Exploring the Validity of the *Oregon Extended Writing Assessment*

Lindy Crawford, *University of Colorado at Colorado Springs*
Gerald Tindal, *University of Oregon*
Dick M. Carpenter II, *University of Colorado at Colorado Springs*

Research into the technical adequacy of statewide alternate assessments is limited. In this study, the authors analyzed 2 years of data from one state's alternate assessment in written language in an attempt to validate current test score interpretations. More than 1,000 students were included in each year. Findings support the test's technical adequacy on two major dimensions: (a) strong convergent and discriminant evidence for tasks on the assessment and (b) evidence that a collection of writing subtasks contribute unique information to the assessment of writing among students in special education. Implications include the use of the data (a) on subtasks to inform instruction in written language and (b) for accountability purposes.

The 1997 amendments to the Individuals with Disabilities Education Act required states to have alternate assessments for students with disabilities in place by 2000; today, every state has either an alternate assessment or guidelines allowing districts to develop their own assessments (Olson, 2004). As researchers at the National Center on Educational Outcomes indicated, however, "Alternate assessments are still very new, and taking the time for thoughtful reexamination is critical" (Quenemoen, Thompson, & Thurlow, 2003, p. 4). After 6 years of implementation, many states are engaged in "thoughtful reexamination" that is focused on improving their alternate assessments. Unfortunately this reexamination, including related research, tends to be "descriptive rather than evaluative" (Browder et al., 2003, p. 51), and technical adequacy data on the assessment of students with disabilities in statewide systems are lacking (Koretz & Barton, 2003).

Challenges Facing Alternate Assessment Systems

Alternate assessments for students with disabilities must yield technically adequate information about individual student performance *and* contribute to overall accountability efforts at the district and state levels (Johnson & Arnold, 2004). An alternate assessment system designed exclusively for accountability purposes falls short of the individualization required by the Individuals with Disabilities Education Improvement Act of 2004; on the other hand, data provided by a purely diagnostic assessment cannot be aggregated for purposes of accountability. Designing an assessment system to achieve both purposes is daunting and "requires rethinking traditional assessment methods" (No Child Left Behind Act of 2001, ¶3). Unfortunately, research supporting the effectiveness of current alternate assessment systems in meeting this dual challenge is limited (Johnson & Arnold, 2004).

Weak Reliability

Currently, states use various types of alternate assessment systems, including checklists, direct observations, samples of student work, portfolios, and performance assessments (Roeber, 2002). Portfolios are the most frequently used technique in the field of alternate assessments (Thompson & Thurlow, 2003), although there is evidence indicating questionable reliability levels (Kleinert, Kearns, & Kennedy, 1997). Traditional statewide portfolio systems also have encountered similar challenges in regards to establishing reliability (Klein, McCaffrey, Stecher, & Koretz, 1995; Koretz & Barron, 1998).

Weak reliability may be a reason why the number of states using portfolios as the alternate assessment system has decreased over the past 6 years (Thompson & Thurlow, 2003). State educational leaders are seeking new methods for collecting meaningful individual *and* reliable group data, including information pertaining to adequate yearly progress (AYP; No Child Left Behind Act of 2001). To date, many school dis-

Address: Lindy Crawford, Dept. of Special Education, 1420 Austin Bluffs Parkway, PO Box 7150, Colorado Springs, CO 80933 (e-mail: mcrawfor@uccs.edu).

tricts have failed to meet AYP targets due solely to the poor performance of students with disabilities (Chaney, 2004; Dabney, 2004). How much of this performance can be attributed to measurement error or low reliability of alternate assessments remains unknown. What is known is that consistent and reliable measures result in more reliable inferences about a student's skills and the success of a school system.

Poor Validity

Without underplaying the importance of reliability, the ultimate goal is to provide valid data for informing instructional *and* policy decisions, a Herculean task still largely unmet in traditional statewide assessment systems (DeBard & Kubow, 2002). Increasing the amount of validity evidence associated with any assessment system begins with establishing its construct validity. One way to increase the construct validity of alternate assessment systems is through alignment with state academic content standards (Roach, Elliott, & Webb, 2005), but for some alternate assessments, the distance is great between the behaviors tested and the actual constructs as defined in a state's academic content standards. This gap between alternate assessment tasks and academic content standards is illustrated in the following two indicators, which were taken from one state's alternate assessment in math: "demonstrates ability to maneuver appropriately in space," and "sits upright in a wheelchair" (Browder et al., 2004, p. 216).

Alternate assessments also receive criticism for underrepresenting the academic constructs they purport to measure. Johnson and Arnold (2004) presented one example of this underrepresentation in their analysis of Washington State's portfolio system. They found instances for which "identify letters" was the only skill measuring progress toward *four* comprehensive state reading standards. Thus, to increase construct validity, alternate assessments need to be broader in scope and better represent the construct of interest.

On-Demand Performance Assessments as a Possible Answer

On-demand performance-based alternate assessments may prove to be more reliable and valid than many of the alternate assessment systems currently in place. We define *on-demand performance-based alternate assessments* as direct assessments of academic skills completed during one test session, or multiple test sessions, that require students to perform authentic tasks aligned with a state's academic content standards. These performances are not limited to paper-and-pencil tasks; for example, students might demonstrate their skills at sequencing events in a story through use of a storyboard. All on-demand assessments—as opposed to indirect assessments, such as teacher checklists or multiple-choice tests—directly assess students' skills.

On-demand performance-based alternate assessments are meant to increase the reliability of collected data through inclusion of a wide-range of predetermined items that remain consistent across time and students. One example of this type of system is the *Colorado Student Assessment Program Alternate* (CSAPA). In this system, students are assessed on a collection of performance indicators aligned with the state's content standards (Colorado Department of Education, 2003). A consistent measurement system, such as the CSAPA, increases the likelihood that reliable decisions that use the collected data will be made. Progress, including AYP, cannot be measured with a moving target—a rule applied to general education measurement systems but not always adhered to in the field of alternate assessments.

On-demand performance assessments may more thoroughly represent the academic construct they purport to measure than systems designed primarily for instructional purposes. Unlike a collection of highly individualized work samples, on-demand performance assessments collect information on a range of skills, each of which contributes unique information about the target construct. Not all students complete all items (as is the case with Oregon's system), nor do they complete each item with the same level of assistance or prompts (as is the case with Colorado's system). Individualization thus resides not in the tasks that compose the assessment but in the items completed by the students and the manner in which they are completed.

Theoretically, on-demand performance assessments may be able to meet both goals of alternate assessment systems: diagnosis and accountability. Proof that on-demand performance assessments provide reliable data and strong validity evidence is needed, however. In our attempt to establish this type of evidence. we examined the technical adequacy of one state's alternate on-demand performance assessment, including its accuracy in measuring the intended construct through a collection of theoretically unique tasks.

Amassing Evidence

Validity evidence of the sort we noted in the previous paragraphs consists of two essential components:

1. a theoretical definition of the construct for which the test was designed and
2. data validating (or invalidating) this theoretical conceptualization.

In the following sections, we provide a conceptual framework for each of these two components.

The Construct of Writing

Berninger (2000) developed a framework for understanding the writing development of younger students or students with low-level writing skills. Berninger's *Functional Writing System* consists of four components: transcription, text genera-

tion, executive functions, and memory. The first two components provide the foundation needed to engage fully in the final two components.

The first component, *transcription,* incorporates both handwriting and spelling as needed tools for transcribing oral language into written text. This theoretical foundation has been supported by factor analyses indicating that handwriting and spelling load on separate but correlated factors (Berninger, 2000). The second component, *text generation,* also measures a separate factor (Abbott & Berninger, as cited in Berninger, 2000). Text generation consists of "translating ideas into visible language" (Berninger, 2000, p. 67) and relies on students' transcription skills. Text generation can be evaluated on different dimensions, including fluency (as measured by words written within a certain amount of time) or quality (as measured by a holistic rating). Researchers have found that fluency and quality are moderately correlated (Tindal & Parker, 1989), even though they load as different factors (Berninger, Cartwright, Yates, Swanson, & Abbott, 1994). The final two stages in the *Functional Writing System* consist of *executive functioning* (globally defined as *self-regulation*) and *memory*. Both are important skills in writing but rely on relative proficiency in the two traits preceding them. The more automatic the low-level transcription skills, the more short-term resources are available for these advanced writing behaviors (McCutchen, 1988; Troia & Graham, 2003). Researchers in the field of curriculum-based measurement also have attempted to define the construct of writing through a collection of subskills that contribute to the larger writing process. Lembke, Deno, and Hall (2003) studied the writing behaviors of elementary school–age students and found that transcription skills, such as dictating words and sentences and copying entire sentences, were moderately to strongly correlated with more advanced text-generation behaviors. They concluded that the measures of sentence copying and dictating showed great promise as indicators of early writing proficiency.

The need for an alternate assessment that would measure basic transcription skills is clear, especially in light of the difficulties that students with disabilities have with spelling, punctuation, capitalization, and handwriting (Graham, 1999; Graham & Harris, 1997; Graham, Harris, MacArthur, & Schwartz, 1991). In addition, the field needs an alternate assessment that measures text generation, because writers with disabilities and poor writers have been found to write less than their general education peers (Englert, 1990; Graham et al., 1991). This is an important finding when one considers the moderate correlations reported between the amount of writing produced under timed conditions and the quality of the writing product (Berninger et al., 1994; Tindal & Parker, 1989).

Validity Evidence Supporting a Test Construct

Once one has established, or clarified, the theoretical construct a test purports to measure, test validation requires producing a clear link between the theoretical construct and the data generated by the actual assessment (Messick, 1995). Validity evidence is strengthened when predicted relationships among the theorized test constructs are empirically supported (Cronbach & Meehl, 1955). These empirical relationships should be strong when components are meant to be related and weak when they are purported to measure different constructs, thus demonstrating both *convergent* and *discriminant evidence*. These are complements, but both must be apparent when establishing construct validity (Messick, 1995). Strong interitem correlations within a particular construct will provide convergent evidence. Similarly, factor analyses are used to support a test's convergent validity (theoretically similar constructs load together) and discriminant validity (unrelated constructs load on separate factors). In the end, each test component should contribute unique information (variance) to a total test score. Evidence of this sort lies at the heart of construct validity.

Purpose of the Study

Alternate assessments have become an integral part of statewide accountability efforts, and the high stakes associated with educational assessment demand that any alternate assessments that are developed should be "conceptually sound and educationally valid" (Browder et al., 2003, p. 57). In the study on which we report here, we used 2 years of data to empirically investigate the reliability and validity of the *Oregon Extended Writing Assessment* (EWA), one component of Oregon's alternate assessment system. Our primary research questions were as follows:

1. Do convergent evidence and discriminant evidence support test reliability?
2. Does the assessment measure discernible writing constructs, and if so, what is the nature of those constructs?

Method

Oregon's Alternate Assessment Program

Oregon's alternate assessment system consists of two prongs: *The Extended Career and Life Role Assessment System* (CLRAS) and three separate *Extended Academic Assessments* (*Extended Reading, Extended Writing,* and *Extended Mathematics*; Oregon Department of Education, 2003a). The CLRAS is designed for use with students who communicate using single words, gestures, or pictures. Students completing the CLRAS often need systematic, individualized instructional and/or technological supports to make progress in learning, whereas the *Extended Academic Assessments* are designed for students who receive instruction in basic academic domains

and require less individualized instruction. The *Extended Academic Assessments* rely on a "behavioral research base" for their development (Quenemoen et al., 2003, p. 46), and measure progress on basic academic skills.

Extended Writing Assessment

Conceptual Framework. The versions of Oregon's EWA analyzed in this study (2001 and 2002) contain six tasks ranging in difficulty from copying letters to story writing. Initial tasks are considered the most basic, with the assessment progressing to more advanced tasks. As alluded to previously, this approach to measuring writing is more behavioral than other definitions of writing (see, e.g., Hayes, 1996) and is largely based on research in the field of curriculum-based measurement (CBM; Tindal et al., 2003). Researchers in the field of CBM have found the writing of students with disabilities to contain enough basic transcription errors to make evaluation of the message as a whole rather challenging. Evaluating a student's ability to communicate in writing through more objective and countable measures thus may be more instructionally useful (Tindal & Parker, 1989).

The intent of the EWA is "to document student skill in composing units of increasing complexity (letters, words, sentences, and stories) to communicate meaning" (Oregon Department of Education, 2003a, p. 7). Scores on six writing tasks assess students' writing proficiency, and final scores reflect the total number of correct items within each task. Students' scores are transposed onto a qualitative scale ranging from 0 to 6 points (0 suggests that the *assessment may be too difficult for the student at that time or that additional adaptations are needed*, 5 or 6 suggests that *the student take the traditional statewide test, possibly with accommodations*). The decision rules used to assign this global score are outlined in Figure 1. Student performance is, therefore, evaluated quantitatively and qualitatively.

Task Administration and Scoring. The six tasks are administered one to one by an evaluator, most often the student's teacher. The evaluator reads the standardized directions for each task, and the student completes the task on his or her scoring protocol. The first four tasks consist of copying and dictating activities. These are not timed, and the student is prompted to copy or write each letter, word, or sentence exactly as it is shown in the test protocol or dictated by the evaluator. Task 1, Copying Letters, consists of 10 individual letters and is scored in the following manner: The evaluator awards 0 to 2 points per letter, depending on its accuracy (models of correct, partially correct, and incorrect letters are provided in the administrator's manual). In Task 2, Copying Words, the student copies eight words (each consisting of two to three letters). Words are awarded 0 to 4 points, for a possible total of 32 points.

In Task 3, Dictating Words, the evaluator reads eight high-frequency words (e.g., *he* or *and*) to the student and scores each written word for correct letter sequences (CLS). A correct letter sequence is defined as "a pair of letters (or spaces and letters) correctly sequenced within a word" (Oregon Department of Education, 2002a, p. 10). Task 3 is worth a total of 29 points. Task 4, Dictating Sentences, is also scored using CLS. It consists of two short sentences and is worth a total of 42 points.

Tasks 5 and 6 call for authentic student writing in response to standardized prompts. In Task 5, Writing Sentences, the evaluator reads a sentence (e.g., "Make up a sentence about school and write it on this line"), and the student responds in writing. Task 5 is not timed. Because students write sentences of different lengths, evaluators record the raw scores for total words written and for correct letter sequences.

Finally, in Task 6, the evaluator asks the student to write a story based on a prompt (e.g., "I would like you to write a story describing something that you do every day"). The evaluator is encouraged to spend some time discussing ideas with the student, and then the student is given 3 min to write his or her story. Task 6 is scored along three dimensions: total words written, correct word sequences (CWS), and a 4-point rubric scale measuring ideas and organization. As noted in the *Extended Writing Administration* (2002), "A CWS is a sequence of adjacent correctly spelled and punctuated words that are judged to be syntactically and semantically correct" (p. 12). See Table 1 for a summary of the tasks and their corresponding number of items.

0	*Nonwriter.* Student copies letters (Task 1) at less than 50% accuracy.
1	*Beginning Writing.* Student copies letters (Task 1) at greater than 50% accuracy.
2	*Emerging Writing.* Student copies words (Task 2) at greater than 50% accuracy.
3	*Early Writing.* Student writes words from dictation (Task 3) at greater than 50% accuracy.
4	*Developing Writing.* Student writes sentences from dictation (Task 4) at greater than 50% accuracy.
5	*Engaged Writing.* Student writes sentences (Task 5) at greater than 50% accuracy on measure of correct letter sequence.
6	*Accomplished Writing.* Student writes stories (Task 6) at greater than 50% accuracy on measure of correct word sequence.

FIGURE 1. Decision rules for assigning global score.

TABLE 1. Tasks Assessed on the Extended Writing Assessment

Task number	Total items	Task name
1	10	Copying Letters
2	8	Copying Words
3	8	Dictating Words
4	2	Dictating Sentences
5.1	1	Writing Sentences (total words)
5.2	1	Writing Sentences (correct letter sequences)
6.1	1	Story Writing (total words)
6.2	1	Story Writing (correct word sequences)
6.3	1	Story Writing (ideas and organization)

Participants

In 2001, a total of 1,861 students completed one of the two alternate assessment options in the state of Oregon (Oregon Department of Education, 2002b). This number represents approximately 1% of the students participating in the statewide assessment system (Oregon Department of Education, 2003b). Our data set for 2001 consisted of those students in Grades 3, 5, 8, and 10 who completed the EWA (1,342 students). For 2002, our database consisted of 1,017 students (see Table 2 for demographic information).

Unfortunately, we were unable to obtain information related to the percentage of students with specific disability types who completed the EWA. However, the decision rule regarding who should take the extended assessments reads as follows: "Student receives instruction primarily at emerging academic skill levels and well below benchmark (Grade 3)" (Oregon Department of Education, 2001, p. 1). Therefore, out

TABLE 2. Student Ethnicity and Gender by Grade: 2001 and 2002

Ethnicity/ gender	Grade 3		Grade 5		Grade 8		Grade 10	
	2001	2002	2001	2002	2001	2002	2001	2002
Caucasian								
Male	251	155	177	131	61	63	66	54
Female	129	69	91	95	35	31	57	32
Unknown	1	1	1	1	0	0	0	0
Latino/a								
Male	38	31	28	32	11	8	11	7
Female	24	20	17	18	11	8	11	0
Unknown	0	0	0	0	1	0	0	0
African-American								
Male	16	17	16	15	6	5	5	5
Female	14	6	9	12	4	4	3	3
Unknown	0	1	0	0	0	0	0	0
American Indian								
Male	9	8	3	7	5	1	1	1
Female	1	4	3	0	3	0	1	0
Unknown	0	0	0	0	1	0	0	0
Asian/Pacific Isl.								
Male	7	9	3	8	2	5	2	3
Female	5	3	3	4	2	1	1	1
Unknown	0	0	0	0	0	0	0	0
Unknown								
Male	28	15	26	9	19	8	5	7
Female	17	2	12	7	7	3	9	1
Unknown	0	1	0	0	1	2	0	2
Totals	540	342	389	339	169	139	172	116

No grade-level data were available for 72 students (2001) and 81 students (2002); not included above but data included in other analyses.

of the 1% of the student population who completed *either* an extended assessment or the *Career and Life Roles Assessment,* only those students receiving academic instruction (as opposed to life skills instruction) completed the EWA.

Not all students completed all tasks, and many students were exempted from individual tasks because their skills were judged by teachers as being too advanced (Not Administered–Proficient [NA-P]) or too limited (Not Administered–Inappropriate [NA-I]); thus, the numbers for each analysis fluctuate. For tasks with a consistent denominator, we assigned students full credit or no credit to match their NA-P or NA-I label.

Research Questions

We explore three questions related to test reliability:

1. What is the interrater reliability for scoring students' performance on individual tasks?
2. What is the interitem reliability *within* individual tasks?
3. What is the correlation between individual items and task totals?

Reliability coefficients, if strong, validate an assessment on one dimension, but equally critical is the collection of validity evidence pertaining to the ability of the assessment to measure the constructs it purports to measure. Therefore, we posed the following questions:

1. Does the assessment measure discernible writing constructs?
2. If so, do those constructs align with the six different tasks measured in the EWA?

Data Analysis

We used the 2001 dataset to conduct an interrater reliability study. We randomly selected 10% of the completed extended assessments, totaling approximately 180 protocols. Not all 180 students completed all three tests in the extended assessment battery (i.e., Math, Reading, and Writing), however, and not all students completed all of the tasks on each test. In the end, out of the 180 extended assessments selected for the interrater reliability analysis, 80 represented fully complete EWAs.

We defined *reliability* as level of agreement in scoring obtained between two scorers. It was calculated by examining the total number of points each scorer awarded a student on each task. For example, if the first scorer gave 14 points for Task 1 and the second scorer gave 15 points for the same task, the overall agreement for the task would be 14/15, or 0.93. After we calculated agreement for each task, we calculated the average agreement of all tasks for *each* student. Finally, we calculated the average agreement of each task across *all* students. For example, if on Task 2 four students had reliabilities of 1.00, 1.00, 0.90, and 0.90, the average reliability across these students would be 1.00 + 1.00 + 0.90 + 0.90/4, or 0.95.

Four graduate students who received the same training that the evaluators received scored the 80 protocols. Training consisted of a full-day workshop with demonstrations on how to administer and score the extended assessments and multiple opportunities for on-site practice with expert feedback. Teachers earned "test certification" by scoring one full administration of a videotaped presentation of the extended assessments. We gave each teacher a videotape to take home, score, and submit when finished. Graduate students attended the training and scored the Writing portion of the videotape. Students were unaware that their scores would be analyzed as part of a study.

We used Cronbach's alpha to investigate the internal consistency of each task so as to provide evidence to support each task's convergent validity. We explored the issue of discriminant validity (i.e., evidence that each task measured a unique writing component) through the use of factor analyses.

For the validity analysis, we performed a principal components analysis two ways: first, with a varimax rotation and, second, by designating a four-factor model. The varimax rotation measured the intercorrelations among the means of Tasks 1 through 4 discretely, 5.1 (Writing Sentences, total words), 5.2 (Writing Sentences, CLS), 6.1 (Story Writing, total words), 6.2 (Story Writing, CWS), and 6.3 (Story Writing, ideas and organization). We performed both analyses separately on the 2001 data, the 2002 data, and the years combined, allowing us to test the reliability of the factors across years.

We used the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity to determine the advisability of proceeding with the factor analyses (Norussis, 1985). The KMO measure of sampling adequacy compares the magnitudes of the observed correlation coefficients to the magnitudes of the partial correlation coefficients. A large KMO measure makes a factor analysis advisable because correlations between pairs of variables can be explained by other variables (Norussis, 1985). The analyses obtained a KMO sampling adequacy measure of .81 for 2001, 2002, and both years combined, which is considered acceptable.

The Bartlett's test of sphericity is used to test the hypothesis that the correlation matrix is an identity matrix (Rivera & Ganaden, 2001); that is, all diagonal terms are 1 and all off-diagonal terms are 0. The current study obtained test-of-sphericity values of 8,097.24 for 2001, 5,484.55 for 2002, and 13,245.93 for both years combined. Given these large values and the associated significance levels of .000 for each, it is unlikely that the correlation matrices indicated an identity. The KMO values and the highly significant levels of the test of sphericity both indicated the data were adequate for factor analysis.

TABLE 3. Mean Scores Across Tasks: 2001 and 2002

Task	Year	# Cases	*M*	*SD*	Minimum	Maximum
1. Copy Letters	2001	1342	18.16	4.9	0	20
	2002	1017	17.00	6.5	0	20
2. Copy Words	2001	1342	28.04	8.7	0	32
	2002	1017	27.44	10.1	0	32
3. Dictate Words	2001	1325	18.87	10.2	0	32
	2002	1017	18.90	11.2	0	32
4. Dictate Sentences	2001	1342	23.39	14.5	0	42
	2002	1017	21.78	14.6	0	42
5.1. Sentences (words)	2001	1342	4.00	3.8	0	33
	2002	1017	5.40	5.3	0	35
5.2. Sentences (CLS)	2001	1342	11.16	10.3	0	65
	2002	1008	14.48	13.3	0	88
6.1. Story (total words)	2001	1267	14.34	14.7	0	99
	2002	1017	13.32	14.2	0	77
6.2. Story (CWS)	2001	1242	8.96	12.6	0	98
	2002	1017	8.97	13.4	0	99
6.3. Story (rubric)	2001	1236	.97	1.0	0	3
	2002	1017	1.00	1.0	0	3

Results

Descriptive Statistics

We first analyzed mean scores on all tasks by grade level. We found some significant differences, but a clear pattern was not established. For example, in the 2001 dataset, a significant difference between grade level occurred for two of the nine tasks: Task 6.1, Story Writing (total words written), $F(3, 1266) = 3.86$, $p = .009$. and Task 6.3, Story Writing (ideas and organization), $F(3, 1166) = 2.77$, $p < .04$. On Task 6.1, the mean differences demonstrated a grade-level pattern, with Grade 3 students writing the fewest number of words ($M = 6.78$), and Grade 10 students writing the most ($M = 14.21$). Grade 5 and Grade 8 students averaged 9.39 and 10.15, respectively. Mean differences on Task 6.3, however, did not follow a grade-level pattern, with mean scores as follows: Grade 3 = .93, Grade 5 = 1.09, Grade 8 = .88, and Grade 10 = .92.

In the 2002 dataset, two tasks revealed a significant difference, but neither supported a grade-level pattern. Mean Scores on Task 1, Copying Letters, were as follows: Grade 3 = 17.66, Grade 5 = 17.77, Grade 8 = 17.24, and Grade 10 = 17.13. Task 2, Copying Words, resulted in these mean scores: Grade 3 = 28.71, Grade 5 = 28.91, Grade 8 = 27.54, and Grade 10 = 27.37. One explanation for the lack of a grade-level pattern may lie in the diversity of skills held by students with disabilities. In special education, an increase in grade level does not necessarily mean an increase in scores. Considering that an increase in grade level resulted in a significant difference in only 1 of the 18 total opportunities, we combined within-year scores of all students on all analyses (see Table 3).

Reliability

The EWA consists of multiple tasks, each task representing 1 to 10 items (see Table 1). Average interrater agreement for each task (across a random selection of 80 protocols) ranged from .89 to .98. Overall agreement across all writing tasks was .93.

For the 2001 data set, we calculated Cronbach's alpha coefficient for the first four tasks. The coefficient for Task 1 was $\alpha = .97$. Similar findings were found for Task 2 ($\alpha = .96$) and Task 3 ($\alpha = .94$), with Task 4 demonstrating a weaker correlation ($\alpha = .77$). Supporting these coefficient alphas were correlations between individual items and their task totals (see Table 4). These analyses were not conducted on Tasks 5 and 6 because each task represented a single item (i.e., a sentence or a story), as opposed to a collection of multiple items, as

TABLE 4. Item to Task Total Correlations: 2001 and 2002

Item	Task 1: Total	Task 2: Total	Task 3: Total	Task 4: Total
		2001		
1	.930	.887	.849	.913
2	.868	.860	.798	.962
3	.920	.910	.796	
4	.889	.918	.882	
5	.914	.919	.859	
6	.915	.926	.862	
7	.909	.900	.865	
8	.923	.871	.884	
9	.911			
10	.848			
		2002		
1	.916	.923	.839	.952
2	.913	.938	.900	.980
3	.923	.923	.904	
4	.920	.928	.890	
5	.917	.938	.905	
6	.914	.929	.856	
7	.923	.928	.854	
8	.910	.939	.874	
9	.923			
10	.929			

was the case with Tasks 1 through 4. Similar to the 2001 findings, alpha coefficients for individual tasks on the 2002 assessment were quite strong ($\propto$ = .80 to $\propto$ = .97), as were the correlations between individual items and task totals (see Table 4).

Validity

In the varimax rotation, we applied the Kaiser criterion. This is also known as the "eigenvalue-greater-than-1" criterion (Stevens, 2002). Results of the procedure in SPSS gave two common factors with eigenvalues greater than 1.00 for all three data sets. The two factors, their corresponding eigenvalues, the percentage of variance each factor contributes to the construct of writing, and the cumulative percentage of variance are presented in Table 5.

The tasks clustered for each factor identically for 2001, 2002, and both years combined. Factor 1 included Tasks 1 through 4 (2001: α = .86, 2002: α = .88, both years combined: α = .87), and Factor 2 included Tasks 5.1 through 6.3 (2001: α = .87, 2002: α = .85, both years combined: α = .86).

These results proved statistically satisfactory, but only moderately so. To begin, the total accumulative variance for each year is at the low end of acceptable standards (Stevens, 2002). Second, the factor loadings in the varimax rotation were not optimal. For example, in 2001, 2002, and both years combined, Factor 2 loadings indicated a strong relationship between Tasks 1 and 2 (2001 = .91 and .92; 2002 = .91 and .92; combined = .91 and .92) and between Tasks 3 and 4 (2001 = .65 and .61; 2002 = .76 and .70; combined = .72 and .67), but a relatively large gap between Tasks 1 and 2 and Tasks 3 and 4.

We therefore speculated that a four-factor model might provide a better representation of the principle components contained in the data and reanalyzed by dictating a four-factor model instead of the varimax/Kaiser method. The results confirmed our speculation. Table 6 includes the four factors, the percentage of variance each factor contributes to the construct of writing, and the cumulative percentage of variance. As the table indicates, the four-factor model accounts for substantively more variance than does the two-factor model. Factor loadings also indicated a closer relationship between variables within each factor.

Moreover, the four factors better align conceptually with both the theory and practice associated with the writing assessment. Factor 1 encompasses Story Writing Tasks 6.1 through 6.3 (2001: α = .86, 2002: α = .80, both years com-

bined: $\alpha = .83$). Factor 2 contains Copying Letters and Words Tasks 1 through 2 (2001: $\alpha = .93$, 2002: $\alpha = .94$, both years combined: $\alpha = .93$). Factor 3 consists of Dictating Words and Sentences Tasks 3 and 4 (2001: $\alpha = .93$, 2002: $\alpha = .87$, both years combined: $\propto = .90$). Factor 4 encompasses Writing Sentences Tasks 5.1 through 5.2 (2001: $\alpha = .74$, 2002: $\alpha = .81$, both years combined: $\alpha = .78$).

TABLE 5. Eigenvalues, Percentage of Variance Contributed, and Cumulative Percentage of Variance for Two Factors

Factor	Eigenvalues	% of Variance	Cumulative %
		2001	
1	3.740	41.552	41.552
2	2.658	29.535	71.086
		2002	
1	3.304	36.716	36.716
2	2.897	32.192	68.908
		Combined	
1	3.484	38.716	38.716
2	2.782	30.913	69.630

TABLE 6. Eigenvalues, Percentage of Variance Contributed, and Cumulative Percentage of Variance for Four Factors

Factor	Eigenvalues	% of Variance	Cumulative %
		2001	
1	2.557	28.409	28.409
2	1.950	21.666	50.075
3	1.820	20.225	70.300
4	1.468	16.308	86.608
		2002	
1	2.196	24.396	24.396
2	1.999	22.212	46.607
3	1.745	19.394	66.001
4	1.669	18.548	84.549
		Combined	
1	2.374	26.378	26.378
2	1.956	21.729	48.107
3	1.774	19.706	67.814
4	1.596	17.735	85.549

Discussion

Our purpose in analyzing 2 years of data from one state's alternate assessment was to provide empirical evidence for an on-demand, performance-based alternate assessment. Overall, the data derived from this study supported the technical adequacy of Oregon's EWA. Alpha coefficients for individual tasks verified their internal consistency, as did the moderate-to-strong relationships apparent in item-to-task total correlations. Both of these analyses contributed evidence for the test's convergent validity in that the items *within* each task grouped together to measure one unified writing construct.

Discriminant evidence, on the other hand, was not as straightforward. Specifically, the six separate tasks on the test were theoretically meant to measure six *different* components of writing. Factor analyses, however, revealed only four distinct factors, each contributing unique variance. Therefore, the assumption inherent in the design of the EWA—that separate tasks represent distinct constructs—was not validated. Although we did not unveil a separate factor for each of the six tasks currently added to arrive at a global score, we did find that these six tasks collapsed into four components that were similar in the amount of variance they contributed: Story Writing (26%), Copying (22%), Dictating (20%), and Sentence Writing (18%), using the combined dataset.

This difference is more statistical than conceptual, however, in that our model of four factors combined Tasks 1 and 2 (Copying Letters and Copying Words) and Tasks 3 and 4 (Dictating Words and Dictating Sentences). This indicated that the two behaviors (copying and dictating) loaded individually but their item complexity (letter, word, or sentence level) did not. An analysis of the validity of the alternate assessment within this four-factor framework indicated that the factors do provide discriminant evidence in that they represent truly different skills (i.e., copying, dictating, sentence writing, and story writing). Convergent evidence was apparent in the moderate-to-strong alphas found between tasks loading within the same factor. Furthermore, these same four factors loaded consistently across two different data sets, representing two different years, which further supported the consistency of the EWA.

Our model did not, however, provide validity evidence for the global score assigned when using the decision-making framework. The decision-making framework is additive and awards more credit to students as they complete tasks "higher" in the continuum (see Figure 1). Two of our findings challenged this framework. First, Tasks 1 and 2, as well as Tasks 3 and 4, did not represent separate constructs; thus, they should not be viewed as additive. Second, each of the four factors contributed relatively similar amounts of variance. A more accurate decision-making framework might be to view tasks as measuring unique constructs and establish cut scores for each to make decisions about a student's readiness to participate in the traditional statewide writing test. This represents a conjunctive, as opposed to a compensatory, scale; a

student must show proficiency in each of the four areas of writing (strength in one area cannot compensate for weakness in another). A decision-making framework using a conjunctive scale incorporates our finding that each construct contributes unique, but relatively similar, amounts of variance.

Limitations: Validity as an Evolving Process

Validity evidence is multifaceted and evolving. Our study provides only initial evidence supporting the construct validity of Oregon's EWA. As a first step, we needed to establish the internal consistency of this assessment and its claims related to the basic constructs it purports to measure. Investigations such as the one we report on here will not fully support a test's construct validity, however. Further investigations, including other previously validated instruments, are needed. Specifically, we did not explore the nature of the relationships between the scores on the alternate assessments and the student's scores on other measures that are assumed to be similar and on measures that are assumed to be different; in not doing so, we have restricted our definitions of convergent evidence and discriminant evidence. Convergent validity evidence would have been strengthened had we triangulated our findings across other independent instruments measuring the same constructs. Similarly, evidence of weak correlations between the EWA and other nonrelated measures would have strengthened discriminant evidence.

A second limitation of our study was absence of an analysis of student performance by disability category. At the time of data collection, the Oregon Department of Education (ODE) had not assigned unique identification numbers to students. Therefore, to respect students' confidentiality, the ODE did not ask teachers to identify students by disability category on the alternate assessment. We thus are unsure whether an analysis of test scores by disability would have provided further validity evidence for the assessment system.

Finally, caution needs to be taken when interpreting interrater reliability data. The number of protocols scored for the interrater reliability analysis was relatively small, and the analysis we used to calculate agreement was rather simple. Accurate scoring of the EWA depends on the skills of individual teachers, and although Oregon has provided extensive training for teachers, more research on the interrater reliability of the test is needed. Also important is analysis of data from direct observations of test administrations.

Implications for Policy and Practice

Although alternate assessments should inform instruction, they can no longer be highly individualized if we are to use their data to measure adequate yearly progress. Advocates of students in special education may challenge the lack of individualization inherent in on-demand performance assessments consisting of a predetermined set of items, but tests are never intended to represent the entirety of what students know and can do. Rather, these tests are designed to provide summary information from which one can infer other skills (Ford, Davern, & Schnorr, 2001). Performance-based alternate assessments can provide diagnostic information about individuals while still maintaining consistency across tasks and in the type of data they provide. A theoretical argument for the use of performance-based alternate assessments is not sufficient, however. Empirical data are tantamount to establishing the credibility of these assessments. Because very few studies have provided empirical evidence for alternate assessments, and in particular for performance-based alternate assessments of writing, our findings have broad implications.

Perhaps most important, our study demonstrated that alternate assessments can reliably assess the writing skills of students with significant disabilities and that these skills can be assessed through multiple tasks contributing unique information. As Gordon, Engelhard, Gabrielson, and Bernknopf (1996), noted, "Adding up separate performances on multiple, diverse, and content-specific tasks leads to valid inferences about the construct" (p. 84). Our analysis demonstrated that after combining Tasks 1 and 2 and Tasks 3 and 4, the EWA does measure a set of unique components contributing to a total score.

A second implication relates to the importance of designing assessments that provide diagnostic information. Unlike tests of reading or math, which are composed of a combination of tasks designed to measure different components of the underlying construct, large-scale writing assessments often require a unified product that *implicitly* measures students' proficiency on multiple components. Many statewide assessments score these writing products using a holistic or analytic (e.g., six-trait writing) scale. These scales, however, may not be sensitive enough to measure the growth of students with disabilities or able to provide diagnostic information (Miller & Crocker, 1990). Furthermore, holistic and analytic scales demand a writing product of some length, and students with disabilities tend to produce limited samples (Englert & Raphael, 1988). Thus, the validity of score interpretation is compromised for students with disabilities when both the writing sample and the scale used to measure it are so limited. Oregon's EWA was thus designed to provide information about writing subcomponents, and we concluded that teachers can use results from the four subcomponents (Story Writing, Copying, Dictating, and Sentence Writing) to identify a student's particular strengths and weaknesses.

Finally, research into traditional statewide assessments has shown an undervaluing of test information by teachers and a reported lack of use of test score data to drive instruction (Abrams, Pedulla, & Madaus, 2003; Vogler, 2002). Little work of this nature has been conducted on alternate assessments. In addition, although the focus of this study was to provide validity evidence for an alternate assessment, a technically ad-

equate test is meaningless if it provides information viewed as irrelevant to its constituents.

REFERENCES

Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the classroom: Tteachers' opinions of statewide testing programs. *Theory Into Practice*, *42*, 18–29.

Berninger, V. W. (2000). Development of language by hand: Its connections with language by ear, mouth, and eye. *Topics in Language Disorders*, *20,* 65–84.

Berninger, V., Cartwright, A., Yates, C., Swanson, H. L., & Abbott, R. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades: Shared and unique variance. *Reading and Writing*.: *An Interdisciplinary Journal*, *6*, 161–196.

Browder, D., Flowers, C., Ahlgrim-Delzell, 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.

Browder, D. M., Spooner, F., Algozzine, R., Ahlgrim-Delzell, L., Flowers, C., & Karvonen, M. (2003). What we know and need to know about alternate assessment. *Exceptional Children, 70,* 45–62.

Chaney, S. (2004, November 28). Thirty-four area schools miss "adequate progress" mark. *Colorado Springs Gazette,* pp. B1, B4.

Colorado Department of Education. (2003). *What you should know about the Colorado Student Assessment Program Alternate (CSAPA).* Retrieved March 1, 2005, from http://www.cde.state.co.us/cdesped/StuDis-Sub2.asp

Cronbach, L. J., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Dabney, M. (2004). August 24). Indiana schools expand special-needs programs under No Child Left Behind law. *Indianapolis Star. Retrieved March*, *4*, 2005. from Newspaper Source database, http://search.epnet.com/login.aspx?direct=true&db=nfh&an=2W642000786915>

DeBard, R., & Kubow, P. K. (2002). From compliance to commitment: The need for constituent discourse in implementing test policy. *Educational Policy, 16,* 387–405.

Englert, C. S. (1990). Unraveling the mysteries of writing through strategy instruction. In T. E. Scruggs & B. Wong (Eds.), *Intervention research in learning disabilities* (pp. 180–223). New York: Springer Verlag.

Englert, C. S., & Raphael, T. (1988). Constructing well-formed prose: Process, structure and metacognitive knowledge. *Exceptional Children, 54,* 513–520.

Ford, A., Davern, L., & Schnorr, R. (2001). Learners with significant disabilities. *Remedial and Special Education, 22,* 214–222.

Gordon, B., Engelhard, G., Gabrielson, S., & Bernknopf, S. (1996). Conceptual issues in equating performance assessments: Lessons from writing assessment. *Journal of Research and Development in Education, 29,* 81–88.

Graham, S. (1999). Handwriting and spelling instruction for students with learning disabilities: A review. *Learning Disability Quarterly, 22,* 78–98.

Graham, S., & Harris, K. R. (1997). It can be taught, but does it develop naturally: Myths and realities in writing instruction. *School Psychology Review, 26,* 414–424.

Graham, S., Harris, K. R., MacArthur, C., & Schwartz, S. (1991). Writing and writing instruction for students with learning disabilities: Review of a research program. *Learning Disability Quarterly, 14,* 89–114.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 1–27). Mahwah, NJ: Erlbaum.

Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C. § 1400 *et seq*.

Individuals with Disabilities Education Improvement Act of 2004), 20 USC § 1400 *et seq*.

Johnson, E., & Arnold, N. (2004). Validating an alternate assessment. *Remedial and Special Education, 25,* 266–275.

Klein, S. P., McCaffrey, D., Stecher, B., & Koretz, D. (1995). The reliability of mathematics portfolio scores: Lessons from the Vermont experience. *Applied Measurement in Education, 8,* 243–260.

Kleinert, H., Kearns, J., & Kennedy, S. (1997). Accountability for all students: Kentucky's alternate portfolio system for students with moderate and severe cognitive disabilities. *The Journal of the Association for Persons with Severe Handicaps, 22,* 88–101.

Koretz, D., & Barron, S. I. (1998*). The validity of gains in scores on the Kentucky Instructional Results Information system (KIRIS).* Santa Monica, CA: Rand Institute on Education and Training.

Koretz, D. M., & Barton, K. (2003). *Assessing students with disabilities: Issues and evidence* (CSE Technical Report 587). Los Angeles: University of California, Center for the Study of Evaluation.

Lembke, E., Deno, S. L., & Hall, K. (2003). Identifying an indicator of growth in early writing proficiency for elementary school students. *Assessment for Effective Intervention*, *28*(3&4), 23–35.

McCutchen, D. (1988). "Functional automaticity" in children's writing: A problem of metacognitive control. *Written Communication*, *5*(3), 306–324.

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

Miller, M. D., & Crocker, L. (1990). Validation methods for direct writing assessment. *Applied Measurement in Education*, *3*, 285–296.

No Child Left Behind Act of 2001 (2002). February 3, 2005, from http://education.umn.edu/nceo/TopicAreas/AlternateAssessments/alt_assess.topic.htm

Norussis, M. J. (1985). *SPSS-X advanced statistics guide*. New York: McGraw-Hill.

Olson, L. (2004). January 8). Enveloping expectations: Federal law demands that schools teach the same content to children they wrote off a quarter-century ago. *Education Week*, *23*(17), 8–20.

Oregon Department of Education. (2001). *Statewide assessment decision: What is the level of student's current curriculum/instruction?* Retrieved March 20, 2005 from http://www.ode.state.or.us/pubs/forms/iep/

Oregon Department of Education. (2002a). *Extended writing administration*. Salem: Author.

Oregon Department of Education. (2002b). *Oregon statewide assessment 2001–2002* (District Participation Report: State Data). Salem: Author.

Oregon Department of Education. (2003a). *Oregon statewide extended assessments: Interpretive guide for individual student reports and class roster.* Salem: Author.

Oregon Department of Education. (2003b). *Part B annual performance report*. Retrieved March 20, 2005, from http://www.ode.state.or.us/pubs/sped/publications.aspx

Quenemoen, R., Thompson, S., & Thurlow, M. (2003). *Measuring academic achievement of students with significant cognitive disabilities: Building understanding of alternate assessment scoring criteria* (Synthesis Report 50). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved March 2, 2004, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis50.html

Rivera, T. C., & Ganaden, M. F. (2001). *The development and validation of a classroom environment scale for Filipinos*. Retrieved March 7, 2005, from http://www.dilnet.upd.edu.ph/~ismed/online/articles/dev/factor.htm

Roach, A. T., Elliott, S. N., & Webb, N. L. (2005). Alignment of an alternate assessment with state academic standards: Evidence for the content validity of the Wisconsin alternate assessment. *The Journal of Special Education*, *38*, 218–231.

Roeber, E. (2002). *Setting standards on alternate assessments* (Synthesis Report 42). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved January 20, 2005, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis42.html

Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Erlbaum.

Thompson, S., & Thurlow, M. (2003). *2003 State special education outcomes: Marching on*. Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved January 18, 2005, from http://education.umn.edu/NCEO/OnlinePubs/2003StateReport.html

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, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. *The Journal of Special Education*, *23*, 169–183.

Troia, G. A., & Graham, S. (2003). Effective writing instruction across the grades: What every educational consultant should know. *Journal of Educational and Psychological Consultation*, *14*, 75–89.

Vogler, K. (2002). The impact of high-stakes, state-mandated student performance assessment on teachers' instructional practices. *Education*, *123*(1), 39–51.