WCER Working Paper No. 2003-2
March 2003

Alignment Analysis and Content Validity of the Wisconsin Alternate Assessment for Students With Disabilities

Andrew T. Roach
Wisconsin Center for Education Research
University of Wisconsin–Madison
atroach@wisc.edu

Stephen N. Elliott
Department of Educational Psychology/
Wisconsin Center for Education Research
University of Wisconsin–Madison
snelliot@facstaff.wisc.edu

Norman L. Webb
Wisconsin Center for Education Research
University of Wisconsin–Madison
nlwebb@facstaff.wisc.edu

Wisconsin Center for Education Research
School of Education • University of Wisconsin–Madison
http://www.wcer.wisc.edu/

The research reported in this paper was supported by the Wisconsin Department of Public Instruction under Agency Award No. PAB00000586 and by the Wisconsin Center for Education Research, School of Education, University of Wisconsin–Madison. Any opinions, findings, or conclusions expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies, WCER, or cooperating institutions.

Alignment Analysis and Content Validity of the Wisconsin Alternate Assessment for Students With Disabilities

Andrew T. Roach, Stephen N. Elliott, and Norman L. Webb

The purpose of this investigation was to provide evidence of the validity of the Wisconsin Alternate Assessment (WAA) for assessing the academic performance of students with significant disabilities. Alternate assessments are intended for use with students who are unable to participate in general state and district assessment systems even with accommodations. In fact, alternate assessments have been described as the "ultimate accommodation" for promoting the inclusion of students with disabilities in standards-based assessment and school reform efforts (Elliott, Braden, & White, 2001).

Alternate assessments are an important component of each state's assessment system and, as such, are required to meet the federal requirements outlined in the Elementary and Secondary Education Act. Specifically, the act, as amended by the No Child Left Behind Act of 2001, mandates that state assessments "be aligned with the State's challenging content and student academic performance standards, and provide coherent information about student attainment of such standards" (Elementary and Secondary Education Act, 2002). Many states have struggled to meet these requirements because (a) the skills and concepts in the state academic standards were deemed inappropriate or irrelevant for students with significant disabilities and (b) the development of the alternate assessment was considered a special education function, precluding the involvement of general education curriculum and measurement experts.

The alignment between an assessment and the content it is meant to assess is an important piece of evidence in any validity argument. Lane (1999) outlined procedures for evaluating the validity of assessments designed to measure students' mastery of state academic standards. According to Lane, two forms of evidence are pertinent to determining the validity of these assessments: (a) the extent to which the state assessment reflects the state's academic standards and (b) the extent to which the curriculum offered to students reflects the academic standards. By establishing the alignment and curricular relevance of the WAA, this investigation provided evidence of the validity of the WAA results as a measure of students' mastery of the academic concepts and skills outlined in the Wisconsin Model Academic Standards. In addition, the investigation demonstrated the use of a formal procedure to establish the alignment of an alternate assessment.

Alternate Assessments: An Element of Inclusive Assessment Systems

According to data collected by the National Center on Educational Outcomes, many students with disabilities traditionally have been excluded from state and district-wide assessment and accountability systems (Ysseldyke & Olsen, 1997). The exclusion of these students is unfortunate because it is impossible to measure the overall effectiveness of instructional and school reform efforts without considering the performance of *all* students. To encourage the inclusion of all students in assessment systems, the 1997 reauthorization of the Individuals With Disabilities Education Act (IDEA '97; Individuals With Disabilities Education Act Amendments of 1997) required states to develop guidelines for the participation of students

with disabilities in state and district standardized testing. According to Thompson, Quenemoen, Thurlow, and Ysseldyke (2001), the movement toward inclusion of children with disabilities in large-scale assessments is important for a variety of reasons. Specifically, the inclusion of students with disabilities

1. provides a more accurate picture of states' and districts' educational systems;
2. allows accurate comparisons of assessment results for schools, districts, and states;
3. is necessary for students with disabilities to be included in standards-based reforms;
4. discourages referrals to special education in order to exclude some children's results from public reporting and accountability;
5. promotes high expectations for what students with disabilities can learn and achieve; and
6. is essential to meet legal requirements. (Thompson et al., 2001, p. 10)

For many students with disabilities, participation in state and district assessment systems involves taking existing standardized tests with testing accommodations. Some students (perhaps .5% to 2% of the student population), however, have disabilities that make their participation in general state and district-wide tests impractical and an inaccurate measure of their academic achievements. For example, a student with a visual impairment may have difficulty completing a test with many visually presented items. A student with Down's syndrome may be unable to understand and respond to items on the same test. For cases such as these, IDEA '97 required states to (a) create and implement alternate assessment systems by July 1, 2000 and (b) include the performance of students participating in alternate assessments in public accountability reporting.

The mandate to create alternate assessments has led states to propose a variety of methods for assessing students with significant disabilities. According to a survey of state special education directors conducted by Thompson and Thurlow (2000), the most common element in alternate assessments is systematic observation (43 of 50), followed by analysis of existing data (32 states), interviews and surveys (27 states), portfolios (26 states), and testing or behavior rating scales (23 states). As the survey responses indicated, states are using multiple data collection methods to increase the validity of their alternate assessment systems. Moreover, many states' alternate assessment systems are in flux as modifications are made to respond to Title I reviews and increase the reliability and validity of inferences based on the alternate assessment results.

Although the alternate assessment methods implemented by different states demonstrate a diversity of approaches, they are strikingly similar in their reliance on teachers' judgments of students' academic progress. Thus, alternate assessments may be subjected to more questions about their reliability and validity than standardized multiple-choice tests, whose objectivity is often assumed. Many researchers have demonstrated, however, that teachers can provide valid and reliable ratings when they are given a meaningful structure for communicating their knowledge about students' achievement (Demaray & Elliott, 1998; Hoge & Coladarci, 1989; Meisels, et al, 2001).

Enhancing Alternate Assessment in Wisconsin

In Wisconsin, the original alternate assessment involved a review of student performance similar to what might typically be part of a reevaluation procedure or an individualized education program (IEP) process. The Wisconsin Department of Public Instruction (1998) stated that the alternate assessment could consist of any of the following elements: school records; the most recent evaluation data; formal and informal assessments conducted by team members; reports by parents, general education teachers, and special education teachers; classroom work samples; and other information available to the IEP team (Elliott, 2001). In addition, for the IEP review process to be considered an alternate assessment, it had to be (a) a recent, representative, and comprehensive review of student performance," (b) conducted in the same general time frame as statewide large-scale testing, and (c) aligned with the state's general education standards.

Although this approach to alternate assessment appeared to meet the IDEA '97 guidelines for the participation of students with disabilities in assessment, some educators and policy makers identified concerns with having students involved in primarily idiographic assessments. As Thurlow et al. (1996) expressed it, "The primary problem with this approach is that attainment of IEP goals cannot be easily aggregated for accountability purposes and IEP goals do not serve as a total curriculum for a student." Moreover, because functional and adaptive behaviors are often the focus of IEP goals for students with significant disabilities, many alternate assessments (under the original approach) would not have reflected the range of knowledge and skills identified by Wisconsin's Model Academic Standards.

In response to these concerns and the questions raised by a Title I review completed by the U.S. Department of Education in February 2001, Wisconsin began the process of designing and implementing an enhanced alternate assessment that will provide more structure to teachers, clearer alignment to the state's academic standards, and more manageable data on students' performance. This enhanced version of the WAA consists of a behavior rating scale based on the state's alternate performance indicators (APIs), a downward extension of the academic standards. In addition, the WAA includes an overall scoring continuum for each core subject area (i.e., reading, language arts, math, social studies, and science), which allows student performance to be categorized in a manner similar to the proficiency levels used to describe students' performance on the Wisconsin Knowledge and Concepts Examinations (WKCE).

Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett's (2001) evaluation of the Work Sampling System (WSS) provides empirical support for the assessment technology used in the enhanced WAA. Like the WAA, the WSS uses a checklist, collection of work samples, and a scoring summary to describe students' academic performance. Meisels et al. found that the WSS was "a dependable predictor of achievement ratings in kindergarten to Grade 3" (p. 91). Moreover, the authors contend their results suggest teacher ratings of performance can serve as a viable alternative to norm-referenced, multiple-choice assessments. A review of the literature on teacher judgments of academic performance by Hoge and Coladarci (1989) provides further support for the validity of teachers' ratings. Specifically, Hoge and Coladarci found direct teacher judgments (i.e., ratings that entailed an explicit link between criterion and judgment) yielded a median correlation of .69. In the same review, studies that included indirect teacher judgments (i.e., ratings of student achievement without explicit definition of the construct to be evaluated) produced a median correlation of .62. In both cases, "the correlations certainly

exceed[ed] the convergent and concurrent validity coefficients normally reported for psychological tests" (p. 308).

To coordinate creation of an enhanced WAA that would provide valid and reliable results, the DPI convened a WAA Leadership Team. This leadership team focused its efforts on achieving the following four objectives: (a) creating standardized materials and scoring guidelines; (b) training teachers (and other stakeholders) in implementing the WAA; (c) developing support materials to assist teachers in administering and scoring the WAA; and (d) gathering validity and reliability evidence to ascertain and, if necessary, improve the WAA's technical qualities. The present study represents part of the research conducted to address these objectives.

Validity: An Essential Characteristic of Good Assessments

Validity refers to the adequacy and appropriateness of the interpretations made from assessments with regard to a particular use (Elliott, Braden, & White, 2001). Validity is of paramount concern in the development and selection of an educational or psychological test. A test is only useful and meaningful if it is valid for its proposed uses. Elliott, Braden, and White (p. 21) outline the following aspects of validity identified by leading measurement experts (Airasian, 1994; Linn & Gronlund, 1995):

- Validity is concerned with the question, "To what extent will this assessment information or test score help me make appropriate decisions?"

- Validity refers to decisions that are made from assessment information, not the assessment approach or test itself.

- Validity is a matter of degree; it does not exist on an all-or-nothing basis. Test consumers should think of assessments in terms of categories: highly valid, moderately valid, and invalid.

- Validity involves an overall evaluative judgment. It requires an evaluation of the degree to which interpretations and uses of assessment results are justified by supporting evidence.

Likewise, *The Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999) define validity as "the degree to which the accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p. 184). Reflecting this definition, the *Standards* treat validity as a unitary concept. In other words, there are not separate forms of validity (e.g., content validity or construct validity); instead, the validity of test interpretations involves collection and analysis of different forms of validity evidence: (a) evidence based on test content; (b) evidence based on response processes; (c) evidence based on relationships to other variables; (d) evidence based on internal structure; and (e) evidence based on the consequences of testing. Table 1 outlines the evidence for validity that was collected by the WAA Leadership Team during the field trials of the WAA rating scale. The evidence utilized in the current investigation is highlighted (shaded cells) in Table 1.

Table 1
Overview of the Validity Evidence in the Preliminary WAA Study

Sources of validity evidence	WAA study #1 research questions	Related evidence collected in WAA study #1
Evidence based on test content	Does the WAA adequately measure the skills and concepts that comprise the curriculum and instruction of students with significant disabilities? Does the WAA adequately measure the concepts and skill areas represented in Wisconsin's Model Academic Standards?	Importance ratings of WAA items by WAA Leadership Team and Field Test Teachers Rates of items used and indicated IEP-aligned during preliminary field test
		Expert panel's ratings as part of WAA alignment study
Evidence based on response processes	Are teachers' and other educators' interpretations and subsequent ratings on WAA items consistent with the intended interpretation of scores?	Teachers' responses to WAA post-administration questionnaire
		Information on WAA process from complete case studies
Evidence based on internal structure	Does the internal structure of the WAA items conform to the underlying factor structure, which is based on the state's academic standards?	Expert panel's ratings as part of WAA alignment study
Evidence based on relations to other variables	What is the relationship between WAA ratings and other measures of students' academic progress (e.g., the Academic Competence Evaluation Scales [ACES])? What is the relationship between WAA ratings and ratings on a measure of social behavior (e.g., Social Skills Rating System [SSRS])?	WAA, ACES–Teacher Form, and SSRS–Teacher Form ratings for individual students
Evidence based on the consequences of testing	Do teachers and parents endorse the WAA as (a) contributing to greater access to the general education curriculum for students with significant disabilities; (b) providing information that contributes to instructional planning for those students; and (c) providing important information on students' present level of performance?	Teachers' and parents' responses to WAA post-administration questionnaire

Alternate Assessments: Measures of Access to the General Curriculum

IDEA '97 clearly mandates that students with disabilities have access to the general education curriculum and academic standards. Specifically, one of the final regulations under

IDEA '97 (34 C.F.R. § 300.347) requires that students' IEPs include consideration of how the student will access the general education curriculum. Moreover, this regulation further requires that (a) all students participate in state and district-wide assessments; and (b) all students have opportunities and instruction that allow them to make progress toward state and district academic standards.

This emphasis on attaining academic achievement represents a dramatic departure from the curriculum and inclusion practices that traditionally have been implemented with many students with significant disabilities. Early considerations of mainstreaming and *least restrictive environment* (LRE) often focused on the socialization and self-esteem benefits for students with significant disabilities. More recent practices have maintained the focus on relationships and self-concept while adding an emphasis on exposure to the general curriculum and the broader school experience (Ford, Davern, & Schnorr, 2001). IDEA '97, however, demands an even greater access to the general education curriculum, according to Pugach and Warger (2001):

> Although the law still maintains the right of each student with disabilities to an individually referenced curriculum, outcomes linked to the general education program have become the optimal target. It is no longer enough for students with disabilities to be present in a general education classroom.

Instead, students must have instruction and accommodations that promote their progress, no matter how modest, toward the educational expectations of the larger student population.

A related concern has been the content and focus of each state's alternate assessment processes. Specifically, test developers and policy makers must determine if assessments for students who are unable to participate in the general assessment systems should be focused on "the content standards (or core learning outcomes) identified for all students; or conversely whether alternate assessments should be based on a separate, more 'functional' set of learner outcomes" (Kleinert & Kearns, 1999, p. 101). If alternate assessments are intended to measure the most salient elements of curriculum and instruction for students with significant disabilities, then an argument can be made that these tests should focus on functional and adaptive behaviors. However, if the alternate assessments are intended to function as one element of a larger accountability system and to measure progress toward the same educational expectations as those applied to the larger student population, then a state's general education academic standards should form the foundation for the alternate assessment. IDEA '97 seems to provide support for the design of alternate assessments as an extension or modification of states' standards-based assessment systems. However, "acknowledging that a central purpose of large-scale assessments is to measure major, agreed-upon outcomes over time does not take away from extensive and ongoing learning that is not captured in these assessments" (Ford et al., 2001, p. 214). Indeed, much as we do not expect multiple-choice standardized tests to measure the entire scope of curriculum and instruction provided to general education students, we should not expect alternate assessments to reflect every element of the school experiences of students with significant disabilities.

Alignment Between Standards, Assessments, and Classroom Practices

Effective schooling is based on the coordination of three components of the educational environment: curriculum, instruction, and assessment (Elliott, Braden, & White, 2001; Webb, 1997; Webb, Horton, & O'Neal, 2002). The process of coordinating these elements is called *alignment* and is the foundation of standards-based education reform. Alignment is the extent "to which expectations and assessments are in agreement and serve in conjunction with one another to guide the system toward students learning what they are expected to know and do" (Webb, Horton, & O'Neal, 2002, p. 1). The development and implementation of large-scale assessment programs represent one approach to aligning classroom instruction with state curriculum standards.

Webb (1997) outlined three methods for determining the alignment between the policy elements of curriculum, instruction, and assessment systems: (a) sequential development, (b) expert review, and (c) document analysis. *Sequential development* involves creation and acceptance of one policy element, which subsequently serves as a "blueprint" for the creation of additional policy elements. For example, a state or district might develop academic standards for mathematics that provide guidance for the selection of a new performance-focused mathematics curriculum and the development of performance-based mathematics assessments. The process of *expert review* involves the convening of a panel of content experts to review the policy elements and determine the extent of their alignment. *Document analysis* involves the coding and analysis of documents that represent the different policy elements. By integrating these three methods, test developers and education policy makers can increase the quality of the alignment process (Webb, 1997).

Sequential development, expert review, and document analysis each contributed to the creation and validation of the WAA. Alternate performance indicators, which were developed based on Wisconsin's academic standards, served as the framework for the development of the original pool of 281 items for the WAA rating scale. Expert review (i.e., review by the WAA Leadership Team) was used to analyze and rate the importance of each item as an educational outcome for students with significant disabilities. This process resulted in the elimination of 153 items from the original pool of items. In addition, expert review and document analysis, conducted according to Webb's (1997) methods for determining alignment between policy elements, were also used in completion of the WAA Alignment Institute held on June 13 and 14, 2002.

Webb (2002) and Webb et al. (2002) represent two applications of Webb's method for analyzing the alignment of assessments and curriculum standards. In these examples, teams of curriculum experts were trained to use a collection of analytic tools and heuristics to rate assessment systems and academic standards on the following criteria (outlined in more detail in Table 2 below): (a) depth-of-knowledge consistency; (b) categorical concurrence; (c) range-of-knowledge correspondence; (d) balance of representation, and (e) source of challenge.

Table 2
Criteria for Evaluating Alignment Between Assessments and Standards

Criterion	Definition
Categorical concurrence	Indicates if the same or consistent categories of content appear in both standards and assessment.
Depth-of-knowledge consistency	Indicates if what is elicited from students on an assessment is as demanding cognitively as what students are expected to know and do as stated in the standards.
Range-of-knowledge correspondence	Indicates whether a span of knowledge expected of students by a standard is the same as, or corresponds to, the span of knowledge needed by students to correctly answer the assessment item or activity.
Balance of representation	Indicates the degree to which one curriculum objective is given more emphasis on the assessment than another.
Source of challenge	Used to identify items on which the major cognitive demand is inadvertently placed and is other than the targeted curriculum skill, concept, or application. Item characteristics may cause some students to get an item partially or totally incorrect, even though they have the understanding and skills being assessed.

Consideration of the extent to which classroom instruction corresponds to academic standards and assessment systems is also an element in evaluating alignment. In the case of the WAA field trial, teachers provided a rating of "not applicable" (NA) for items outside the scope of instruction and curriculum appropriate for their students. Moreover, as part of the WAA field testing, teachers indicated which WAA items were IEP-aligned (i.e., items that represented goals and objectives on the student's IEP). Analysis of this information provided insights into the correspondence between the state's academic standards, the WAA, and students' classroom experiences.

Research Questions

By focusing on the alignment of the WAA with Wisconsin's academic standards, students' IEPs, and curriculum, on the one hand, and on parents' and teachers' reactions to the WAA, on the other, the current investigation provided content and consequential evidence for the validity of the WAA. As Table 1 indicates, however, the following questions and predictions provide only a part of the validity evidence required to make an informed judgment about the technical adequacy of the WAA.

Question #1: Does the WAA adequately measure the skills and concepts that comprise the curriculum and instruction of students with significant disabilities? The current investigation considered multiple sources of evidence in answering this question. First, importance ratings of WAA items by the WAA Leadership Team and teachers who participated in WAA field testing provided two expert reviews of items considered "essential outcomes" for students with significant disabilities. Additional data about WAA items that were representative of actual

students' IEPs and classroom curriculum were provided by (a) the percentage of field trial cases in which each WAA item was rated applicable to the student's curriculum and instruction; and (b) the percentage of field trial cases in which each item was identified as aligned with the student's IEP. The correlation of item importance ratings and actual item applicability and IEP alignment was considered evidence of the correspondence between the curricular and instructional priorities driving the development of the WAA development process and the curriculum and instruction in students' actual classrooms. Those items considered important by the WAA Leadership Team and teachers in the field-testing group (i.e., the "expert reviews") were expected to be the items most frequently rated (i.e., given a rating other than "not applicable") by teachers and endorsed as IEP-aligned during implementation.

In addition, survey responses of teachers and parents of students in the field trial provided additional insight into the correspondence between the collection of WAA items and the content of students' curriculum and instruction.

Question #2: Does the WAA adequately measure the concepts and skill areas represented in Wisconsin's Model Academic Standards? The ratings of the expert panel that participated in the WAA Alignment Institute provided information about the correspondence between WAA items and the Wisconsin Model Academic Standards. The expert panel's responses were expected to indicate that the WAA generally conforms to Webb's (1997) model for alignment of assessments and curriculum expectations. Specifically, we predicted that the expert panel's ratings would indicate that each WAA subject domain scale met the criteria for categorical concurrence, range of knowledge, and balance of representation.

On the other hand, we expected the panel responses to indicate a low overall depth-of-knowledge rating for the WAA subject domain scales. The low overall depth-of-knowledge rating would represent a departure from previous alignment studies using expert panel ratings (Webb, 2002; Webb et al., 2002). Although it is desirable that depth-of-knowledge ratings for curriculum objectives and assessment items be similar, items on alternate assessments are believed to generally demand less depth of knowledge than items in the general education academic standards and on the corresponding large-scale assessment. Thus, although WAA items represent the range of concepts and skills outlined in the state academic standards, these items are presented at a lower level of complexity or prerequisite skill that allows access for students with significant disabilities.

In addition to the results from the alignment process, students' teachers completed two additional ratings of student's academic functioning—the Academic Competence Evaluation Scales (ACES) and the Academic Competence scale on the Social Skills Rating System (SSRS). The converging and diverging relationship between these measures and performance on the WAA provided an additional index of the extent to which the WAA measured the core academic domains. Finally, responses on teacher and parent surveys also provided additional information about the correspondence of the WAA items to the state academic standards.

Given the stated purpose of the investigation and the two research questions, a method was needed whereby the content validity and instructional utility of the WAA, as well as its alignment to Wisconsin's Model Academic Standards, could be evaluated. Such a method is described in the next section of this paper.

Method

The current investigation integrated data from two components of the overall WAA validation effort. WAA item ratings, teachers' and WAA Leadership Team members' item importance ratings, and teachers' and parents' perceptions of the WAA process were gathered during the field trial study of the WAA conducted during the spring of 2002. Additional information about the alignment of the WAA instrument to Wisconsin's Model Academic Standards was collected during the WAA Alignment Institute conducted on June 13 and 14, 2002.

WAA Field Trial Study

Participants

Teachers. Special education teachers (N = 40) from elementary and secondary schools across Wisconsin participated in the WAA Field Trial Study. Overall, teachers who participated in field testing had extensive teaching experience (M = 13.5 years; SD = 9.85), and the majority (56.4%) had completed advanced degrees (e.g., Master's of Education). Each teacher was responsible for obtaining consent for participation from students' parents and (if possible) the students themselves. All teachers consented to participate and received monetary compensation (a $240 honorarium) to attend a 1-day WAA administration training session and to complete a case study on the WAA process.

Students and their parents. Teachers identified one student with a significant disability (N = 40) to participate in the WAA field test. Teachers used the Wisconsin Alternate Assessment Participation Checklist (see Appendix A) to determine the appropriateness of using the WAA with their students. The participation checklist requires a student's IEP team to verify that the student meets the following criteria for each subject domain:

1. The student's curriculum and daily instruction focus on knowledge and skills *significantly different* from those represented by the state's academic standards for students of the same age.

2. The student's present level of educational performance (PLOEP) significantly impedes participation in and completion of the general education curriculum even with significant program modifications.

3. The student requires extensive direct instruction to accomplish the acquisition, application, and transfer of knowledge and skills.

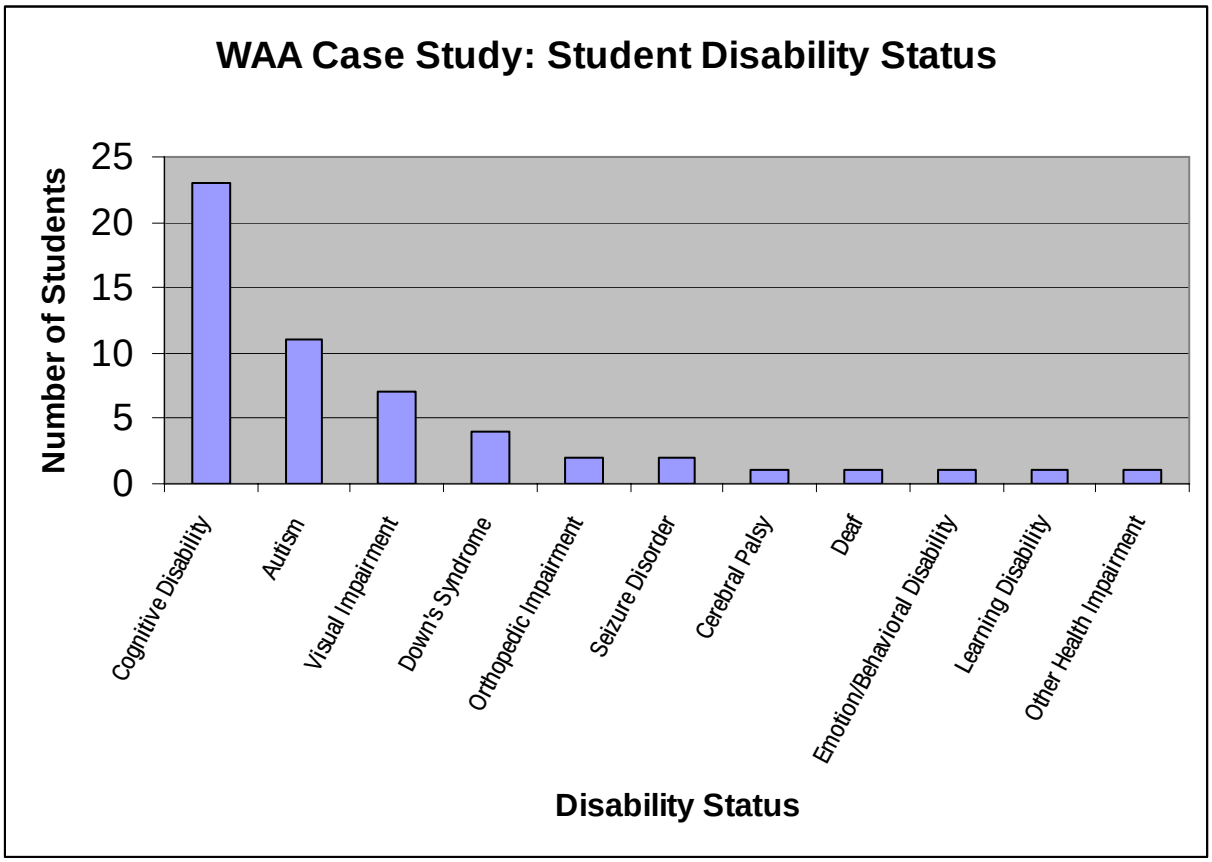
4. The student's difficulty with the regular curriculum demands is primarily due to his or her disabilities, rather than to extensive absences unrelated to the disability or social, cultural or environmental factors.

The final sample of student participants (N = 40) included 14 females and 26 males. These students represented individuals with various disabilities (see Figure 1). The students also represented varying grade levels: 18 elementary students, 13 middle school students, and 9 high

school students. The mean performance for the field trial participants on the ACES Academic Skills subscale (M = 34.7) indicated the group's academic functioning was within the 1st decile when compared to the national norming sample of their grade-level peers. Approximately a quarter of the students involved were described as having two or more disabilities (thus, the number of students in Figure 1 adds to more than 40). Upon completion of the WAA process, students' parents were asked to complete a short questionnaire regarding their impressions of the assessment. A majority (N = 32) of the students' parents responded to the questionnaire.

Participant consent. Teachers, parents, and students were provided with information regarding the nature of the WAA Field Trial Study, potential drawbacks and benefits of their participation, and provisions to protect their confidentiality. Teachers and parents signed a written consent form, indicating they understood the nature of the research and the ramifications of their participation. In addition, if possible, informed consent was obtained from the students themselves prior to their participation in the study. Participants were further informed that their participation was voluntary and that they could withdraw from the investigation at any time. For teachers, withdrawal from the investigation before completing the case study resulted in forfeiture of a portion of their honorarium.

Figure 1. Student disability status of field trial study participants.



Instruments

The Wisconsin Alternate Assessment (WAA). The WAA is a part of the Wisconsin Student Assessment System and is designed to assess the educational performance of students with disabilities who cannot meaningfully participate in the general test (WKCE), even with accommodations. The WAA used in the spring 2002 field trial consisted of 128 Likert-scale items that required teachers to rate students' performance of a skill or understanding of a concept on a 4-point scale ranging from "non-existent" (0) to "proficient/generalized" (3). In addition, teachers could rate items "not applicable" (NA) if they determined the item was "not relevant to the student's educational needs." The WAA items were organized into five scales that assessed students' performance in each core academic subject: Reading, Language Arts, Math, Social Studies, and Science.

The WAA Leadership Team developed items for the WAA during the summer of 2001. These items were based on DPI's alternate performance indicators. The APIs represent a downward extension of Wisconsin's Model Academic Standards. Further information about the item development and selection process is provided in the subsequent description of the procedures used in this investigation.

Using the results from the spring 2002 field trial of the WAA, we generated statistics for each of the separate WAA scales to describe aspects of central tendencies, score distributions, and reliability (i.e., internal consistency and standard error of measurement). These results should be interpreted with caution given that they are based on the ratings of a relatively small sample (40 students). Table 3 below presents a summary of statistics that describe the technical characteristics and utilization of each WAA content domain scale resulting from the spring 2002 field trial.

Academic Competence Evaluation Scales (ACES). The ACES (DiPerna & Elliott, 2000) is a multirater assessment device designed to measure academic skills (i.e., reading/language arts, mathematics, and critical thinking) and academic enablers (i.e., motivation, study skills, engagement, and interpersonal skills). Only the ACES–Teacher Form was used in this study. Teachers generally provide two ratings for each item on the ACES: (a) the proficiency (or frequency) of a behavior, skill, or attitude and (b) the importance of that behavior, skill, or attitude for classroom achievement. Because the research focus of the current investigation was on the students' relative level of academic achievement, field trial teachers completed only the proficiency ratings.

The ACES *Manual K-12* (DiPerna & Elliott, 2000) provides substantial evidence for the reliability and validity of the ACES–Teacher Form. Internal consistency (alpha = .94 to .99) and test-retest reliability (r = .88 to .97) suggest that the ACES subscales are highly reliable measures of academic skills and enablers. In addition, the Academic Skills scale was highly correlated with standardized test performance (Iowa Test of Basic Skills, r = .80 [reading] and r = .86 [mathematics]) and with students' GPAs (r = .90). Although the validity evidence for the ACES exceeds the level generally found on behavior rating scales, the authors do caution that some evidence (i.e., interrater agreement) should be "interpreted with caution given the rather small and varying sample size" in validity studies (p. 80).

Table 3
Descriptive Statistics for WAA Subject Domain Scales

Descriptive and statistical indices	Reading	Language arts	Mathematics	Science	Social studies
Total number of items / maximum possible score	23 / 69	26 / 78	29 / 87	21 / 63	29 / 87
Mean raw score	24.3	28.6	28.7	15.1	27.7
Median	20.0	24.0	23.5	14.5	26.0
Standard deviation	19.0	19.3	22.4	12.5	21.8
Standard error of measurement	3.7	3.7	3.6	3.7	3.7
Percentiles					
25th percentile	6.0	12.0	7.8	3.0	7.0
50th percentile	20.0	24.0	23.5	14.5	26.0
75th percentile	44.0	49.0	49.0	23.0	43.0
Performance levels					
Prerequisite skill 1	30.8%	30.8%	28.2%	53.8%	46.2%
Prerequisite skill 2	30.8%	33.3%	30.8%	25.6%	25.6%
Prerequisite skill 3	28.2%	33.3%	35.9%	7.7%	17.9%
Prerequisite skill 4	10.3%	2.6%	2.6%	0%	0%
Coefficient alpha	0.98	0.97	0.98	0.96	0.98

Social Skills Rating System (SSRS). The SSRS (Gresham & Elliott, 1990) is a multirater assessment device providing norm-referenced information regarding students' social behaviors, problem behaviors, and academic competence. Only the SSRS Teacher Form was used in the current investigation. The SSRS asks teachers to provide two ratings for each item: (a) the proficiency (or frequency) of a behavior, skill, or attitude and (b) the importance of that behavior, skill, or attitude within the classroom. The *Social Skills Rating System Manual* (Gresham & Elliott, 1990) provides substantial information regarding the technical adequacy of the SSRS. The coefficient alphas for the Social Skills (.83 to .94), Problem Behaviors (.81 to .88), and Academic Competence scales (.95) are consistently high, as is test-retest stability for teacher ratings (Social Skills = .85, Problem Behavior = .93, and Academic Competence = .84).

WAA teacher and parent surveys and case studies. A short written survey was administered to teachers and parents to gather their perceptions concerning the acceptability and utility of the WAA process. A copy of both surveys is included in Appendix B. As part of their case studies, teachers also completed written responses to a series of open-ended prompts concerning their impressions of the WAA assessment process.

Procedure

Item importance ratings, completed by the WAA Leadership Team, were used to select the 128 items that appeared on the WAA rating scale. Initial item development involved revising the APIs to include more objective behavioral descriptions and to enhance the likelihood that students' skills and knowledge could be demonstrated in a variety of ways. This process resulted in the creation of 281 items. To reduce the number of items for the WAA rating scale, leadership team members (N = 9) were given a list of the API-based items and asked to rate the importance of each item in contributing to students' learning and academic progress. Item ratings were based on a 4-point scale, ranging from 0 ("not important") to 3 ("critical/very important"). Following collection of the leadership team's ratings, the mean overall rating for each item was calculated. The first and second authors used the mean overall ratings and a series of decision-making rules to reduce the original list of items. The process used in determining the items that were included on the final scale is outlined in Table 4.

Table 4
WAA Item Selection Process and Decision-Making Rules

Decision-making rule	Result
Delete items with mean importance ratings below 2.0	194 items remaining (from the original 281 developed from APIs)
Delete items with mean importance ratings below 2.2	139 items remaining
Revise list of remaining items to • Be representative within each subscale (at least 50% of original APIs) • Reduce redundancy with items in other subject domains • Eliminate poorly written items	128 items selected for inclusion on WAA rating scale

*Special education teachers (*N *= 40) field-tested the WAA rating scale with one of their students with significant disabilities*. The field trial occurred during the spring of 2002. Each teacher completed a case study for his or her student that included the following elements: (a) a completed WAA rating scale, (b) a WAA Participation Checklist (see Appendix A), (c) narrative responses to questions about the WAA process, (d) collected work samples, (e) the student's most recent IEP, (f) the ACES, (g) the SSRS, and (h) a signed parent consent form. Teachers participated in a 1-day training session on the administration and use of the WAA prior to completion of their case studies.

In addition to completing a case study, teachers were required to have another credentialed staff member familiar with the student rate each IEP-aligned item and the student's overall performance level scores. An interrater agreement of 80% for IEP-aligned items in each subject domain had to be attained before the proficiency score summary for that area was considered reportable. An agreement of 100% was required before overall performance level scores for each subject domain were reportable.

Data on the applicability of each WAA item was gathered from the completed WAA rating scales. In addition, teachers were asked to indicate which WAA items were aligned with their students' IEPs. Frequency of IEP alignment also was calculated for each of the WAA items using data from the field trial cases. Following completion of their case studies, teachers were mailed a survey that asked them to rate the importance of each item in contributing to the learning and academic progress of students with significant disabilities. Item ratings were based on a 4-point scale, ranging from 0 ("not important") to 3 ("critical/very important").

Research Questions and Statistical Analysis

Question #1: Does the WAA adequately measure the skills and concepts that comprise the curriculum and instruction of students with significant disabilities? The variables considered in answering Question #1 were (a) the percentage of items rated applicable and (b) the percentage of items rated aligned to students' IEP goals. A correlational design was used to examine the strength of the relationship between the item importance ratings of WAA Leadership Team members and field trial teachers and the frequency of item applicability and IEP alignment. Pearson correlations between the leadership team's and field trial teachers' mean item importance ratings and rates of IEP alignment (i.e., percentage of cases in which an item was marked IEP-aligned) and applicability (i.e., percentage of cases in which an item was rated applicable to the student's curriculum and instruction) were calculated for each WAA item.

Teachers and parents were also asked to rate the instructional utility of the WAA and the meaningfulness of results for describing students. Descriptive statistics (frequencies for response options and mean responses) were calculated for the following items from the teacher survey:

> Item 4: "The results of the WAA were useful to me and others who make instructional plans for students with significant disabilities."
>
> Item 8: "The student results appeared to be an accurate representation of the student's skills that were measured."

In addition, descriptive statistics (frequencies for response options and mean responses) were calculated for the following items from the parent survey:

> Item 4: "I was confident in the results about my child's functioning that the teachers provided to me."
>
> Item 5: "I believe the time spent by teachers conducting an alternate assessment is important to their teaching of my child."

Question #2: Does the WAA adequately measure the concepts and skill areas represented in Wisconsin's Model Academic Standards? Responses on teacher and parent surveys provided information about the correspondence of WAA items to Wisconsin's Model Academic Standards and to the academic subjects tested on the WKCE assessment administered to the general student population. Descriptive statistics (frequencies for response options and mean responses) were calculated for the following items from the teacher survey:

Item 5: "The WAA items were well aligned with the state's general education academic standards."

Item 6: "By conducting the alternate assessment, I learned more about Wisconsin's academic standards and statewide assessment system."

Descriptive statistics (frequencies for response options and mean responses) were calculated for the following items from the parent survey:

Item 2: "I was pleased to know that the assessment was aligned with the state's academic standards."

Item 3: "I think it is good that all students in the state participate in an assessment that focuses on their achievement in basic areas such as reading, writing, and math."

Additional information about the correspondence of WAA items to the Wisconsin Model Academic Standards and to the academic subjects tested on the WKCE assessment administered to the general student population was gathered during the WAA Alignment Institute conducted June 13–14, 2002.

WAA Alignment Institute

The purpose of the WAA Alignment Institute was to determine whether the WAA adequately measured the skills and concepts represented in Wisconsin's Model Academic Standards. The variables in the investigation were: (a) panel members' depth-of-knowledge ratings for content standard objectives and individual WAA items; and (b) panel members' identification of the one or two objectives that corresponded to each WAA item.

Participants

The alignment review panel (*N = 10*) consisted of special education teachers, personnel from DPI, and graduate students who participated in a 2-day WAA Alignment Institute conducted at the University of Wisconsin–Madison June 13–14, 2002.

Procedure

The alignment coding process entailed panel members' rating the WAA items and Wisconsin's Model Academic Standards using depth-of-knowledge, categorical concurrence, range-of-knowledge, and balance-of-knowledge criteria. The primary role of the panel members was to complete the following three tasks:

1. Rate the depth-of-knowledge level of each objective in the Model Academic Standards.

2. Rate the depth-of-knowledge level of each item on the WAA rating scale.

3. Identify the one or two objectives to which each WAA item corresponds.

Panel members' responses were recorded on a series of coding sheets (see Appendix C), which provided columns for (a) rating each WAA item on the depth-of-knowledge criteria; (b) indicating the corresponding objective(s) for each item; and (c) identifying potential sources of challenge (e.g., visual identification items for students who are visually impaired). WAA items were presented in random order instead of by subject domain as they appear on the WAA rating scale.

Before completing their ratings, panel members were trained to identify the depth-of-knowledge level for curriculum objectives (i.e., performance standards) and WAA items. This training included a review of the four general depth-of-knowledge levels outlined in Table 5.

Table 5
Depth-of-Knowledge Levels

Level	Description
Level 1: Recall	Level 1 includes the recall of information, such as a fact, definition, term, or simple procedure, as well as performing a simple algorithm or applying a formula.
Level 2: Skill/concept	Level 2 includes the engagement of some mental processing beyond a habitual response. A Level 2 assessment item requires students to make some decisions about how to approach a problem or activity. Keywords that distinguish a Level 2 item or task include *classify*, *organize*, *estimate*, *make observations*, *collect and display data*, and *compare data*.
Level 3: Strategic thinking	Level 3 includes items that require reasoning, planning, using evidence, and a higher level of thinking than the previous two levels. In most instances, requiring students to explain their thinking is a Level 3 attribute. Students might also be required to make conjectures or determine a solution to a problem with multiple correct answers at this level.
Level 4: Extended thinking	Level 4 includes items that require complex reasoning, planning, developing, and thinking most likely over an extended period of time. At Level 4, the cognitive demands of the task should be high, and the work should be very complex. Students should be required to make connections both within and between subject domains. Level 4 activities include designing and conducting experiments; making connections between a finding and related concepts; combining and synthesizing ideas into new concepts; and critiquing literary pieces and experimental designs.

Note. From Webb (2002). Adapted with permission.

Specific descriptions for depth-of-knowledge levels for each of the subject domains covered by the WAA were developed, using examples from previous alignment analyses conducted on large-scale assessments (Webb, 2002; Webb et al., 2002) as models (see Appendix D). Panel members rated the depth-of-knowledge levels for a series of sample items before completing their individual ratings of the curriculum objectives and WAA items using the criteria. These

practice items provided an opportunity for discussion of the criteria and "calibration" of panel members' understanding of the depth-of-knowledge rating process (Webb, 2002).

Following the "calibration" process, panel members were asked to assign a depth-of-knowledge rating to each objective (i.e., performance standard) in Wisconsin's Model Academic Standards and to each assessment item on a randomly ordered list of WAA items. If panel members had difficulty deciding between two levels for an objective or a WAA item (e.g., between a rating of 1 or 2), they were instructed to choose the higher of the two levels. During the alignment institute, the panel reached consensus about the depth of knowledge for the curriculum objectives before individually completing the ratings of the WAA items. After completing the depth-of-knowledge ratings, panel members completed the coding sheets by identifying the one or two objectives that corresponded to each WAA item.

According to Webb (2002), the alignment coding process is not designed to produce exact agreement between members of the expert panel. In fact, variance in ratings "are considered valid differences in opinion that are a result of a lack of clarity in how the objectives were written and/or the robustness of an item that may legitimately correspond to more than one objective" (p. 3).

The alignment analysis completed by panel members provided descriptive statistics for the four criteria underlying Webb's alignment model: (a) categorical concurrence, (b) depth-of-knowledge consistency, (c) range-of-knowledge correspondence, and (d) balance of representation. Webb's criteria for determining alignment between assessments and curricular expectations are outlined in Table 6 below.

Research Questions and Statistical Analysis

The expert panel's responses were expected to indicate the WAA generally conforms to Webb's model for alignment of assessments and curriculum expectations. Specifically, the expert panel was expected to indicate that each WAA subject domain scale meets the criteria for categorical concurrence, range of knowledge, and balance of representation. However, the panel responses were expected to indicate a low overall depth-of-knowledge rating for the WAA items and subject domain scales. Because the WAA is intended to be an assessment for students with significant disabilities, it was anticipated that the items would not demand the level of mastery expected from students who take the regular large-scale assessment. Thus, the anticipated percentage of items meeting the depth-of-knowledge criteria was less than 50%.

Results

Question #1: Does the WAA adequately measure the skills and concepts that comprise the curriculum and instruction of students with significant disabilities? Data gathered during the WAA spring 2002 Field Trial Study was used to determine the extent to which the WAA adequately measures the skills and concepts that comprise the curriculum and instruction of students with significant disabilities. The following forms of data were considered in establishing this relationship: (a) the percentage of field trial cases in which each WAA item was rated applicable to the student's curriculum and instruction; (b) the percentage of field trial cases in which each item was identified as aligned with the student's IEP; (c) field trial teachers' and

Table 6
Summary of Webb's Criteria for Alignment

Criteria	Description
Categorical concurrence	An assessment must have at least six items measuring content for each standard in order to demonstrate an acceptable categorical concurrence between the standard and the assessment. "The number of items, six, is based on estimating the number of items that could produce a reasonably reliable subscale for estimating students' mastery of content on that subscale. . . . Using a procedure developed by Subkoviak (1988) and assuming that the cutoff score is the mean and the reliability of one item is .1, it was estimated that six items would produce an agreement coefficient of at least .63" (Webb, 2002, p. 4).
Range of knowledge	At least 50% of the objectives for a standard corresponded with at least one related WAA item based on the ratings of alignment institute panel members. The range-of-knowledge criterion is based on the assumption that an assessment should test students' understanding or mastery of the majority of the knowledge (i.e., more than half the objectives) represented by any given standard (Webb, 2002).
Balance of representation	A balance index score was computed to judge the distribution of assessment items. "The balance index compares the proportion of items for each objective to the proportion if the items were evenly distributed among all possible objectives" (Webb et al., 2002). An index value of .7 or greater indicated that WAA items are distributed among all objectives to an acceptable degree.
Depth of knowledge	"For consistency between the assessment and standard . . . at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective" (Webb, 2002, p. 4). Meeting this criterion suggests a test demands adequate depth of understanding and sufficient mastery of the knowledge and skills covered in the corresponding academic standards.

leadership team members' ratings of the importance of each WAA item in contributing to the learning and academic progress of students with significant disabilities; and (d) responses to items on parent and teacher surveys concerning the instructional utility and acceptability of WAA results.

Descriptive statistics concerning the mean number of items rated applicable to field trial students' curricula and aligned with their IEPs are reported in Table 7 for each WAA subject domain scale. The results indicate the majority of items on WAA scales were identified by field trial teachers as applicable to the curriculum and instruction of their students. With the exception of the Science scale (62.4 % applicable items), approximately 75% of the items were rated applicable for each subject domain scale.

On average, approximately four items on the Reading, Language Arts, and Mathematics scales were rated as aligned to field trial students' IEP goals, objectives, and benchmarks. In comparison, fewer Social Studies (M = 1.9) and Science (M = 0.7) items were designated IEP-aligned.

Table 7
Mean Items Rated Applicable and IEP-Aligned on Spring 2002 Field Trial Cases

Subject domain	Applicable items		IEP-aligned items	
	Mean	SD	Mean	SD
Reading (23 items)	16.9	7.9	3.8	3.0
Language arts (26 items)	20.1	7.4	4.2	3.4
Mathematics (29 items)	21.2	10.3	4.1	3.6
Science (21 items)	13.1	8.0	0.7	1.1
Social studies (29 items)	20.6	9.9	1.9	1.7

Pearson correlations were calculated between the leadership team members' and field trial teachers' mean item importance ratings and the rates of IEP alignment and applicability for each WAA item. The resulting correlations are reported by subject domain (see Table 8). Across subject domains, the results indicated a strong positive correlation between field trial teachers' mean item importance ratings and the percentage of cases in which an item was rated applicable to the student's curriculum and instruction (r = .70 to .94). Moreover, correlations between the leadership team's mean item importance ratings and rates of item applicability on the separate subject domain scales were in the moderate to strong positive range (r = .35 to .73). These results suggest the WAA items considered by teachers and education leaders as the most relevant curriculum outcomes for students with significant disabilities were likely to be viewed as applicable to the curriculum and instruction of students (i.e., unlikely to be rated "not applicable" by teachers) in the WAA Field Trial Study.

With the exception of the WAA Reading scale, the results also indicated a moderate positive correlation between field trial teachers' mean item importance ratings and the percentage of cases in which an item was rated aligned with the student's IEP goals, objectives, or benchmarks (r = .41 to .52). Similarly, correlations between the leadership team's mean item importance ratings and rates of IEP alignment were in the moderate positive range (r = .28 to .59). These results suggest that the WAA items viewed by teachers and education leaders as the most relevant curriculum outcomes for students with significant disabilities were more likely to be identified as aligned with field trial students' IEPs. A similar relationship was not observed on the WAA Reading scale (see Table 8). The correlations were negligible between teachers' (r = .11) and leadership team members' (r = -.07) item importance ratings and rates of IEP alignment on the WAA Reading scale.

Table 8
Pearson Correlations Between WAA Item Importance Ratings and the Percentage of Field Trial Cases in Which WAA Items Were Rated IEP-Aligned and Applicable

	WAA items	
Importance ratings (*M, SD*)	IEP-aligned	Applicable
Reading scale		
WAA Leadership Team (2.6, .19)	*r* = -.07	*r* = .41*
Field Trial Teachers (2.1, .35)	*r* = .11	*r* = .84**
Language Arts scale		
WAA Leadership Team (2.6, .23)	*r* = .28	*r* = .71**
Field Trial Teachers (2.0, .51)	*r* = .44*	*r* = .94**
Mathematics scale		
WAA Leadership Team (2.5, .26)	*r* = .59**	*r* = .35
Field Trial Teachers (2.0, .37)	*r* = .53**	*r* = .70**
Science scale		
WAA Leadership Team (2.3, .22)	*r* = .59**	*r* = .73**
Field Trial Teachers (1.6, .35)	*r* = .41*	*r* = .74**
Social Studies scale (29 items)		
WAA Leadership Team (2.5, .22)	*r* = .38*	*r* = .49**
Field Trial Teachers (2.2, .43)	*r* = .52**	*r* = .86**

Note. WAA Leadership Team (*N = 9)* and Field Trial Teachers *(N = 40).*
* $p < .05$.
** $p < .01$.

Field trial teachers' mean responses to survey items concerning the instructional utility and acceptability of WAA results are reported in Table 9. Overall, field trial teachers generally endorsed the WAA as an accurate measure of their students' academic skills (Item 8). However, support for the instructional utility of the WAA (Item 4) was less strong, with the field trial teachers' mean response in the neutral range.

Field trial parents' mean responses to survey items concerning the instructional utility and acceptability of WAA results are reported in Table 10. Field trial parents generally indicated strong support for the acceptability of WAA results as an indicator of their students' academic functioning (Item 4). In comparison to the field trial teachers, parents were more supportive of the instructional utility of the WAA (Item 5), with a mean response in the moderate agreement range.

Table 9
Field Trial Teachers' Frequency of Responses to Survey Items

Survey item	Mean (SD)	Response options 1	2	3	4	5
Item 4: "The results of the WAA were useful to me and others who make instructional plans for students with significant disabilities."	3.2 (1.4)	8	5	10	9	10
Item 8: "The student results appeared to be an accurate representation of the student's skills that were measured."	4.2 (1.0)	2	1	2	19	17

Note. 1 = "Strongly Disagree"; 5 = "Strongly Agree."

Table 10
Field Trial Parents' Frequency of Responses to Survey Items

Survey item	Mean (SD)	Response options 1	2	3	4	5
Item 4: "I was confident in the results about my child's functioning that the teachers provided to me."	4.5 (0.7)	0	0	4	9	19
Item 5: "I believe the time spent by teachers conducting an alternate assessment is important to their teaching of my child."	3.9 (1.2)	2	1	8	7	14

Note. 1 = "Strongly Disagree"; 5 = "Strongly Agree."

In summary, results from the spring 2002 WAA Field Trial Study provide evidence for the WAA rating scale as an acceptable measure of the skills and concepts that comprise the curriculum and instruction of students with significant disabilities.

Question #2: Does the WAA adequately measure the concepts and skill areas represented in Wisconsin's Model Academic Standards? Data gathered during the WAA spring 2002 Field Trial Study and the WAA Alignment Institute were used to determine the extent to which the WAA adequately measures the skills and concepts represented in Wisconsin's Model Academic Standards. The following forms of data were considered in establishing this relationship: (a) responses to items on parent and teacher surveys concerning the WAA rating scale's focus on the state's academic standards; (b) field trial students' performance on the WAA rating scales and two additional measures of academic performance, the ACES and the SSRS–Academic Competence scale; and (c) the WAA Alignment Institute expert panel members' ratings of the alignment between the WAA and the state academic standards.

Field trial teachers' mean responses to survey items concerning the WAA rating scale's relationship to the state academic standards are reported in Table 11. Overall, field trial teachers strongly endorsed the WAA as well aligned to Wisconsin's Model Academic Standards (Item 5). Moreover, the mean field trial teachers' response indicated that completing the WAA rating scale helped them feel more familiar with the state's academic standards (Item 6).

Table 11
Field Trial Teachers' Frequency of Responses to Survey Items

Survey item	Mean (SD)	Response options				
		1	2	3	4	5
Item 5: "The WAA items were well aligned with the state's general education academic standards."	4.4 (0.6)	0	0	4	16	21
Item 6: "By conducting the alternate assessment, I learned more about Wisconsin's academic standards and statewide assessment system."	4.0 (1.1)	2	3	5	15	17

Note. 1 = "Strongly Disagree"; 5 = "Strongly Agree."

Field trial parents' mean responses to survey items concerning the relation of the WAA rating scale to the state's academic standards are reported in Table 12. Field trial parents generally indicated moderate support for the acceptability of assessing *all* students' performance in reading, math, and writing (Item 3). Parents were also moderately supportive of the alignment of the WAA to Wisconsin's Model Academic Standards (Item 5).

Table 12
Field Trial Parents' Frequency of Responses to Survey Items

Survey item	Mean (SD)	Response options 1	2	3	4	5
Item 2: "I was pleased to know that the assessment was aligned with the state's academic standards."	4.1 (1.1)	1	1	8	4	17
Item 3: "I think it is good that all students in the state participate in an assessment that focuses on their achievement in basic areas such as reading, writing, and math."	3.9 (1.3)	3	1	6	7	15

Note. 1 = "Strongly Disagree"; 5 = "Strongly Agree."

Given that the WAA rating scale was designed to measure basic knowledge and skills in the five content domains represented in Wisconsin's Model Academic Standards, it was expected to correlate strongly with other measures of academic performance. In the WAA Field Trial Study, two additional teacher judgment measures of academic functioning—the ACES and the SSRS Academic Competence scale—were used to examine the concurrent and convergent validity of the WAA. The SSRS Social Skills and Problem Behavior scales provided measures of behavior that were expected to be less strongly related to the pre-academic and academic skills measured by the WAA. Table 13 addresses the degree to which the various WAA scales measure similar skills and behaviors measured by well-established and nationally normed rating scales completed by teachers.

An examination of the correlations in Table 13 indicates that the magnitude of the correlations among measures (regardless of whether the WAA total raw score or performance level score was used) was generally in the moderate positive range (i.e., r = .30 to .60). The moderate correlations between the measures, ACES with WAA and SSRS with WAA, suggest the instruments measure similar, if not the same, underlying knowledge and skills. It also can be observed that the correlations for the Reading, Language Arts, and Mathematics scales are consistently higher than the correlations for the Science and Social Studies scales. This result was expected given that the item content of the ACES and SSRS–Academic Competence scale is primarily concerned with reading and mathematics.

Additional information about the extent to which the WAA adequately measures the skills and concepts represented in Wisconsin's Model Academic Standards was provided by an analysis of the following data gathered as part of WAA Alignment Institute: (a) panel members' rating of the depth-of-knowledge level of each objective in the academic standards, (b) panel members' ratings of the depth-of-knowledge level of each item on the WAA rating scale; and (c) the objectives identified by panel members as corresponding with each WAA item.

Table 13
Correlation Matrix of Concurrent Rating Scale Measures of Students' Academic Skills

	Reading (Raw score/ performance level)	Language arts (Raw score/ performance level)	Mathematics (Raw score/ performance level)	Science (Raw score/ performance level)	Social studies (Raw score/ performance level)
ACES Total academic skills	.55 /.60	.51 /.50	.54 /.52	.43 /.37	.54 /.42
ACES Reading/ language arts	.47 /.56	.40 /.31	.50 /.45	.46 /.35	.57 /.44
ACES Mathematics	.40 /.40	.34 /.37	.45 /.42	.33 /.32	.46 /.36
ACES Critical thinking	.58 /.56	.58 /.49	.52 /.47	.46 /.35	.52 /.36
SSRS Total social skills	.63 /.57	.61 /.46	.52 /.48	.57 /.33	.60 /.25
SSRS Total problem behaviors	.41 /.35	.38 /.28	.37 /.26	.55 /.43	.48 /.40
SSRS Total academic competence	.47 /.46	.42 /.52	.54 /.46	.39 /.28	.54 /.21

Note. (a) All correlations > .40 are statistically significant at the $p < .01$ level. (b) The SSRS Academic Competence scale only measures reading and mathematics.

Alignment Institute panel members reached consensus on the depth-of-knowledge level ratings for the objectives (i.e., performance standards) for the Reading, Language Arts, and

Mathematics scales. Because of time constraints, the panels' most common depth-of-knowledge rating (i.e., the mode) was assigned to the objectives in social studies and science. Panel members independently rated the depth-of-knowledge levels of individual WAA items with moderate to high consistency. The average measure of intraclass correlations (Shrout & Fleiss, 1979), which compared the ratings of the 10 reviewers, was consistently .85 or higher (Table 14).

Table 14
Reliability of Depth-of-Knowledge Level Ratings of WAA Items

Subject domain	Number of reviewers	Number of items	Alpha	95% confidence interval
Reading	10	23	.95	.92–.98
Language arts	10	26	.94	.89–.97
Mathematics	10	29	.90	.83–.95
Science	10	21	.86	.74–.93
Social studies	10	29	.89	.82–.94

Categorical concurrence. One aspect of alignment between standards and assessments is whether both documents address similar content. The categorical concurrence criterion provides a very general analysis of the content match. Analysis of the results from the Alignment Institute indicates that the WAA scales demonstrated varying levels of categorical concurrence across subject domains (Table 15). Academic standards with an acceptable level of categorical concurrence were judged by panel members to have at least six corresponding items on the WAA scale. Those items could, if necessary, "produce a reasonably reliable subscale for estimating students' progress" on the specific skills and concepts outlined on the corresponding academic standards (Webb, 2002, p. 4). The categorical concurrence of academic standards with five corresponding WAA *i*tems, according to panel members' ratings, was considered weak.

The WAA Language Arts and Science scales achieve categorical concurrence for less than 50% of academic standards. Although this result is less than optimal, it is important to emphasize that attaining the categorical concurrence criterion only indicates there are sufficient items to create subscales within a particular academic area. Because the WAA reports only total scale scores for each subject domain, meeting this criterion was desirable but not necessary for determining the validity and usability of the assessment.

Range-of-knowledge consistency. When standards and assessment are aligned, they cover a comparable breadth of knowledge. The range-of-knowledge criterion measures the number of objectives (i.e., performance standards) with at least one corresponding assessment item. At least 50% of the objectives for a standard must have at least one corresponding WAA item to meet this criterion. When 40%–50% of a standard's objectives were rated as corresponding to an item, the range of knowledge was considered weak (Webb, 2002).

Table 15
Categorical Concurrence for WAA Subject Domain Scales

Subject domain	Academic standards	Objectives (performance standards)	Number of WAA items	Number of hits (mean)	% of academic standards acceptable *
Reading	1	4	23	33.1	100%
Language arts	5	14	26	39.3	40%
Math	6	32	29	42.2	50%
Science	8	41	21	26.2	13%
Social studies	5	47	29	39.1	60%

*Includes standards with weak categorical concurrence.

The results of the WAA alignment indicate the range-of-knowledge criterion was met for the Reading and Language Arts scales (Table 16). According to the panel members' ratings, 100% of the Reading and Language Arts objectives (i.e., performance standards) had a corresponding WAA item. The mean number of objective hits (4.2 for Reading and 1.1 for Language Arts Standard F) indicates that some panel members rated items as corresponding to the larger content standard without indicating a specific corresponding objective.

The range-of-knowledge criterion was also met for the Mathematics, Social Studies, and Science scales, although the panel members' ratings indicate the WAA items only weakly met the criterion for the majority of standards. This result is attributable to the numerous academic standards for these subject domains and the relative brevity of the WAA subject domain scales. For example, the low levels of range-of-knowledge consistency between the Social Studies Standards B and E and the WAA Social Studies scale reflect the numerous objectives for those standards. Although the panel members' ratings indicated multiple items on the WAA Social Studies scale corresponded to Standards B and E, the range of item hits was not expansive enough to strongly meet the range-of-knowledge criterion.

Balance of representation. Whereas the range-of-knowledge criterion measures an assessment's breadth of content, the balance of representation is related to the degree of emphasis. As stated by Webb (2002), "The underlying assumption is that items should be evenly spread among the objectives for a standard. . . . If an objective is to be weighed more heavily on an assessment, teachers and students should be informed of this emphasis" (p. 14). The analysis of the balance of representation included the use of the *balance index* developed by Webb, which provides scores ranging from 0 (a large percentage of items correspond to one or two objectives) to 1 (equal distribution of the items). Index values of .7 or higher indicate that the WAA items are distributed among all the objectives to an acceptable degree and that the balance-of-representation criterion was met. The balance of representation for all the subject domain scales was rated as acceptable (see Table 17). This result is attributable to the concise format of the WAA rating scale in comparison to many individually administered standardized tests. The

Table 16
Range of Knowledge (ROK) for WAA Subject Domain Scales

Subject domain	Academic standard	Objectives rated (mean)	Objectives hit (mean)	Objectives hit (%)	ROK acceptable?	% of standards acceptable
Reading	A. Reading/literature	4.2	4.2	100%	Yes	100%
Language arts	B. Writing	3.0	3.0	100%	Yes	100%
	C. Oral language	3.0	3.0	100%	Yes	
	D. Language	2.0	1.5	75%	Yes	
	E. Media and technology	5.0	2.7	54%	Yes	
	F. Research and inquiry	1.1	1.1	100%	Yes	
Math	A. Mathematical processes	5.2	4.5	87%	Yes	66%
	B. Number operations/relationships	7.3	6.3	86%	Yes	
	C. Geometry	4.0	1.7	43%	Weak	
	D. Measurement	5.1	4.4	86%	Yes	
	E. Statistics/probability	5.0	1.1	22%	No	
	F. Algebraic relationships	6.1	1.7	28%	No	
Science	A. Science connections	5.3	2.3	44%	Weak	75%
	B. Nature of science	3.1	1.1	35%	No	
	C. Science inquiry	8.2	4.8	59%	Yes	
	D. Physical science	8.2	3.5	43%	Weak	
	E. Earth and space science	8.1	3.3	41%	Weak	
	F. Life and environmental science	4.0	2.3	58%	Yes	
	G. Science applications	5.1	2.5	49%	Weak	
	H. Science in social/personal perspectives	4.1	1.1	27%	No	
Social studies	A. Geography	9.2	5.3	58%	Yes	80%
	B. History	10.1	2.4	24%	No	
	C. Political science	6.2	4.2	79%	Yes	
	D. Economics	7.1	4.8	67%	Yes	
	E. Behavioral sciences	15.1	9.6	42%	Weak	

Note. The Wisconsin English – Language Arts Model Academic Standards are represented by items on both the WAA Reading and Language Arts scales.

Table 17
Balance of Representation for WAA Subject Domain Scales

Subject domain	Academic standard	Objectives rated (mean)	Balance index mean (SD)	Balance of representation acceptable?	% of standards acceptable
Reading	A. Reading/literature	4.2	.79 (.07)	Yes	100%
Language arts	B. Writing	3.0	.83 (.07)	Yes	100%
	C. Oral language	3.0	.88 (.06)	Yes	
	D. Language	2.0	.94 (.11)	Yes	
	E. Media and technology	5.0	.86 (.11)	Yes	
	F. Research and inquiry	1.1	.98 (.08)	Yes	
Math	A. Mathematical processes	5.2	.73 (.13)	Yes	100%
	B. Number operations/relationships	7.3	.70 (.05)	Yes	
	C. Geometry	4.0	.96 (.07)	Yes	
	D. Measurement	5.1	.84 (.05)	Yes	
	E. Statistics/probability	5.0	.98 (.06)	Yes	
	F. Algebraic relationships	6.1	.90 (.11)	Yes	
Science	A. Science connections	5.3	.95 (.08)	Yes	100%
	B. Nature of science	3.1	.98 (.05)	Yes	
	C. Science inquiry	8.2	.85 (.06)	Yes	
	D. Physical science	8.2	.92 (.08)	Yes	
	E. Earth and space science	8.1	.85 (.06)	Yes	
	F. Life and environmental science	4.0	.98 (.05)	Yes	
	G. Science applications	5.1	.92 (.08)	Yes	
	H. Science in social/personal perspectives	4.1	1.00 (.00)	Yes	
Social studies	A. Geography	9.2	.84 (.03)	Yes	100%
	B. History	10.1	.91 (.10)	Yes	
	C. Political science	6.2	.82 (.10)	Yes	
	D. Economics	7.1	.80 (.05)	Yes	
	E. Behavioral sciences	15.1	.82 (.04)	Yes	

Note. The Wisconsin English – Language Arts Model Academic Standards are represented by items on both the WAA Reading and Language Arts scales.

limited number of items for each subject domain scale demanded that the scale developers evenly distribute items among the objectives. The panel members' ratings provided confirmation that the item development process resulted in a well-balanced scale for assessing students' performance.

Depth-of-knowledge consistency. In addition to evaluating the correspondence between the skills and concepts addressed in the academic standards and on the WAA instrument, the Alignment Institute results also provide a measure of the complexity of knowledge required by both documents. Depth-of-knowledge consistency describes the alignment between the skills and understanding students are expected to possess as stated in the standards and the skills and understanding necessary to successfully complete the WAA. According to Webb (2002), "For consistency to exist between an assessment and the standards, as judged in this analysis, at least 50% of the items corresponding to an objective had to be at or above the level of knowledge of the objective" (p. 4). If between 40% and 50% of the items were at or above the level of knowledge required by the objective, the depth-of-knowledge criterion was considered weakly attained.

Although it is generally desirable to obtain similar depth-of-knowledge ratings for curriculum objectives and assessment items, many alternate assessments' items may demand less depth of knowledge than items in the general education academic standards and on the corresponding large-scale assessment. WAA items represent the range of concepts and skills outlined in Wisconsin's Model Academic Standards, but these items are presented at a lower level of complexity that allows access for students with significant disabilities. Therefore, the WAA was not expected to demonstrate acceptable depth-of-knowledge consistency. The acceptance of a low overall depth-of-knowledge rating represents a departure from previous alignment studies using expert panel ratings (Webb, 2002; Webb et al., 2002). The results of the WAA Alignment Institute, however, indicate a generally acceptable level of depth-of-knowledge consistency for each subject domain scale (see Table 18).

When considered together, the results from the spring 2002 WAA Field Trial Study and the WAA Alignment Institute provide support for the WAA rating scale as a measure of student achievement and performance on the skills and concepts represented in Wisconsin's Model Academic Standards.

Discussion

The purpose of this investigation was to provide evidence of the validity of an enhancement of the Wisconsin Alternate Assessment for assessing the academic performance of students with significant disabilities. Specifically, the investigation provided evidence for the extent to which the WAA rating scale measures (a) the skills and concepts that make up the curriculum and instruction of students with significant disabilities and (b) the skill areas and concept areas represented in Wisconsin's Model Academic Standards.

Interpretation of Major Findings and Relation to Previous Research

Analysis of data gathered during the spring 2002 WAA Field Trial provided evidence for the adequacy of the WAA rating scale for students' mastery of the skills and concepts that make

Table 18
Depth-of-Knowledge (DOK) Consistency for the WAA Subject Domain Scales

Subject domain	Academic standard	% of items below DOK for objectives	% of items at (equal to) DOK for objectives	% of items above DOK for objectives	DOK acceptable?
Reading	A. Reading/literature	42%	47%	11%	Yes
Language arts	B. Writing	48%	44%	9%	Yes
	C. Oral language	75%	24%	1%	No
	D. Language	81%	19%	0%	No
	E. Media and technology	41%	52%	7%	Yes
	F. Research and inquiry	93%	7%	0%	No
Math	A. Mathematical processes	45%	53%	2%	Yes
	B. Number operations/relationships	6%	89%	5%	Yes
	C. Geometry	36%	59%	5%	Yes
	D. Measurement	2%	88%	10%	Yes
	E. Statistics/probability	77%	23%	0%	No
	F. Algebraic relationships	14%	79%	7%	Yes
Science	A. Science connections	78%	22%	0%	No
	B. Nature of science	70%	25%	5%	No
	C. Science inquiry	33%	47%	20%	Yes
	D. Physical science	43%	53%	5%	Yes
	E. Earth and space science	24%	64%	11%	Yes
	F. Life and environmental science	38%	38%	25%	Yes
	G. Science applications	6%	71%	23%	Yes
	H. Science in social/personal perspectives	27%	64%	9%	Yes
Social studies	A. Geography	44%	55%	1%	Yes
	B. History	25%	56%	6%	Yes
	C. Political science	59%	27%	14%	Weak
	D. Economics	82%	17%	1%	No
	E. Behavioral sciences	59%	34%	6%	Weak

Note. The Wisconsin English – Language Arts Model Academic Standards are represented by items on both the WAA Reading and Language Arts scales.

up their classroom curriculum. Specifically, field trial teachers on average rated between 60% and 70% of items on each WAA subject domain scale as applicable to their students' educational needs and outcomes. Although the mean number of items identified as aligned to field trial students' IEPs was significantly lower (approximately 15 items for the entire WAA rating scale), it is important to remember that IEPs are not meant to represent the totality of the curriculum and instruction provided to students with significant disabilities. In addition, relatively few items on the Science scale and Social Studies scale were designated as IEP-aligned.

A recent study of Kentucky's statewide alternate assessment (Turner, Baldwin, Kleinert, & Kearns, 2000) found a similar lack of relationship between students' alternate assessment scores and the content focus of students' IEP goals and objectives. However, these results may reflect the quality of students' IEPs rather than the technical adequacy of states' alternate assessment procedures. For example, an analysis of the IEPs of 46 students from nine different states (Giangrecco & Dennis, 1994) indicated that IEP goals and objectives are often broad or general in nature and inadequately referenced to the general education curriculum. The results of the current investigation indicate the need for further examination of the relationship among the state academic standards, the WAA, and the IEPs and day-to-day curriculum and instruction of students with significant disabilities.

In addition to rates of item applicability and IEP alignment, ratings of the importance of individual WAA items to the education goals of students with significant disabilities were gathered from the field trial teachers and the WAA Leadership Team. The relationship between the importance ratings and rates of IEP alignment and item applicability was considered additional evidence for the adequacy of the WAA as a measure of students' curriculum and instruction. The expectation was that those items that were considered most important by teachers and leadership team members would be the items (i.e., skills and concepts) most likely to be considered applicable to students' curriculum and included in their IEPs. The resulting Pearson correlations generally supported this expectation. The most notable exceptions were the negligible correlations between both groups' item importance ratings for the WAA Reading scale and the rates of IEP alignment for Reading items. This result may be attributable to the restricted range of field trial teachers and leadership team members' item importance ratings. Because all the WAA reading items were identified as important educational outcomes for students with disabilities, the resulting correlations with rates of item applicability and IEP alignment were not significant.

Although a follow-up survey of field trial teachers and parents indicated that both groups generally viewed WAA results as accurate measures of the academic performance of students with significant disabilities, the teachers expressed less confidence in the instructional utility of the WAA rating scales. A previous study of teachers' perceptions of the Kentucky alternate assessment system (Kampfer, Horvath, Kleinert, & Kearns, 2001) produced similar results. Teachers in the Kentucky investigation expressed low to moderate support for the benefit of student participation in the alternate assessment. The teachers did, however, indicate they were generally able to "embed" alternate assessment elements into instruction. In contrast to teachers in both studies, WAA field trial parents in the current investigation generally indicated stronger support of the instructional utility of the alternate assessment results.

Analysis of data gathered during the spring 2002 WAA Field Trial Study and WAA Alignment Institute indicated the WAA rating scale is an adequate measure of the skills and concepts represented by Wisconsin's Model Academic Standards. This result is significant in light of a recent alignment study by Kleinert and Kearns (1999), in which 9 of 44 participants questioned the focus of the Kentucky alternate assessment on functional living skills and instead recommended aligning the instrument with general education curricular standards.

In the current investigation, a follow-up survey of teachers and parents who participated in the WAA Field Trial Study indicated strong to moderate support among both groups for the importance of (a) assessing the reading, writing, and mathematics performance of all students and (b) aligning the WAA with the state academic standards. Moreover, field trial teachers indicated that completing the alternate assessment process helped to familiarize them with Wisconsin's Model Academic Standards. This capacity-building aspect of the WAA is important in light of special educators' traditional isolation and lack of knowledge concerning the general education curriculum and standards (Pugach & Warger, 1993).

Additional evidence of the degree to which the various WAA scales measure skills and behaviors similar to those outlined in state's academic standards was provided by the correlation between students' WAA results and their performance on two nationally normed rating scales of academic competence completed by field trial teachers. The correlations among measures were generally in the moderate positive range. The moderate positive correlations between the WAA, the SSRS Academic Competence scale, and the ACES suggest the instruments measure similar underlying knowledge and skills. Moreover, the correlations for the Reading, Language Arts, and Mathematics scales were consistently higher than those for the Science and Social Studies scales. This result was expected given that the item content of the ACES and SSRS-Academic Competence scale is primarily concerned with reading and mathematics.

The expert panel's responses during the WAA Alignment Institute indicate the WAA rating scale is as an adequate measure of the skills and knowledge represented by Wisconsin's Model Academic Standards. In fact, the performance of the WAA on the four criteria that make up Webb's (1997) alignment model met or exceeded the performance of many states' general education assessments. In comparison, 60% of the special education experts surveyed as part of a recent alignment analysis of APIs from 42 states indicated that most states had not adequately assessed the general education curriculum standards with their alternate performance indicators (Browder et al., 2002).

The WAA rating scale was not expected to demonstrate acceptable depth-of-knowledge consistency using Webb's alignment procedures; in fact, meeting the depth-of-knowledge criterion could be considered an indication that some WAA items were too difficult for the population of students for whom the test was developed. The results of the WAA Alignment Institute, however, indicated a generally acceptable level of depth-of-knowledge consistency between the WAA and the majority of academic standards in Reading, Mathematics, and Social Studies. There are multiple plausible explanations for this unexpected result: (a) the wording of the WAA items is general enough to allow for more complex interpretations of the tasks; (b) panel members felt that the items tapped the same skills and knowledge expected in the objectives in a way that made them accessible to students with severe disabilities; and (c) the

skills and concepts expected in the state's academic standards primarily focus on recall and simple application of knowledge.

Limitations of the Current Investigation and Directions for Future Research

The data examined in this investigation are from the initial WAA field trial and alignment studies. These data should be interpreted with caution but can be used to guide alternate assessment refinement efforts and the implementation of the WAA statewide.

Specifically, the results of this investigation are based on a relatively small sample of students with significant disabilities and their teachers. Although an effort was made to ensure the diversity of the sample in terms of geography, type of disability, and grade level, the participants in this study may not be entirely representative of the typical population of students who will take the WAA. Because the 2002–2003 school year represents the initial implementation year for the WAA, data collected in fall 2002 will provide (a) insight into the representativeness of participants in the current investigation and (b) the opportunity for additional validity studies based on a larger number of cases drawn from the overall population of students with significant disabilities.

In addition, because teachers in the WAA Field Trial Study were asked to volunteer for the investigation, the sample may have included teachers who were more interested in standards-based assessment of students with disabilities and instructional improvement than their peers. Thus, it is possible that the IEPs and classroom curricula of students in the field trial study were more aligned with Wisconsin's Model Academic Standards than those of the broader population of students with significant disabilities. Moreover, teachers in the field trial study were asked to complete only one alternate assessment over the period of approximately 2 months. During actual implementation of the WAA, teachers may be asked to complete multiple alternate assessments in a shorter time period. Therefore, teachers may not be able to give as close attention to the alignment and completion of the WAA rating scale as the teachers in our field trial. Conversely, one might argue that completing more than one WAA will lead to greater understanding of and comfort with the instrument and better insight into the IEP alignment process. Gathering additional case studies from the first few years of implementation will provide additional information about the relationship of the alternate assessment to students' IEPs and classroom curricula.

Participants in the WAA Alignment Institute were primarily Department of Public Instruction (DPI) Special Education Team administrators and graduate students in educational psychology. Although this expert panel had extensive understanding of testing and measurement, special education policy, and education accountability systems, the addition of other constituencies to the expert panel may have produced different alignment results. For example, the inclusion of special education researchers or practitioners on the expert panel would have provided additional insight into the curriculum and instruction of students with significant disabilities. In addition, previous applications of Webb's methods have utilized subject domain specialists when analyzing the alignment of tests and standards in a specific curricular area. Replication of the methods used in this investigation with other alternate assessments would provide additional evidence of the methods' applicability to the behavior ratings scales,

checklists, and portfolios generally employed to assess the academic performance of students with significant disabilities.

The correlations among the WAA rating scale, ACES, and SSRS were based on individual teachers' completion of all three measures on individual students. This common source of information for each student may have resulted in somewhat higher correlations than one would expect to find if different assessment methods or different raters had been used. Future validity studies should collect direct measures of pre-academic and academic skills (e.g., curriculum-based measurement probes), ACES and SSRS ratings from parents or a second educator, and adaptive behavior scales to further examine the relationships among these measures and the WAA results for students with significant disabilities.

The content covered by the WAA appears to be well aligned with Wisconsin's Model Academic Standards, but it is more comprehensive than most students' IEP objectives. The degree of alignment between students' IEP goals and the WAA items was greater (but by no means extensive) in the core academic areas of Reading, Language Arts, and Mathematics and lower in the areas of Social Studies and Science. Additional investigation of teachers' decision-making process for aligning WAA items could provide insight into the relatively low level of IEP alignment observed in this investigation, especially in the areas of Science and Social Studies. Moreover, further analyses of the IEPs of students with significant disabilities would determine whether they include standard-based goals, objectives, and benchmarks in multiple subject domains.

Feedback on a post-administration survey indicated field trial teachers were neutral regarding the instructional utility of the WAA rating scale. Although the WAA is well aligned to the state's academic standards, an improved sense of its effect on students' curriculum and instruction is important to determine its efficacy as an element of standards-based reform and accountability. Thus, longitudinal investigations of the mean number of IEP-aligned items on the WAA and the standards-based content of students' annual IEPs should be conducted to provide additional evidence of the consequential validity of the alternate assessment process.

Implications for Policy and Practice

The findings from the current investigation have implications for WAA policy and practice. In general, the results suggest the alternate assessment procedures are working well and can yield valid and useful scores. Based on the results of the field trial and alignment studies, the following actions were recommended to improve the validity and utility of WAA results: (a) add items to the WAA Science scale to improve its alignment to the state's academic standards and to the IEPs and classroom curriculum of students with significant disabilities; (b) provide additional training to special education teachers to increase their understanding of Wisconsin's Model Academic Standards and their use of the WAA; and (c) attempt to replicate these findings with a larger sample drawn from WAA rating forms completed during the first year of implementation (2002–2003).

Following completion of the field test and alignment studies, three new items were added to the WAA Science scale. These items were developed from the state's alternate performance

indicators and were selected to represent content areas on the Science Standards identified by the alignment study as not well represented on the WAA rating scale.

The Wisconsin DPI conducted extensive training on the use of the WAA during August and September 2002. Participant evaluations and implementation data will provide information on the efficacy of these efforts. In addition, teachers should receive additional support to understand the state's academic standards. Providing access to the general education curriculum requires training that raises special educators' awareness of the state's academic standards and their applicability to the curriculum and instruction of students with disabilities (Walsh, 2001).

Collecting data from a larger sample of students with significant disabilities will provide the opportunity to conduct additional analyses concerning the construct validity of the various scales. In particular, conducting a confirmatory factor analysis would help to establish the underlying five-factor structure of the WAA rating scale.

Conclusions

The current investigation provided evidence that the WAA rating scale represents a meaningful method of assessing the performance of students with significant disabilities. Specifically, the results suggest that performance on the WAA can serve as an index of achievement for (a) the skills and concepts that make up the curriculum and instruction of students with significant disabilities and (b) the concepts and knowledge represented in Wisconsin's Model Academic Standards.

In a recent presentation to the Alternate Assessment Forum at the CCSSO National Conference on Large Scale Assessment, Ken Warlick from the Office of Special Education Programs, U.S. Department of Education, discussed federal provisions concerning students with disabilities and state and district assessments. In particular, Warlick affirmed the need to create alternate assessments that measure students' progress toward the goals and standards held for all students:

> The purpose of an alternate assessment should reasonably match, at a minimum, the purpose of the assessment for which it is an alternate. One might ask, 'If an alternate assessment is based on totally different or alternate standards, or a totally separate curriculum, what is the alternate assessment an alternate to?' (Quenemoen, Massanari, Thompson, & Thurlow, 2000, p. 15)

The results of the WAA Alignment Institute and Field Trial Study suggest that the WAA is an appropriate alternate measure of the general subject domains included on the WKCE, Wisconsin's general statewide assessment.

Aligning to a state's general education academic standards is only one aspect, however, of creating a meaningful alternate assessment. Ysseldyke and Olsen (1997) suggest that, in addition to aligning to academic standards, alternate assessments should be curriculum-relevant, measuring what students with significant disabilities are learning and doing in their classrooms. In many cases, the curriculum and instruction of students who participate in an alternate assessment differ significantly from those of other students. Therefore, test developers must determine the alignment between alternate assessments and the curriculum and instruction

provided to students with significant disabilities. The results of the WAA Field Trial Study provide evidence for this relationship but also suggest the need for additional work to improve the match between students' IEPs and the state's alternate assessment and academic standards. Strengthening this aspect of alignment will help ensure that students with significant disabilities are included in instructional improvement efforts and standards-based reform in a meaningful way.

References

Airasian, P. W. (1994). *Classroom assessment* (2nd ed.). Boston: McGraw-Hill.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *The standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Browder, D., Flowers, C., Ahlgrim-Delzill, L., Karvonen, M., Spooner, F., & Algozzine, R. (2002, April). *Curricular implications of alternate assessments.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Demaray, M. K., & Elliott, S. N. (1998). Teachers' judgments of students' academic functioning: A comparison of actual and predicted performance. *School Psychology Quarterly, 13,* 8–24.

DiPerna, J. C., & Elliott, S. N. (2000). *Academic competence evaluation scales.* San Antonio, TX: The Psychological Corporation.

Elementary and Secondary Education Act, 20 U.S.C. ə 6311(b)(3)(C)(ii) (2002).

Elliott, S. N. (2001). *Enhancing the Wisconsin Alternate Assessment project proposal*. Madison, WI: Wisconsin Department of Public Instruction.

Elliott, S. N., Braden, J. B., & White, J. L. (2001). *Assessing one and all: Educational accountability for students with disabilities.* Arlington, VA: Council for Exceptional Children.

Ford, A., Davern, L, & Schnorr, R. (2001). Learners with significant disabilities: Curricular relevance in an era of standards-based reform. *Remedial and Special Education, 22,* 214–222.

Giangrecco, M. F., & Dennis, R. E. (1994). Dressing your IEPs for the general education climate: An analysis of IEP goals and objectives for students with multiple disabilities. *Remedial & Special Education, 15,* 288–297.

Gresham, F. M., & Elliott, S. N. (1990). *Social skills rating system.* Circle Pines, MN: American Guidance Service.

Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59,* 297–313.

Individuals With Disabilities Education Act Amendments of 1997, Pub. L. No. 105-17, 111 Stat. 37 (codified as amended at 20 U.S.C. § 1400 *et seq.*).

Kampfer, S. H., Horvath, L. S., Kleinert, H. L., & Kearns, J. F. (2001). Teachers' perceptions of one state's alternate assessment: Implications for practice and preparation. *Exceptional Children, 67,* 361–374.

Kleinert, H. L., & Kearns, J. F. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's alternate assessment for students with significant disabilities. *The Journal of the Association for Persons With Severe Handicaps, 24,* 100–110.

Lane, S. (1999, October). *Validity evidence for assessments.* Paper presented at the Edward F. Reidy Interactive Lecture Series, The National Center for the Improvement of Educational Assessment, Providence, RI.

Linn, R. L., & Gronlund, N. E. (1995). *Measurement and assessment in teaching* (7th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in kindergarten to grade 3. *American Educational Research Journal, 38,* 73–95.

Pugach, M. C., & Warger, C. L. (2001). Curriculum matters: Raising expectations for students with disabilities. *Remedial and Special Education, 22,* 194–196.

Quenemoen, R., Massanari, C., Thompson, S., & Thurlow, M. (2000). *Alternate assessment forum: Connecting into a whole.* Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 5, 2003, from http://education.umn.edu/NCEO/OnlinePubs/Forum2000/ForumReport2000.htm

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420 – 428.

Thompson, S. J., Quenemoen, R. F., Thurlow, M. L., & Ysseldyke, J. E. (2001). *Alternate assessments for students with disabilities.* Thousand Oaks, CA: Corwin Press.

Thompson, S. J., & Thurlow, J. L. (2000). *State alternate assessments: Status as IDEA alternate assessment requirements take effect* (NCEO Synthesis Report No. 35). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved February 5, 2003, from http://education.umn.edu/nceo/OnlinePubs/Synthesis35.html

Thurlow, M., Olsen, K., Elliott, J., Ysseldyke, J., Erickson, R., & Ahern, E. (1996). Alternate assessments for students with disabilities. *NCEO Policy Directions, 5,* 1–6.

Turner, M. D., Baldwin, L., Kleinert, H. L., & Kearns, J. F. (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.

Walsh, J. M. (2001). Getting the "big picture" of IEP goals and state standards. *Teaching Exceptional Children, 5,* 18–26.

Webb, N. L. (1997). *Criteria for alignment of expectations and assessments in mathematics and science education* (NISE Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Webb, N. L. (2002, April). *An analysis of the alignment between mathematics standards and assessments for three states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Webb, N. L., Horton, M., & O'Neal, S. (2002, April). *An analysis of the alignment between language arts standards and assessments in four states.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Wisconsin Department of Public Instruction. (1998). *Guidelines for complying with the assessment provisions of the Individuals with Disabilities Education Act* (Bulletin 98.14). Madison: Author.

Ysseldyke, J. E., & Olsen, K. R. (1997). *Putting alternate assessments into practice: What to measure and possible sources of data* (NCEO Synthesis Report No. 28). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes. Retrieved February 5, 2003, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis28.htm

Appendix A

WAA Participation Checklist

WISCONSIN ALTERNATE ASSESSMENT PARTICIPATION CHECKLIST

(Final Draft 6/22/01)

IEP team members are responsible for deciding which students with disabilities participate in the regular assessment, with or without testing accommodations, or in the state's alternate assessment. To facilitate informed and equitable decision-making, IEP teams should address each of the following statements **for each of the 4 content areas** when considering an alternate assessment. Check all that apply.

NOTE: If the IEP team concurs that all four of the statements below accurately characterize a student's current educational situation, then an alternate assessment should be used to provide a meaningful evaluation of the student's current academic achievement.

Participation Criteria	***Reading/ Language Arts***	***Math***	***Science***	***Social Studies***
1. The student's curriculum and daily instruction focus on knowledge and skills significantly below those represented by the state's content standards for students of the same chronological age.				
2. The student's present level of educational performance (PLOEP) significantly impedes participation and completion of the general education curriculum even with significant program modifications.				
3. The student requires extensive direct instruction to accomplish the acquisition, application, and transfer of knowledge and skills.				
4. The student's difficulty with the regular curriculum demands is primarily due to his/her disabilities, and not to excessive absences unrelated to the disability, or social, cultural or environmental factors.				

ASSUMPTIONS:

- The IEP team has knowledge of the **student's PLOEP** in reference to the Wisconsin Model Academic Standards.
- The IEP team has working knowledge of the test format and what skills and knowledge **are** being measured by the statewide assessments.
- The IEP team is knowledgeable of state testing guidelines and the use of appropriate testing accommodations.

Appendix B

WAA Teacher and Parent Surveys

Teacher Survey
Wisconsin Alternate Assessment Project
2002

After you have completed your case study, please take a few minutes an answer each of the 10 questions below by checking the box that best represents your position. Please do not skip any items.

	1 Strongly Disagree	2 Mildly Disagree	3 Neutral	4 Mildly Agree	5 Strongly Agree
1. The WAA was easy to use.					
2. The WAA facilitated the participation in the state's assessment system of students who historically would have been left out.					
3. The amount of time needed to administer the WAA will be reasonable after this initial assessment year.					
4. The results of the WAA were useful to me and others who make instructional plans for students with significant disabilities.					
5. The WAA items were well aligned with the state's general education academic standards.					
6. By conducting an alternate assessment, I learned more about Wisconsin's academic standards and statewide assessment system.					
7. The results of the WAA that I conducted appeared to be statistically sound.					
8. The student results appeared to be an accurate representation of the student's skills that were measured.					
9. The Participation Decision Checklist was helpful.					
10. The scores and performance level information resulting from the WAA are meaningful.					

Comments:

_____

_____

_____

Please return this survey to: Thomas Kratochwill, Department of Educational Psychology, 1025 W. Johnson Street., 333 Education Sciences, UW-Madison, Madison WI 53706-1796

Parent Survey
Wisconsin Alternate Assessment Project
2002

In the forthcoming school year, the state of Wisconsin will be implementing an enhanced Alternate Assessment (WAA) for students who cannot meaningfully participate in the regular statewide test. If your child was included in the WAA Field Test during this spring, please take 5 minutes to complete the following items. Your feedback is valued and will contribute to the evaluation of the WAA and possible improvement of this assessment.

For items #1 through #5, circle the number that best characterizes your perception of or reaction to the WAA. Please use the following responses: 1 = strongly disagree, 2 = mildly disagree, 3 = neutral, 4 = mildly agree, and 5 = strongly agree.

1. I found the alternate assessment to be useful.	1 2 3 4 5
2. I was pleased to know that the assessment was aligned with the state's academic standards.	1 2 3 4 5
3. I think it is good that all students in the state participate in an assessment that focuses on their achievement in basic areas such as reading, writing, and math.	1 2 3 4 5
4. I was confident in the results about my child's functioning that the teachers provided me.	1 2 3 4 5
5. I believe the time spent by teachers conducting an alternate assessment is important to their teaching of my child.	1 2 3 4 5

If you have any of thoughts or reactions about the alternate assessment that was completed with your child, please let us know below:

_____
_____
_____
_____

Thank you very much. Return this survey to your child's teacher by May 1, 2002.

Teacher's Name: _____ Address: _____

Appendix C

WAA Alignment Institute Item Coding Forms

Wisconsin Alternate Assessment—Reading

Item Review Form Reviewer: _____

Item	DOK	P Std/ Obj	S1 Std/ Obj	S2 Std/ Obj	Source of Challenge	Notes
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						

Wisconsin Alternate Assessment—Language Arts

Item Review Form Reviewer: _____

Item	DOK	P Std/ Obj	S1 Std/ Obj	S2 Std/ Obj	Source of Challenge	Notes
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						

Wisconsin Alternate Assessment—Mathematics

Item Review Form Reviewer: _____

Item	DOK	P Std/ Obj	S1 Std/ Obj	S2 Std/ Obj	Source of Challenge	Notes
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						

Wisconsin Alternate Assessment—Science

Item Review Form Reviewer: _____

Item	DOK	P Std/ Obj	S1 Std/ Obj	S2 Std/ Obj	Source of Challenge	Notes
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						

Wisconsin Alternate Assessment—Social Studies

Item Review Form Reviewer: _____

Item	DOK	P Std/ Obj	S1 Std/ Obj	S2 Std/ Obj	Source of Challenge	Notes
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						

Appendix D

WAA Alignment Institute Depth-of-Knowledge Criteria

Depth of Knowledge for Reading
Wisconsin Alternate Assessment
Alignment Analysis
June 2002

Reading Level 1

Level 1 items and standards require students to receive or recite facts or to use simple skills or abilities. Oral reading and basic comprehension of the text are included in Level 1. Items require only a shallow understanding of the text presented and often consist of verbatim recall of information from the text or simple understanding of a single word or phrase. Verbs such as "identify," "recall," "recognize," "use" and "remember" are typically used in Level 1 items or standards. Some examples that represent but do not constitute, all of Level 1 performance are:

- Student uses a dictionary to find the meaning of words.
- Student remembers and recalls details from a text.
- Student identifies main characters in a reading passage.

Reading Level 2

Items in Level 2 require the engagement of some mental processing beyond recalling or reproducing a response; these items require both comprehension and subsequent processing of texts or portions of text. Some important concepts may be covered but not in a complex way. Items and standards at this level may include words such as "summarize," "interpret," "infer," "classify," "organize," "collect," "display," "compare," or "determine whether fact or opinion." Literal main ideas may be identified. A Level 2 item may require students to apply some of the skills and concepts that are covered in Level 1. Some examples that represent but do not constitute all of Level 2 performance are:

- Student uses context cues to identify the meaning of unfamiliar words.
- Student predicts the logical outcome based on information in a reading selection.
- Student retells a story, including the major events in the narrative.

Reading Level 3

Deep knowledge becomes a greater focus at Level 3. Students are encouraged to go beyond the text; however, they are still required to show understanding of the ideas in the text. Students may be asked to explain, generalize, or connect ideas. Standards and items at Level 3 require reasoning and planning. Students must support or explain their thinking. Items may involve abstract theme identification or students' application of prior knowledge and experience. Items may call for simple comparisons between texts. Some examples that represent but do not constitute all of Level 3 performance are:

- Student determines the author's purpose for writing a text and understands how different texts have different purposes (e.g., letter, journals, reports, or stories).
- Student researches a topic using a variety of sources and summarizes their findings.
- Student can identify different types of literature and can classify a text after reading it.

Reading Level 4

Higher order thinking is central and knowledge is deep at Level 4. The standard or item at this level will probably be an extended activity, requiring extensive time to complete. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require application of significant conceptual understanding and higher order thinking. Students take information from at least one text and are ask to apply it to complete a new task. They may also be asked to develop hypotheses and perform complex analyses of the connections among the texts. Some examples that represent but do not constitute all of Level 4 performance are:

- Student can analyze and synthesize information from multiple written sources.
- Student uses graphic organizers to organize and analyze information from multiple texts.
- Student can extend or identify the themes and concepts in multiple texts and draw connections between them.

Depth of Knowledge for Language Arts
Wisconsin Alternate Assessment
Alignment Analysis
June 2002

Language Arts Level 1

Level 1 items and standards require students to write or recite simple texts or narrative and to use basic skills and abilities. The writing or recitation does not include complex synthesis or analysis. Level 1 only requires students to follow a set procedure (like a recipe) or perform a clearly defined series of steps. Verbs such as "identify," "recall," "recognize," "use" and "remember" are typically used in Level 1 items or standards. Students are expected to write or speak using Standard English conventions. This includes using appropriate grammar, punctuation, capitalization, and spelling. Some examples that represent but do not constitute all of Level 1 performance are:

- Student uses punctuation correctly.
- Student lists ideas or words (i.e., brainstorms) about a simple topic.
- Student identifies and uses Standard English grammatical structures in oral communication.

Language Arts Level 2

Items in Level 2 require the engagement of some mental processing beyond recalling or reproducing a response. Students are engaged in first-draft writing or brief extemporaneous speaking for a limited number of purposes or audiences. Students are beginning to connect ideas, using a simple organizational structure. Items and standards at this level may include words such as "summarize," "interpret," "infer," "classify," "organize," "collect," "display," and "compare." Students demonstrate a basic understanding and appropriate use of such reference materials as a dictionary, thesaurus, or Web site. A Level 2 item may require students to apply some of the skills and concepts that are covered in Level 1. Some examples that represent but do not constitute all of Level 2 performance are:

- Student constructs compound sentences.
- Student uses simple organizational structures to organize written work or oral presentations.
- Student engages in note taking, outlining, or simple summaries of oral presentations and written texts.

Language Arts Level 3

Level 3 requires some higher level mental processing. Students are engaged in developing substantial oral presentations and write compositions that include multiple paragraphs. Students' work includes complex sentence structure, and they are expected to demonstrate some synthesis and analysis. Standards and items at Level 3 are more cognitively demanding, requiring reasoning and planning on the part of the student. Students may be asked to support or explain their thinking. Items may involve application of prior knowledge and experience or the inclusion of supporting facts and details in an informational report. At this level, students are engaged in editing and revising to improve the quality of their oral presentations and written work. Some examples that represent but do not constitute all of Level 3 performance are:

- Student supports ideas with details and examples.
- Student uses appropriate compositional elements (e.g., addressing chronological order in a narrative or showing an awareness of the audience) to improve his or her ability to communicate orally or in writing.
- Student can edit and revise written work and oral presentations to produce a logical progression of ideas.

Language Arts Level 4

Higher order thinking is central and knowledge is deep at Level 4. The standard or item at this level will probably be an extended activity, requiring extensive time to complete. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require application of significant conceptual understanding and higher order thinking. Level 4 tasks have high cognitive demands and are very complex. Students are expected to create oral presentations and written compositions that demonstrate a distinct voice and that stimulate the reader or listener to consider new perspectives to address ideas and themes. Some examples that represent but do not constitute all of Level 4 performance are:

- Student creates oral presentations and written works that analyze and synthesize information from multiple sources.
- Student produces written text and oral reports that include hypotheses and supporting evidence.
- Student writes a multi-paragraph composition that demonstrates synthesis and analysis of complex idea and themes.

Depth of Knowledge for Math
Wisconsin Alternate Assessment
Alignment Analysis
June 2002

Math Level 1

Level 1 items and standards require students to recall facts, definitions, and terms, or to use simple skills or abilities. Simple algorithmic procedures are included in Level 1. Items require only a shallow understanding of the math concepts and typically only require students to solve a basic math problem or identify or recognize the "right" answer. Simple word problems that can be directly translated into a number sentence and solved by computation are considered a Level 1 task. Verbs such as "identify," "recall," "recognize," "compute," "measure," and "use" are typically used in Level 1 items or standards. Some examples that represent but do not constitute all of Level 1 performance are:

- Student performs a routine procedure such as measuring weight.
- Student recognizes math words or symbols.
- Student performs a simple computational algorithm.

Math Level 2

Items in Level 2 require the engagement of some mental processing beyond recalling or reproducing a response; these items require students to make some decision about how to approach a problem or activity. Items and standards at this level may include words such as "classify," "organize," "estimate," "compare," "display," "make observations," or "collect data." Level 2 items and standards often imply actions that include more than one step. A Level 2 item may require students to apply some of the skills and concepts that are covered in Level 1. Some examples that represent but do not constitute all of Level 2 performance are:

- Student reads a basic chart or graph and applies the information in a different context.
- Student picks the appropriate strategy and solves a simple word problem.
- Student solves a math problem with more than one step, beyond just applying a simple algorithm.
- Student decides which operation to use and then uses the operation.

Math Level 3

Reasoning, planning, and using evidence become a greater focus at Level 3. Students are presented with tasks that can be solved more than one way and may have more than one right answer. Students may be asked to explain, generalize, or connect ideas and concepts. Students may be asked to support, explain, and justify how they solved a problem. Items and standards at Level 3 are more complex and abstract. This complexity results not only from the fact that there are multiple answers (a possibility for items at lower levels), but also from the fact that the task requires more demanding reasoning. Some examples that represent but do not constitute all of Level 3 performance are:

- Student analyzes data from charts and graphs and uses it to solve problems.
- Student decides on the appropriate tools, selects the appropriate units of measurement, and uses measurement to solve a problem.

- Student identifies connections between math ideas and concepts (e.g., student demonstrates understanding of the relationship between addition and multiplication and applies this relationship in a different or novel context).

Math Level 4

Higher order thinking is central and knowledge is deep at Level 4. The standard or item at this level will probably be an extended activity, requiring extensive time to complete. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require application of significant conceptual understanding and higher order thinking. At Level 4, the cognitive demands of the task should be high and the work should be very complex. Students should be required to make several connections—relate ideas within the content area or among content areas—and to select one approach among many alternatives on how a problem can be solved. They may also be asked to develop a hypothesis and perform complex analyses of the connections among concepts and ideas. Some examples that represent but do not constitute all of Level 4 performance are:

- Student uses data from multiple sources to solve a complex, real-world problem.
- Student constructs a visual representation of a complex mathematical concept.
- Student uses physical models to examine the relationship between two- and three-dimensional figures.

Depth of Knowledge for Science
Wisconsin Alternate Assessment
Alignment Analysis
June 2002

Science Level 1
Level 1 requires students to recall information, such as a fact, definition, term, or a simple procedure, and to perform a simple science process or procedure. Level 1 only requires students to demonstrate a rote response, use a well-known formula, follow a set procedure (like a recipe), or perform a clearly defined series of steps. A "simple" procedure is well defined and typically involves only one step. Verbs such as "identify," "recall," "recognize," "use," "calculate," and "measure" generally represent cognitive work at the recall and reproduction level. Some examples that represent but do not constitute all of Level 1 performance are:

- Student recalls or recognizes a fact, term, or property.
- Student represents in words or diagrams a scientific concept or relationship.
- Student provides or recognizes a standard scientific representation for simple phenomenon.
- Student performs a routine procedure such as measuring length.

Science Level 2
Level 2 includes the engagement of some mental processing beyond recalling or reproducing a response. The content knowledge or process involved is more complex than in Level 1. Items require students to make some decisions as to how to approach the question or problem. Keywords that generally distinguish a Level 2 item include "classify," "organize," "estimate," "make observations," "collect and display data," and "compare data." These actions imply more than one step. For example, to compare data requires first identifying characteristics of the objects or phenomenon and then grouping or ordering the objects. Some action verbs, such as "explain," "describe," or "interpret," could be classified at different depth-of-knowledge levels, depending on the complexity of the action. For example, interpreting information from a simple graph, requiring reading information from the graph, is a Level 2 activity. Some examples that represent but do not constitute all of Level 2 performance are:

- Student specifies and explains the relationship between facts, terms, properties, or variables.
- Student makes observations, collects data, and displays data in tables, graphs, and charts.
- Student compares and classifies animals or plants according to multiple characteristics.
- Student selects a procedure according to specified criteria and performs it.
- Student formulates a routine problem given data and conditions.

Science Level 3
Level 3 requires reasoning, planning, using evidence, and demonstrating a higher level of thinking than the previous two levels. The cognitive demands at Level 3 are complex and abstract. The complexity results not only from the fact that there are multiple answers (a possibility for both Levels 1 and 2), but also from the fact that the multi-step task requires more demanding reasoning. An activity that has more than one possible answer and requires students to justify the response they give would most likely be a Level 3 activity. Experimental designs in

Level 3 typically involves more than one dependent variable. Other Level 3 activities include drawing conclusions from observations; citing evidence and developing a logical argument for concepts; explaining phenomena in terms of concepts; and using concepts to solve non-routine problems. Some examples that represent but do not constitute all of Level 3 performance are:

- Student identifies research questions and designs investigations for a scientific problem.
- Student solves non-routine problems.
- Student develops a scientific model for a complex situation.
- Student forms conclusions from experimental data.

Science Level 4

Level 4 tasks have high cognitive demands and are very complex. Students are required to make several connections—relate ideas *within* the content area or *among* content areas—and to select or devise one among many possible solutions. Level 4 requires complex reasoning, experimental design and planning, and probably an extended period of time either for the science investigation required by an objective or for completion of the multiple steps of an assessment item. However, the extended time period is not a distinguishing factor if the required work is only repetitive and does not require applying significant conceptual understanding and higher order thinking. For example, if a student has to take the water temperature from a river each day for a month and then construct a graph, this would be classified as a Level 2 activity. However, if the student conducts a river study that requires taking into consideration a number of variables, this would be a Level 4 activity. Some examples that represent but do not constitute all of Level 4 performance are:

- Based on data from a complex experiment that is novel to the student, the student deducts the fundamental relationship between several controlled variables.
- The student conducts an investigation, from specifying a problem, to designing and carrying out an experiment, to analyzing its data and forming conclusions.

Depth of Knowledge for Social Studies
Wisconsin Alternate Assessment
Alignment Analysis
June 2002

Social Studies Level 1

Level 1 items and standards require students to recall facts, terms, concepts, trends, generalizations, and theories. Students are expected to recognize and identify specific information in the graphics. The items at this level usually ask the student to recall who, what, when, and where. Verbs such as "identify," "recall," "recognize," "use" and "remember" are typically used in Level 1 items or standards. Some examples that represent but do not constitute all of Level 1 performance are:

- Student uses a map to identify basic landforms or geopolitical features.
- Student recognizes important historical figures and their accomplishments.
- Student uses basic charts or pictures to answer factual questions.

Social Studies Level 2

Items in Level 2 require the engagement of some mental processing beyond recalling or reproducing a response. This level generally requires students to contrast or compare people, places, events, and concepts. Students are expected to convert information from one form to another, give examples, and classify and sort items into meaningful categories. A Level 2 item may require students to apply some of the skills and concepts that are covered in Level 1. Level 2 items require students to make some decisions as to how to approach a question or problem. Some examples that represent but do not constitute all of Level 2 performance are:

- Student explains important issues in the classroom, school, or community.
- Student presents an explanation of the various causes that contributed to a historical event.
- Student explains the significance of individuals from history or current events.

Social Studies Level 3

Level 3 requires some higher level mental processing. Students go beyond explaining or describing "how and why" to justifying "how and why" through application and evidence. Standards and items at Level 3 may require strategic thinking and planning. The cognitive demands of Level 3 are complex and abstract. This complexity result not only from the fact that there are multiple correct responses (a possibility for items at lower levels), but also from the fact that the task requires more demanding reasoning. Students are asked to support or explain their thinking. Items may involve application of prior knowledge and experience or the inclusion of supporting facts and details. Some examples that represent but do not constitute all of Level 3 performance are:

- Student makes connections between, compares, and contrasts individuals or events across time and place.
- Student proposes and justifies a solution to a prominent social problem.

Social Studies Level 4

Higher order thinking is central and knowledge is deep at Level 4. The standard or item at this level will probably be an extended activity, requiring extensive time to complete. The extended time period is not a distinguishing factor if the required work is only repetitive and does not require application of significant conceptual understanding and higher order thinking. Tasks and standards at Level 4 should have high cognitive demands and be very complex. Students are expected to connect and relate ideas and concepts within the content area and among content areas in order to be at this highest level. Some examples that represent but do not constitute all of Level 4 performance are:

- Student analyzes and synthesizes information from multiple sources to support his or her own hypotheses about a historical event.
- Student examines alternative perspectives on a social dilemma using information from a variety of sources.
- Student describes and illustrates how a common theme is evident in both a historical event and a literary work.