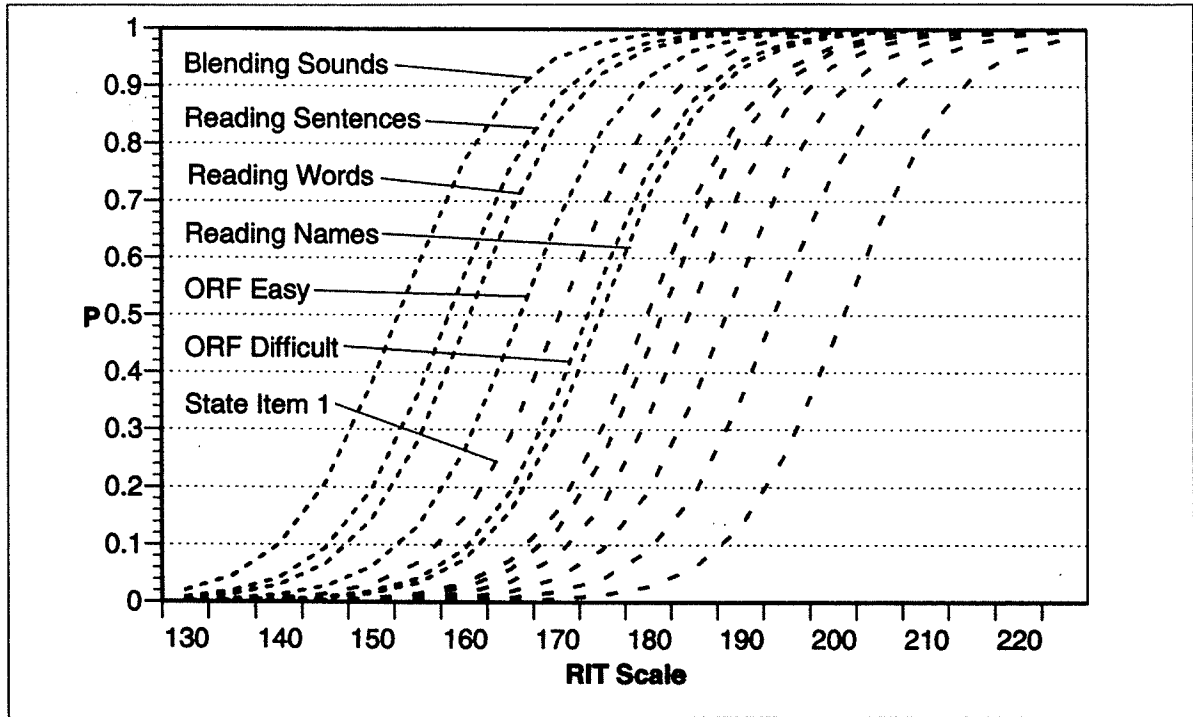
dents, on average, score higher than the special education students on the performance tasks (Table 5), when controlling for these differences, we expect the item difficulties to be the same for both groups. We tested the differences in Rasch estimated item difficulties, as reported in Table 6. The results are very encouraging. Only for one performance task, reading sentences, do we observe DIF. For the special education students and general education students the item difficulty is -2.14 and $-.81$, respectively. The item is significantly less difficult for the special education students, when we

control for ability. As Gierl (2005) points out, explaining DIF can be very challenging. DIF is probably a function of multidimensionality in the measurement, and identifying the *construct irrelevant component* requires considerable thought and possibly additional research.

DISCUSSION

Our research on scaling an early reading alternate assessment tests hypotheses underlying the psy-

FIGURE 1
Reading One-Parameter Logistic Item Characteristic Curves



chometric potential for carefully designed performance assessments in the context of large-scale, high-stakes testing. The results are particularly important for educators concerned with young students and students with low skills who may not be able to perform on the statewide tests. For students from this population who can complete the general statewide test, estimated scale scores

may be invalid. Many of the items composing the general assessment are not informative for students with low ability. If valid scale scores for this population of students are desired, then technically adequate alternate assessments are necessary.

As hypothesized, we found that performance tasks can be used to extend the Oregon benchmark on statewide assessment in reading. We used

TABLE 5
Comparison of Special Education and General Education Students Passing Each of Six Performance Tasks by Sample

Performance Task	Sample				$\chi^2_{(df=1)}$
	Special Education		General Education		
	N	%	N	%	
Task 1. Reading words	87	53.4	729	86.8	100.46*
Task 2. Blending sounds	105	64.4	745	88.7	62.22*
Task 3. Reading names	36	22.1	607	72.3	149.36*
Task 4. Reading sentences	99	60.7	630	75.0	13.99*
Task 5. Oral reading fluency (easy)	86	52.8	743	88.5	121.28*
Task 6. Oral reading fluency (difficult)	71	43.6	705	83.9	127.06*

* $p \leq 0.01$

TABLE 6*Summary of Differential Item Analysis of Performance Tasks*

Performance Task	Sample		DIF Contrast	S.E.	t (df = 347)
	Special Education	General Education			
Task 1. Reading words	-0.12	-0.52	0.40	0.27	1.45
Task 2. Blending sounds	-1.52	-1.52	0.00	0.33	0.01
Task 3. Reading names	1.67	1.48	0.19	0.27	0.72
Task 4. Reading sentences	-2.14	-0.81	-1.34	0.35	-3.86*
Task 5. Oral reading fluency (easy)	0.60	0.21	0.39	0.26	1.50
Task 6. Oral reading fluency (difficult)	1.18	1.27	-0.09	0.26	-0.35

* $p \leq 0.01$

six well-controlled early reading performance tasks that had been highlighted as very important by the U.S. National Reading Panel (2000) and previously validated by a number of different researchers. The reading performance tasks combined with statewide test items tend to be essentially unidimensional, though multidimensional item format factors (constructed response and selected response) are somewhat apparent. Using IRT, specifically the Rasch model for dichotomously scored items, we scaled performance on these tasks in relation to proficiency on the Oregon statewide testing metric. Testing for model fit and performance task difficulty, our hypotheses regarding the Rasch model and the relative ease of the tasks were confirmed. In addition to being less difficult, for our early reading alternate assessment we were able to confirm that most tasks function without bias regarding special and general education populations. Collectively, these results support our argument that the alternate assessment tasks can be used to more accurately scale low ability performance interpretable in terms of the statewide assessment scale.

In addition to being less difficult, for our early reading alternate assessment we were able to confirm that most tasks function without bias regarding special and general education populations.

Our findings aside, a unique measurement challenge emerges relative to the standards estab-

lished by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (1999). Whereas there is a call for technical adequacy of the alternate assessment, what is the context of psychometric viability, and how can measurement specialists best conceptualize these issues (Tindal and Haladyna, 2002)? Previous literature on performance tasks, particularly curriculum-based measurement (CBM) suggests that task variance is not influenced to a great extent by the specific task sampling (e.g. Hintz, Owen, Shapiro, & Daly, 2000; Hintz & Petite, 2001). On the contrary, we emphasize that careful task sampling cannot be overstated. Furthermore, we have taken a next step toward interpretation of early reading test performance by situating the alternate assessment in the context of large-scale testing. Having scaled performance in terms of the statewide measurement system, alternate assessment performance is informative for understanding how students with serious disability are performing in terms of the statewide metric. In this manner, we can focus on making inferences to a larger behavioral construct. In this process, then, we begin to build common measures among uncommon tests (Koretz, Bertenthal, & Green, 1999).

Our findings offer a promising option for the challenge educators of early school-age children currently face. First, there is a way to better conceptualize the equitable measurement of students with significant disability without compromising the performance comparison with other student populations (e.g., general education students). Second, the design and data collection for prop-

erly sampling performance tasks was based on the U.S. National Reading Panel's (2000) identification of critical reading skills, and careful alignment with state Benchmark 1 content standards, and the annual statewide assessment for general education students. Paying close attention to articulated skill competencies and test content at the statewide level does provide clear direction for task design and sampling. The use of performance tasks in the context of alternate assessments is defensible.

We now better understand some of the requisite analytic needs for developing our technically adequate alternate assessments. There are some fundamental psychometric procedures that can provide evidence supporting use of the measures constructed for special education populations and low ability levels. First, our measures should be unidimensional. This is not a novel idea, though new factor models and analytic strategies are developing constantly. Second, there is value in thinking hard about the generating process underlying our obtained responses and carefully fitting measurement models in the spirit of hypothesis testing. Again, this is not new, but our models and software are improving. Finally, understanding measurement bias and how group membership sometimes interacts with item psychometric quality is an exciting aspect of establishing technical adequacy for our alternate assessments.

IMPLICATIONS FOR THE FIELD

These results are consistent with widely recognized measurement needs pertaining to high-stakes alternate assessments (Tindal & Haladyna, 2002), and yet they are inconsistent with some conclusions of the National Research Council. Properly constructed performance tasks can be aligned not only with respect to content, but performance scale, too. This psychometric feature is as important as commonly reported reliability and criterion validity statistics. Our research indicates that it is possible to combine constructed response items with standardized selected response items, at least for the purpose of extending the measure range and appropriately testing low performing populations.

We are not proposing that the two item types be combined in single standardized test adminis-

tration. Protocols for administration of alternate assessments do prescribe careful consideration of who is administered which tests, applying appropriate content and performance standards. The essence of our research results confirm our hopes that alternate assessments, when properly constructed and administered, can be scaled onto a metric defined by a statewide assessment.

Regarding the use of performance tasks, they are often considered appropriate primarily for measuring higher-order skills (Bennett & Ward, 1993; Messick, 1996). To the contrary, in this research the performance tasks can be developed to measure low-level reading skills. When compared to standardized, multiple-choice items, they measure more directly the *latent skills* presumed to be at the core of early reading. However, measurement validity and reliability do not come easy. It is not clear that our results with early reading will generalize to other content areas (e.g., math, writing, language). These issues need to be addressed in well-designed studies. Nevertheless, these findings have merit as they suggest potential where traditional standardized methods have fallen short.

LIMITATIONS AND FUTURE RESEARCH

Our research has several noteworthy limitations, and it is important to take research on alternate assessments forward in several ways. First, the data analyzed in this investigation were not obtained as part of carefully planned research for the purposes to which they have been used. The problem with missing data and lack of clarity regarding missing responses renders the design somewhat weak. Although all care was taken to use only credible data, there may be bias in the data used. Second, the dichotomization of performance assessments constitutes a breach of one purpose for which they were developed. The constructed response performance assessment is intended to be a rich, flexible assessment not bound necessarily by binary correct/incorrect scoring. For example, we note that there is a very important accumulating body of research regarding passing scores and categorical decisions based on performance assessments (e.g., Berk, 1986; Cizek, 2001; Kane, Crooks, & Cohen, 1999; Livingston & Zieky, 1982). Performance task scoring is usually partial credit at least. The use of score pat-

terns for IRT modeling may be more defensible than dichotomous scoring of the performance assessments (Rosa, Swygert, Nelson, & Thissen, 2001). In addition, elimination of variance when dichotomizing performance assessment responses is rarely defensible (MacCallum, Zhang, Preacher, & Rucker, 2002). Finally, the interpretation of these results pivots around the use of the performance assessments for a very specific population of students in a very specific skill area. The results cannot be generalized immediately to measurement of other populations and other skill areas.

A valuable line of research pertaining to the use and scaling of alternate assessments for students with serious disabilities includes the investigation of specific types of *response patterns*. Measurement models impose well-specified rules governing patterns of expected responses. Identification of response patterns that do not fit these rules has value for understanding classes of response tendencies that are not well explained by the hypothetical measurement model (Levine & Drasgow, 1983). Li and Olejnik (1997) explain the value of person-fit statistics for the Rasch model, such as we applied to the data in our research reported here. With these investigations, it is possible to discern the effects that misfitting response patterns have on ability estimation and the impact these errors have on validity. Either the performance measures can be modified or classes of unusual response tendencies can be identified. Meijer and Sijsma (2001) review current methods for assessing person fit to the data and delineation of characteristic response patterns.

Collectively, the reported research and suggested lines of future research can add to the technical adequacy of alternate assessments. Based on persistent issues surrounding high-stakes alternate assessments, we recommend continued research along the following lines. First, partial credit models should be explored for psychometric appropriateness and simplicity of use in large scale testing. Second, multidimensional models should be explored with the same issues in mind (Thissen, Nelson, Rosa, & McLeod, 2001). Third, considerable effort should be made to identify recurring response patterns that are inconsistent with the hypothesized measurement model. Finally, the use of performance assess-

ments in skill areas other than reading should be examined.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1–28). Hillsdale, NJ: Lawrence Erlbaum.
- Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum.
- Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research*, 56, 137–172.
- Boomsma, A., van Duijn, M. A. J., & Snijders, T. A. B. Eds. (2001). *Essays on item response theory*. New York: Springer.
- Browder, D., Flowers, C., Ahlgrim-Dezell, L., Karvonen, M., Spooner, F., & Algozzine, R. (2004). The alignment of alternate assessment content with academic and functional curricula. *The Journal of Special Education*, 37, 211–223.
- Cizek, G. J. (Ed.). (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum.
- Ford, A., Davern, L., & Schnorr, R. (2001). Learners with significant disabilities. *Remedial and Special Education*, 22, 214–222.
- Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, 57, 423–436.
- Gierl, M. J. (2005). Using dimensionality-based DIF analyses to identify and interpret constructs that elicit group differences. *Educational Measurement Issues and Practice*, 24, 3–14.
- Hasbrouck, J., & Tindal, G. (1992). Curriculum-based oral reading fluency norms for students in grades 2–5. *TEACHING Exceptional Children*, 24, 41–44.

- Hintz, J. M., Owen, S. V., Shapiro, E. S., & Daly, E. J. (2000). Generalizability of oral reading fluency measures: Application of G theory to curriculum-based measurement. *School Psychology Quarterly, 15*(1), 52–68.
- Hintz, J. M., & Petitte, H. A. P. (2001). The generalizability of CBM oral reading fluency measures across general and special education. *Journal of Psychoeducational Assessment, 19*, 158–170.
- Holland, P. & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.
- Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Hanzel procedure. In H. Wainer & H. Brau (Eds.), *Test validity* (pp. 129–145), Hillsdale, NJ: Lawrence Erlbaum.
- Kampfer, S., Horvath, L., Kleinhert, H., & Kearns, J. (2001). Teachers' perceptions of one state's alternate assessment: Implications for practice and preparation. *Exceptional Children, 67*, 361–374.
- Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5–17.
- Kim, S., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement, 22*, 131–143.
- Kleinhert, H., Kearns, J., & Kennedy, S. (1997). Accountability for all students: Kentucky's alternate portfolio assessment for students with moderate and severe cognitive disabilities. *Journal of the Association for Persons With Severe Handicaps, 24*, 88–101.
- Kleinhert, H., & Kearns, J. F. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's alternate assessment for students with significant disabilities. *Journal of the Association for Persons With Severe Handicaps, 24*, 100–110.
- Kleinhert, H., Kennedy, S., & Kearns, J. (1999). The impact of alternate assessments: A statewide teacher survey. *Journal of Special Education, 33*, 93–102.
- Kolen, M. J. (1981). Comparison of traditional and item response methods for equating tests. *Journal of Educational Measurement, 19*, 1–11.
- Kolen, M. J., & Brennan, R. L. (1995). *Test equating: Methods and practices*, New York: Springer-Verlag.
- Koretz, D. M., Bertenthal, M. W., & Green, B. F. (Eds.). (1999). *Embedding questions. The pursuit of common measure in uncommon tests*. Washington, DC: National Academy Press.
- Levine, M. V., & Drasgow, F. (1983). Appropriateness measurement: Validating studies and variable ability models. In D. J. Weiss (Ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 109–131). New York: Academic Press.
- Li, M. F., & Olejnik, S. (1977). The power of the Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement, 21*, 215–231.
- Linacre, M. (2005). WINSTEPS (Version 3.55) [Computer software]. Chicago: MESA Press.
- Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.
- Little, R. J. A., & Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, & M. E. Sobel (Eds.), *Handbook of statistical modeling for the social and behavioral sciences* (pp. 339–369). New York: Plenum.
- Livingston, A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods, 7*, 19–40.
- Masters, G., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika, 49*, 529–544.
- Meijer, R. R., & Sijtsma, K. (2001). Methodological review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107–135.
- Messick, S. (1996). Validity of performance assessment. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment* (pp. 1–18). Washington, DC: National Center for Education Statistics.
- Muraki, E., Hombro, C. M., & Yong-Won Lee. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*, 325–337.
- Nandakumar, R. (1993). Assessing essential unidimensionality for real data. *Applied Psychological Measurement, 17*, 29–38.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.
- Robinson, S. P. (1993). The politics of multiple-choice versus free-response assessment. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 313–324). Hillsdale, NJ: Lawrence Erlbaum.

- Rosa, K., Swygert, K. A., Nelson, L., & Thissen, D. (2001). Item response theory applied to combinations of multiple-choice and constructed-response items—scale scores for patterns of summed scores. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 253–292). Hillsdale, NJ: Lawrence Erlbaum.
- Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement, 1*, 199–218.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201–210.
- Thissen, D., Nelson, L., Rosa, K., & McLeod, L. D. (2001). Item response theory for items scored in more than two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 141–186). Hillsdale, NJ: Lawrence Erlbaum.
- Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Hillsdale, NJ: Lawrence Erlbaum.
- Thissen, D., Wainer, H., & Wang, X. (1994). Are tests comprising both multiple-choice and free-response items necessarily less unidimensional than multiple-choice tests? An analysis of two tests. *Journal of Educational Measurement, 31*, 113–123.
- Tindal, G. (2005). *Alignment of alternate assessments using the Webb system*. Washington, DC: Council of Chief State School Officers.
- Tindal, G., Glasgow, A., Gall, J., VanLoo, D., & Chow, E. (2002). *Oregon Extended Assessment: Technical adequacy analysis summary*. Salem: Oregon Department of Education.
- Tindal, G., McDonald, M., Tedesco, M., Glasgow, A., Almond, P., Crawford, L. et al. (2003). Alternate assessments in reading and math: Development and validation for students with significant disabilities. *Exceptional Children, 69*, 481–494.
- Tindal, J., & Haladyna, T. M. (Eds.). (2002). *Large-scale assessment programs for all students*. Mahwah, NJ: Lawrence Erlbaum.
- Townsend, R., Zhang, L., Vesterman, B., Wise, L., Tindal, G., Winter, P. et al. (2004, June). *Meeting alignment challenges: A computer-based alignment analysis tool with models for analyzing vertical alignment and alternate assessments*. Symposium conducted at the meeting of the National Conference on Large-Scale Assessment, Boston, MA.
- Traub, R. E., & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple choice tests. *Applied Psychological Measurement, 1*(3), 355–369.
- Turner, M., Baldwin, L., Kleinhert, H., & Kearns, J. (2000). The relation of a statewide alternate assessment for students with severe disabilities to other measures of instructional effectiveness. *Journal of Special Education, 34*, 69–76.
- U.S. Department of Education. Title I—Improving the Academic Achievement of the Disadvantaged; Final Rule. Fed. Reg. 68,236 (Dec. 9, 2003).
- U.S. Department of Education. Title I—Improving the Academic Achievement of the Disadvantaged; Final Rule. 68 Fed. Reg. 68,699 (Dec. 9, 2003).
- U.S. National Reading Panel. (2000). *Teaching children to read*. Washington, DC: National Institute of Child Health and Human Development, National Institutes of Health.
- Vale, D. C. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement, 10*, 333–344.
- van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states*. Washington, DC: Council of Chief State School Officers.
- Wilson, D. T., Wood, R., & Gibbons, R. (2003). *TESTFACT: Test scoring, item statistics, and item factor analysis* (Version 4.0.2) [Computer software]. Chicago: Scientific Software International.
- Ysseldyke, J. E., & Olsen, K. R. (1997). Putting alternate assessments into practice: What to measure and possible sources of data. *Exceptional Children, 65*, 175–185.

ABOUT THE AUTHORS

PAUL YOVANOFF (CEC OR Federation), Associate Professor/Senior Research Associate; and **GERALD TINDAL** (CEC OR Federation), Castle-McIntosh-Knight Professor of Education, Educational Leadership Area Head, College of Education, University of Oregon, Eugene.

Address all correspondence to Paul Yovanoff, College of Education, 5253 University of Oregon, Eugene, OR 97403-5253 (e-mail: yovanoff@uoregon.edu).

Manuscript received March 2005; accepted March 2006.