



Validity of Test Results

IN THE MOST BASIC OF TERMS, tests are said to be valid if they do what they are supposed to do. Unfortunately, it is far easier to define validity than it is to demonstrate conclusively that a particular test is indeed valid. In part, this is because validity is at heart a relative rather than an absolute concept. A test's validity will vary according to the purpose for which its results are being given and the types of individuals tested. Therefore, a test's validity must be investigated again and again until a conclusive body of research has accumulated. The analysis and interpretation of the results of this entire body of research are necessary before the status of a test's validity can be known with any degree of certainty. The study of any test's validity is an ongoing process.

Most authors of current textbooks dealing with educational and psychological measurement (e.g., Aiken, 1994; Anastasi & Urbina, 1997; Linn & Gronlund, 1995; Salvia & Ysseldyke, 1998; and Wallace, Larsen, & Elksnin, 1992) suggest that those who develop tests should provide evidence of at least three types of validity: content description, criterion prediction, and construct identification. The particular terms used here are from Anastasi and Urbina (1997). Other sources refer to content validity, criterion-related validity, and construct validity. Although the terms differ somewhat, the concepts they represent are identical.

Content-Description Validity

“Content-description validation procedures involve the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured” (Anastasi & Urbina, 1997, pp. 114–115). Obviously, this kind

of validity has to be built into the test at the time that subtests are conceptualized and items constructed.

Those who build tests usually deal with content validity by showing that the abilities chosen to be measured are consistent with the current knowledge about a particular area and by showing that the items hold up statistically.

Three demonstrations of content validity are offered for the SAGES-2 subtests. First, a rationale for the content and the format is presented. Second, the validity of the items is ultimately supported by the results of “classical” item analysis procedures used to choose items during the developmental stages of test construction. Third, the validity of the items is reinforced by the results of differential item functioning analysis used to show the absence of bias in a test’s items.

Rationale Underlying the Selection of Formats and Items for the Subtests

Before designing the items for the SAGES-2, we reviewed the following tests used for screening gifted students in intelligence and in achievement:

- *Iowa Test of Basic Skills, Form M* (Hoover, Hieronymus, Frisbie, & Dunbar, 1996)
- *Kaufman Assessment Battery for Children* (Kaufman & Kaufman, 1984)
- *Kaufman Test of Educational Achievement* (Kaufman & Kaufman, 1998)
- *Metropolitan Readiness Tests* (Nurss, 1995)
- *Naglieri Nonverbal Ability Test* (Naglieri, 1996)
- *Otis-Lennon School Ability Test* (Otis & Lennon, 1996)
- *Peabody Individual Achievement Test-Revised* (Markwardt, 1989)
- *Slosson Intelligence Test-Revised* (SIT-R; Slosson, Nicholson, & Hibpshman, 1991)
- *Test of Mathematical Abilities for Gifted Students* (Ryser & Johnsen, 1998)
- *Test of Reading Comprehension* (V. L. Brown, Hammill, & Wiederholt, 1995)
- *Test of Nonverbal Intelligence-Third Edition* (L. Brown, Sherbenou, & Johnsen, 1997)
- *Wechsler Intelligence Scale for Children-Third Edition* (Wechsler, 1991)
- *Wechsler Preschool and Primary Scale of Intelligence-Revised* (Wechsler, 1989)
- *Wide Range Achievement Test 3* (Wilkinson, 1993)
- *Woodcock Reading Mastery Tests-Revised* (Woodcock, 1998)

We reviewed the national standards in each of the core subject areas, using the following sources:

- *Curriculum and Evaluation Standards for School Mathematics* (National Council of Teachers of Mathematics, 1989)
- *History for Grades K–4: Expanding Children’s World in Time and Space* (National Center for History in the Schools, n.d.-a)
- *Science Education Standards* (National Research Council, 1996)
- *National Standards for World History: Exploring Paths to the Present* (National Center for History in the Schools, n.d.-b)
- *Standards for the English Language Arts* (National Council of Teachers of English & International Reading Association, 1996)

We examined many textbooks in the academic areas. We found these books particularly helpful:

- *Geography, People and Places in a Changing World* (Moerbe & Henderson, 1995)
- *Harcourt Brace Jovanovich Language* (Strickland, Abrahamson, Farr, McGee, & Roser, 1990)
- *Just Past the Possible: Part 2* (Aoki et al., 1993)
- *Mathematics Activities for Elementary School Teachers* (Dolan & Williamson, 1990)
- *Mathematics Unlimited 7* (Fennell, Reys, Reys, & Webb, 1991)
- *Macmillan Earth Science Laboratory and Skills Manual* (Danielson & Denecke, 1986)
- *World Adventures in Time and Place* (Banks et al., 1997)

Finally, we observed the types of activities that occur in classrooms for gifted and talented students. The rationale underlying each of the three subtests is described next, followed by a description of the item analysis.

Format Selection In reviewing the literature, we selected those formats that are most familiar to students in Grades K–8 and that appeared suitable for measuring the SAGES–2 constructs (school-acquired information in four core academic areas—math, science, language arts, and social sciences—and reasoning).

Subtest 1: Mathematics/Science. This subtest samples achievement in mathematics and science, the two of the four core academic areas whose foundation is more logical or technical in nature. It requires the student to respond to questions in a familiar

multiple-choice format. Because the subtest is not timed, a student may take as much time as necessary to formulate a response to each item. The content for this subtest was drawn from current texts, professional literature, books, and the national standards, and is sequenced according to difficulty.

Mathematics items are closely related to the *Curriculum and Evaluation Standards for School Mathematics* set forth by the National Council of Teachers of Mathematics (NCTM, 1989). The first four—mathematics as problem solving, mathematics as communication, mathematics as reasoning, and mathematical connections—are standards for kindergarten through Grade 8. At least 62% of the items at each of the two SAGES-2 levels relate to these four standards. In addition, at both levels, all concepts and mathematical operations using whole and fractional numbers are systematically addressed. The remaining items relate to other NCTM standards, including number sense and numeration, number and number relationships, computation and estimation, measurement, patterns and relationships, and algebra.

Science items relate to the National Research Council's (1996) *Science Education Standards*. The vast majority of the items relate to the four standards that are frequently addressed in textbooks and classrooms: science as inquiry, physical science, life science, and earth and space science.

Subtest 2: Language Arts/Social Studies. This subtest samples achievement in language arts and social studies. Again, the subtest is not timed. A student may take as much time as necessary to formulate a response to each item. This subtest is similar to Subtest 1 in that the student responds to multiple-choice questions. These items reflect knowledge in the two of the four core academic areas whose foundation is more linguistic in nature. When given in combination, Subtests 1 and 2 sample the academic areas that are most frequently addressed in classrooms and in academic programs for gifted students. The content for this subtest was drawn from current texts, professional literature, books, and the national standards, and is sequenced according to difficulty.

The language arts items relate to the broad standards that are defined by the National Council of Teachers of English and International Reading Association (1996). These organizations state that these standards “are not distinct and separable; they are, in fact, interrelated and should be considered as a whole” (p. 3). Indeed, we found that the SAGES-2 items relate to most of the 12 identified standards, particularly those that require students to read and acquire information from a wide range of print and non-print texts (Standard 1); “apply a wide range of strategies to comprehend, interpret, evaluate, and appreciate texts” (Standard 3, p. 3); “apply knowledge of language structure, language conventions, media techniques, figurative language, and genre to create, critique, and discuss print and nonprint text” (Standard 6, p. 3); and “gather, evaluate, and synthesize data” (Standard 7, p. 3).

Social studies items related to the standards of the Center for Civic Education (1994) and those of the National Center for History in the Schools (n.d.-b). The social studies items address all five areas suggested by the National Center for History in the Schools: chronological thinking, historical comprehension, historical analysis and interpretation, historical research capabilities, and historical issues—analysis and decision making. In addition, social studies also relate to government, democracy, relationships among nations, and citizen roles that are included within the *National Standards for Civics and Government* set forth by the Center for Civic Education (1994).

Subtest 3: Reasoning. The Reasoning subtest samples one aspect of intelligence or aptitude: problem solving or analogical reasoning. When gifted students are identified for programs emphasizing general intellectual ability, some measure of aptitude is often included. The Reasoning subtest was designed to estimate aptitude, that is, a student's capacity to learn the kinds of information necessary to achieve in programs designed for gifted students. Aptitude is not related to abilities that are formally taught in school. This subtest requires the student to solve new problems by identifying relationships among figures and pictures. For each analogy item, the student is shown three pictures or three figures, two of which are related, and a series of five pictures or five figures. The student is to point to or mark which of the five pictures or figures relates to the third unrelated picture or figure in the same way that the first two pictures or two figures are related. Since the subtest is not timed, the student may take as much time as needed to think about his or her choices.

The items are constructed to vary characteristics related to shading, function, size, shape, position, direction, movement, and mathematical concepts (i.e., number, addition, part-whole). Special care was taken to include items that require flexible and novel kinds of thinking while maintaining an emphasis on convergent skills. For example, Item 29 for Grades K through 3, which is the same as Item 27 for Grades 4 through 8, requires the student to identify a new relationship for a "sailboat" that is similar to the relationship between "flashlight" and "radio." In this case, the relationship in common is the source of energy. Sternberg (1982) labels intelligence as the ability to deal with "nonentrenched" tasks that require seeing "old things in new ways" (p. 64). The student must use existing knowledge or skills to solve problems that are unfamiliar or strange. While this knowledge may be affected by previous experience, the inclusion of nonverbal items such as pictures and figures allows the examiner an opportunity to see the student's reasoning ability with content that is least affected by cultural factors.

Although a great number of items have been designed to measure intelligence, analogies have been extremely popular because of their strength in discriminating among abilities. Analogies are tasks that are found in most tests of intellectual aptitude. In fact, Spearman (1923) used analogies as the prototype for intelligent performance.

Piagetian and information processing theorists of intelligence also currently use these tasks because they require the ability to see “second-order relations” (Sternberg, 1982, 1985b; Sternberg & Rifkin, 1979). For young gifted children, White (1985) recommended the use of analogy problems to determine the presence of advanced cognitive capabilities. His earlier studies had indicated that 4- and 5-year-olds not only were able to solve geometric analogy problems, but also were able to verbally justify their responses. Analogies also incorporate many of the behaviors associated with intelligence, such as classification, discrimination, induction, deduction, detail recognition, and, in particular, problem solving (Salvia & Ysseldyke, 1998). Problem solving with analogies has been identified as a general component of intelligent behavior (Mayer, 1992; Resnick & Glaser, 1976; Sternberg, 1982; Sternberg & Detterman, 1986). So while analogical reasoning is one of many behaviors associated with intelligence, analogical reasoning also reflects the level of intellectual functioning of the problem solver.

Item Selection Following the review of tests and the professional literature, we designed an experimental edition of the test with 25 items in each of the core areas for each of three levels: K through 2, 3 through 5, and 6 through 8. We also designed 40 items for the Reasoning subtest for each of three levels: K through 2, 3 through 5, and 6 through 8. We submitted these items to university professors, graduate students, teachers of the gifted, gifted students, and other professionals for critical review. These original items were administered to 1,465 gifted and normal students in Grades K through 2, 1,500 gifted and normal students in Grades 3 through 5, and 1,485 gifted and normal students in Grades 6 through 8. Students identified as gifted were those who were currently enrolled in classrooms for the gifted.

The resulting data were analyzed using the techniques described in the next section. Item discriminating power and item difficulty were ascertained for each item at each of the three levels. Following that analysis, items were revised or discarded. Consequently, a norming version was created that consisted of two levels, Level 1 for students in Grades K through 3 and Level 2 for students in Grades 4 through 8. The norming version for Grades K through 3 consisted of 36 items in Subtest 1: Mathematics/Science, 35 items in Subtest 2: Language Arts/Social Studies, and 40 items in Subtest 3: Reasoning. The norming version for Grades 4 through 8 consisted of 44 items in Subtest 1: Mathematics/Science, 42 items in Subtest 2: Language Arts/Social Studies, and 38 items in Subtest 3: Reasoning. After a second item analysis, 28 items were retained in Subtest 1: Mathematics/Science, 26 items were retained in Subtest 2: Language Arts/Social Studies, and 30 items were retained in Subtest 3: Reasoning for Grades K through 3. For Grades 4 through 8, 30 items were retained in Subtest 1: Mathematics/Science, 30 items were retained in Subtest 2: Language Arts/Social Studies, and 35 items were retained in Subtest 3: Reasoning.

Conventional Item Analysis and Item-Response Theory Modeling

In previous sections, we provided qualitative evidence for the SAGES–2’s content validity. In this section, we provide quantitative evidence for content validity. We report the results of traditional, time-tested procedures used to select good (i.e., valid) items for a test. These procedures focus on the study of an item’s discriminating power and its difficulty.

Item discrimination (sometimes called discriminating power or item validity) refers to “the degree to which an item differentiates correctly among test takers in the behavior that the test is designed to measure” (Anastasi & Urbina, 1997, p. 179). The item discrimination index is actually a correlation coefficient that represents a relationship between a particular item and the other items on the test.

Over 50 different indexes of item discrimination have been developed for use in building tests. In regard to selecting an appropriate index, Anastasi and Urbina (1997), Guilford and Fruchter (1978), and Oosterhof (1976) have observed that, for most purposes, it does not matter which kind of coefficient is used because they all provide similar results.

In the past, test builders have preferred the point-biserial index (probably because it is fairly easy to calculate). Since the development of high-speed computers, however, the item–total-score Pearson correlation index has become increasingly popular and was the method we chose to select items. Ebel (1972) and Pyrczak (1973) suggested that discrimination indexes of .35 or higher are acceptable; Anastasi and Urbina (1997) and Garrett (1965) pointed out that indexes as low as .20 are all right under some circumstances.

The value of using the discrimination index to select good items cannot be overemphasized. A test comprised of too many items that have low indexes of discrimination will very likely have low reliability as well, and a test having low reliability is unlikely to be valid.

Item difficulty (i.e., the percentage of examinees who pass a given item) is determined to identify items that are too easy or too difficult and to arrange items in an easy-to-difficult order. Anastasi and Urbina (1997) wrote that an average difficulty should approximate 50% and have a fairly large dispersion. Items distributed between 15% and 85% are generally considered acceptable. However, for a test such as the SAGES–2, which is a screening test designed for gifted students, items should have difficulty values that come closest to the desired selection ratio (Anastasi & Urbina, 1997). Therefore, the items on the SAGES–2 should be more difficult for the average population. As can be seen in Tables 6.1 (for the SAGES–2:K–3) and 6.2 (for the SAGES–2:4–8), the median item difficulties for the normal normative sample at most ages are below .50.

Table 6.1
 Median Item Difficulties of SAGES-2:K-3 at Five Age Intervals
 (Decimals Omitted)

Sample	SAGES-2 Subtest	Age				
		5	6	7	8	9
Normal	Mathematics/Science	15	6	30	49	75
	Language Arts/Social Studies	18	13	32	55	65
	Reasoning	6	7	9	21	49
Gifted	Mathematics/Science	26	36	64	85	74
	Language Arts/Social Studies	25	34	64	69	71
	Reasoning	40	31	47	58	78

Table 6.2
 Median Item Difficulties of SAGES-2:4-8 at Six Age Intervals
 (Decimals Omitted)

Sample	SAGES-2 Subtest	Age					
		9	10	11	12	13	14
Normal	Mathematics/Science	6	9	14	15	15	15
	Language Arts/Social Studies	7	11	20	23	36	32
	Reasoning	32	27	32	38	43	40
Gifted	Mathematics/Science	18	27	39	43	49	60
	Language Arts/Social Studies	20	28	42	57	69	62
	Reasoning	42	46	56	63	64	42

To demonstrate that the item characteristics of these items were satisfactory, an item analysis was undertaken using the entire normative sample as subjects. The resulting item discrimination coefficients (corrected for part-whole effect) are reported in Tables 6.3 and 6.4 for both forms of the SAGES-2 and both normative samples. In accordance with accepted practice, the statistics reported in these tables are computed only on items that have some variance. On average, the test items satisfy the requirements previously described and provide evidence of content validity.

More recently, item-response theory (IRT) models have increasingly been used for test development (Hambleton & Swaminathan, 1985; Thisser & Wainer, 1990). Parameters of IRT models are available that correspond to the traditional item statistics just described. Item intercept, also known as item location or threshold, corresponds to item difficulty in conventional item analyses. Item slope corresponds to item discrimination. Finally, for tests in which guessing is possible (e.g., multiple-choice formats), a lower asymptote parameter is available that corresponds to the probability of obtaining a correct response by chance.

The procedures just described were used to select items for the SAGES-2. Based on the item difficulty and item discrimination statistics, the corresponding parameters in the IRT models, and an examination of item and test information, unsatisfactory items (i.e., those that did not satisfy the criteria described above) were deleted from the test.

Table 6.3
Median Discriminating Powers of SAGES-2:K-3 at Five Age Intervals
(Decimals Omitted)

Sample	SAGES-2 Subtest	Age				
		5	6	7	8	9
Normal	Mathematics/Science	41	41	55	56	51
	Language Arts/Social Studies	43	45	51	49	47
	Reasoning	63	49	58	56	53
Gifted	Mathematics/Science	51	59	58	50	41
	Language Arts/Social Studies	57	58	58	55	53
	Reasoning	61	57	50	53	53

Table 6.4
Median Discriminating Powers of SAGES-2:4-8 at Six Age Intervals
(Decimals Omitted)

Sample	SAGES-2 Subtest	Age					
		9	10	11	12	13	14
Normal	Mathematics/Science	53	49	53	51	56	46
	Language Arts/Social Studies	43	49	44	50	48	58
	Reasoning	47	44	40	30	31	33
Gifted	Mathematics/Science	52	55	50	47	50	49
	Language Arts/Social Studies	40	47	52	58	63	68
	Reasoning	40	34	56	63	64	42

The “good” items (i.e., those that satisfied the item discrimination and item difficulty criteria) were placed in easy-to-difficult order and compose the final version of the test.

Differential Item Functioning Analysis

The two item-analysis techniques just described (i.e., the study of item difficulty and item discrimination) are traditional and popular. However, no matter how good these techniques are in showing that a test’s items do in fact capture the variance involved in “intelligence,” they are still insufficient. Camilli and Shepard (1994) recommended that test developers need to go further and employ statistical techniques to detect item bias (i.e., use techniques that identify items that give advantages to one group over another group).

To study bias in the SAGES-2 items, we chose to use the logistic regression procedure for detecting differential item functioning (DIF) introduced in 1990 by Swaminathan and Rogers. The logistic regression procedure for detecting DIF is of particular importance because it provides a method for making comparisons between groups when the probabilities of obtaining a correct response for the groups is different at varying ability levels (Mellenberg, 1983). The strategy used in this technique is to compare the full model (i.e., ability, group membership, and the interaction between ability and group membership) with the restricted model (i.e., ability alone) to determine whether

the full model provides a significantly better solution than the restricted model in predicting the score on the item. If the full model is not significantly better at predicting item performance than the restricted model, then the item is measuring differences in ability and does not appear to be influenced by group membership (i.e., the item is not biased).

Logistic regression, when used in the detection of DIF, is a regression technique in which the dependent variable, the item, is scored dichotomously (i.e., correct = 1, incorrect = 0). The full model consists of estimated coefficients for ability (i.e., test score), group membership (e.g., male vs. female), and the interaction between ability and group membership. The restricted model consists of an estimated coefficient for ability only. In most cases, the ability is estimated from the number of correct responses that the examinee has achieved on the test. Because the coefficients in logistic regression are estimated by the maximum likelihood method, the model comparison hypothesis is tested using likelihood ratio statistics. The statistic has a chi-square distribution with 2 degrees of freedom. For our purposes, alpha is set at .01. The authors and two PRO-ED staff members reviewed each item for which comparison between ability and group membership was significant, to determine if the content of the item appeared to be biased against one group. In all, 41 of the original 235 items were eliminated from the final SAGES–2 because the item content was suspect (the others were removed because of low item discrimination indexes). The numbers of items retained in the subtests that were found to be significant at the .01 level of confidence are listed in Tables 6.5 (SAGES–2:K–3) and 6.6 (SAGES–2:4–8) for three dichotomous groups: males versus females, African Americans versus all other ethnic groups, and Hispanic Americans versus all other ethnic groups.

Criterion-Prediction Validity

In the latest edition of their book, Anastasi and Urbina (1997) refer to criterion-prediction validity instead of criterion-related validity. The definition for the new term is the same as that used previously for criterion-related validity, namely “criterion-prediction validation procedures indicate the effectiveness of a test in predicting an individual’s performance in specific activities” (p. 118).

They state that performance on a test is checked against a criterion that can be either a direct or an indirect measure of what the test is designed to predict. Thus, if it is indeed valid, a test like the SAGES–2, which is presumed to measure reasoning ability and academic ability, should correlate well with other tests that are also known or presumed to measure the same abilities.

The correlations may be either concurrent or predictive depending on the amount of time lapsed between the administration of the criterion test and the test being validated.

Table 6.5
Number of Significant Indexes of Bias Relative to
Three Dichotomous Groups for the SAGES-2:K-3

Subtests	Number of Items	Dichotomous Groups		
		Male/ Female	African American/ Non-African American	Hispanic American/ Non-Hispanic American
Mathematics/Science	28	1	0	2
Language Arts/Social Studies	26	1	0	1
Reasoning	30	1	1	2

Table 6.6
Number of Significant Indexes of Bias Relative to Three Dichotomous
Groups for the SAGES-2:4-8

Subtests	Number of Items	Dichotomous Groups		
		Male/ Female	African American/ Non-African American	Hispanic American/ Non-Hispanic American
Mathematics/Science	30	1	2	1
Language Arts/Social Studies	30	2	3	0
Reasoning	35	0	1	0

For example, the correlation between the SAGES-2 and the *Wechsler Intelligence Scale for Children-Third Edition* (Wechsler, 1991) in a situation where one test is given immediately after the other is called concurrent. Anastasi and Urbina (1997) point out that, for certain uses of psychological tests (specifically those uses of the SAGES-2), concurrent validation is the most appropriate type of criterion-prediction validation.

In this section, the results of a number of studies are discussed in terms of their relation to the criterion-prediction validity of the SAGES-2. The characteristics of the studies referred to in this section are discussed below.

In the first study, the criterion-prediction of the SAGES-2:K-3 and the *Gifted and Talented Evaluation Scales* (GATES; Gilliam, Carpenter, & Christensen, 1996) was investigated for 40 students in Brookings, Oregon. The students ranged in age from 6

to 11 years and were in Grades 1 through 5. Eighty-five percent of the students were European American and the other 15% were Hispanic American. Ten of the 40 students were identified as gifted and talented using local school district criteria. Sixty percent of the sample consisted of males; the other 40% were females. The GATES is a behavioral checklist used to identify persons who are gifted and talented. There are five scales on the checklist: Intellectual Ability, Academic Skills, Creativity, Leadership, and Artistic Talent. Each scale has 10 items. Because the SAGES–2 is a measure of reasoning and academic ability, only two of the five scales—Intellectual Ability and Academic Skills—were applicable for our study. These two scales evaluate general intellectual aptitude and academic aptitude, respectively. The SAGES–2 was administered to the students in the Spring of 1999. The GATES was administered approximately 4 months after the SAGES–2 in the Fall of 1999. Teachers rated the students using the GATES. The three subtests of the SAGES–2 were correlated with the two scales of the GATES. Because the GATES has a normative sample consisting only of identified gifted students, all SAGES–2 raw scores were converted to standard scores based on the gifted normative sample before correlating the scores with the GATES.

In the second study, the SAGES–2:K–3 and Total School Ability Index (SAI) of the *Otis–Lennon School Ability Test* (OLSAT; Otis & Lennon, 1996) were correlated. The OLSAT Total SAI examines verbal comprehension, verbal reasoning, pictorial reasoning, figural reasoning, and quantitative reasoning. The subjects of this study, 33 students

Table 6.7
Correlations Between SAGES–2:K–3 Subtests and Criterion Measures

Criterion Measures	SAGES–2 Subtests		
	Mathematics/ Science	Language Arts/ Social Studies	Reasoning
<i>Gifted and Talented Evaluation Scales</i> (Gilliam et al., 1996)			
Intellectual Ability	.32	n.s.	.46
Academic Skills	.38	n.s.	.53
<i>Otis–Lennon School Ability Test</i> (Otis & Lennon, 1996)			
Total School Ability Index	.50	.45	.83

Note. n.s. = not significant.

residing in Arkansas, Illinois, Missouri, and Texas, ranged in age from 7 through 9 years. Thirty-three percent were males, 64% were European American, 15% were African American, and 21% were from other ethnic groups. All students were identified as gifted by their local school districts. The results of these two studies are summarized in Table 6.7.

In the third study, 36 students' scores on the SAGES-2:K-3 and the SAGES-2:4-8 were correlated with their scores on the *Wechsler Intelligence Scale for Children-Third Edition* (WISC-III; Wechsler, 1991). The children in the study ranged in age from 9 through 14 years and resided in Alabama, Georgia, and Mississippi. Sixty-one percent of the students were male, all students were European American, and all students were identified by their school districts as gifted. The results of this study are summarized in Table 6.8.

In the fourth study, 52 students' scores on the SAGES-2:4-8 were correlated with their scores on the OLSAT Nonverbal and Verbal subtests. The OLSAT Nonverbal subtest was correlated with the SAGES-2:4-8 Mathematics/Science and Reasoning subtests and the OLSAT Verbal subtest was correlated with the SAGES-2:4-8 Language/Social Studies subtest. Fifty-two children from Georgia, North Dakota, and Oklahoma, ranging in age from 9 through 13 years, participated in the study. Forty-four percent of the participants were male, 56% were European American, 29% were African American, 12% were Hispanic American, 4% were Asian American, and 21% were identified as gifted by their local school districts. The results of this study are reported in Table 6.9.

The final study examined the criterion prediction validity of the SAGES-2:4-8 scores and the *Stanford Achievement Test-Ninth Edition* (SAT-9; Harcourt Brace Educational Measurement, 1997) Complete Battery scores. The participants of the study were 76 students from Vermont, ranging in age from 9 through 12 years. Fifty-four percent of the students were male, 97% were European American, and 8% were identified as gifted by their local school districts. The results of this study also are summarized in Table 6.9.

Table 6.8

Correlations Between SAGES-2 Subtests and *Wechsler Intelligence Scale for Children-Third Edition* (WISC-III) Full Scale

SAGES-2 Subtests	WISC-III Full Scale
Mathematics/Science	.71
Language Arts/Social Studies	.86
Reasoning	.89

Table 6.9
Correlations Between SAGES–2:4–8 Subtests and Criterion Measures

Criterion Measures	SAGES–2 Subtests		
	Mathematics/ Science	Language Arts/ Social Studies	Reasoning
<i>Otis–Lennon School Ability Test</i> (Otis & Lennon, 1996)			
Verbal	.49	.50	.54
Nonverbal	.61	.61	.64
<i>Stanford Achievement Test–Ninth Edition</i> (Harcourt Brace Educational Measurement, 1997)			
Complete Battery	.57	.47	.53

In all of these studies, raw scores were converted to standard scores that were correlated with the standard scores of the criterion-related tests. As can readily be seen by examining Tables 6.7 through 6.9, the coefficients are high enough to give support for the validity of the SAGES–2 scores.

Construct-Identification Validity

“The construct-identification validity of a test is the extent to which the test may be said to measure a theoretical construct or trait” (Anastasi & Urbina, 1997, p. 126). As such, it relates to the degree to which the underlying traits of the test can be identified and to the extent to which these traits reflect the theoretical model on which the test is based. Linn and Gronlund (1995) offered a three-step procedure for demonstrating this kind of validity. First, several constructs presumed to account for test performance are identified. Second, hypotheses are generated that are based on the identified constructs. Third, the hypotheses are verified by logical or empirical methods. Four basic constructs thought to underlie the SAGES–2 and four testable hypotheses that correspond to these constructs are discussed in the remainder of this chapter:

1. *Age Differentiation*—Because achievement and aptitude are developmental in nature, performance on the SAGES–2 should be strongly correlated to chronological age.

2. *Group Differentiation*—Because the SAGES-2 measures giftedness, its results should differentiate between groups of people known to be average and those identified as gifted.
3. *Subtest Interrelationships*—Because the SAGES-2 subtests measure giftedness (but different aspects of giftedness), they should correlate significantly with each other, but only to a low or moderate degree.
4. *Item Validity*—Because the items of a particular subtest measure similar traits, the items of each subtest should be highly correlated with the total score of that subtest.

Age Differentiation

The means and standard deviations for the SAGES-2:K-3 subtests at five age intervals and the SAGES-2:4-8 subtests at six age intervals are presented in Tables 6.10 and 6.11, respectively. Coefficients showing the relationship of age to performance on the subtests are also found in those tables. The contents of the tables demonstrate that the SAGES-2 subtests are strongly related to age in that their means become larger as the subjects grow older. This observation is verified by the coefficients in the last line of each table, which, according to MacEachron's (1982) rule of thumb interpretations, range in size from moderate to high. These coefficients are high enough to demonstrate the developmental nature of the subtests' contents. Because a relationship with age is a long-acknowledged characteristic of achievement and aptitude, the data found in this table support the construct validity of the SAGES-2.

Group Differentiation

One way of establishing a test's validity is to study the performances of different groups of people on the test. Each group's results should make sense, given what is known about the relationship of the test's content to the group. Thus, in the case of the SAGES-2, which is a test to identify giftedness, one would expect that individuals who have disabilities affecting intelligence and academic performance would do less well than individuals who are identified as gifted. We would certainly anticipate that individuals who are diagnosed as having a learning disability would do more poorly on the test compared with other individuals.

Because being disadvantaged does indeed adversely affect intellectual development in all groups of people, one would assume that groups who are the most disadvantaged would have lower test scores than groups who are less disadvantaged. However, in a test such as the SAGES-2, which was built to minimize the effects of cultural, linguistic, racial, and ethnic bias, any differences among these groups should be minimal and the mean scores of these groups should be within the "normal" (i.e., average) range.

Table 6.10

Means (and Standard Deviations) for SAGES–2:K–3 Subtests at Five Age Intervals and Correlations with Age for Both Normative Samples

Age Interval	SAGES–2 Subtests		
	Mathematics/Science	Language Arts/Social Studies	Reasoning
Normal Sample			
5	6 (3)	5 (3)	5 (5)
6	8 (4)	8 (5)	8 (6)
7	13 (6)	13 (6)	11 (8)
8	17 (6)	17 (6)	16 (9)
9	20 (5)	20 (5)	20 (8)
Correlation with Age	.67	.67	.52
Gifted Sample			
5	9 (4)	10 (3)	9 (8)
6	15 (5)	13 (6)	16 (7)
7	19 (6)	19 (6)	20 (8)
8	22 (3)	21 (4)	24 (6)
9	24 (3)	23 (3)	26 (5)
Correlation with Age	.75	.64	.62

The mean standard scores for both samples used to norm the SAGES–2:K–3 and SAGES–2:4–8 are listed in Tables 6.12 and 6.13, respectively. In addition, the mean standard scores for six subgroups from the normal normative sample are listed in this table. Included are two gender groups (males, females), three ethnic groups (European Americans, African Americans, and Hispanic Americans), and one disability group (learning disability).

The mean standard scores for each gender and ethnic group are all within the normal range (i.e., between 90 and 110). The mean standard scores for the African

Table 6.11

Means (and Standard Deviations) for SAGES-2:4-8 Subtests at Six Age Intervals and Correlations with Age for the Normal Normative Sample

Age Interval	SAGES-2 Subtests		
	Mathematics/Science	Language Arts/Social Studies	Reasoning
Normal Sample			
9	4 (5)	5 (4)	12 (6)
10	5 (4)	6 (5)	14 (6)
11	8 (6)	9 (7)	15 (6)
12	10 (7)	10 (7)	16 (5)
13	12 (8)	11 (8)	17 (5)
14	12 (7)	11 (8)	17 (5)
Correlation with Age	.41	.37	.29
Gifted Sample			
9	8 (6)	8 (5)	16 (5)
10	11 (7)	9 (6)	17 (4)
11	13 (6)	12 (7)	18 (5)
12	16 (6)	15 (7)	18 (5)
13	18 (7)	17 (7)	20 (4)
14	19 (7)	17 (7)	20 (4)
Correlation with Age	.47	.46	.29

American and Hispanic American groups are particularly noteworthy because mean standard scores for these groups are often reported (Neisser et al., 1996) to be a standard deviation or more below average (possibly as a consequence of test bias against these groups). Our findings that the subgroups performed in the normal range on the SAGES-2 are consistent with the findings of Kaufman and Kaufman (1984) in their

Table 6.12
 Standard Score Means for the Entire SAGES-2:K-3
 Normative Sample and Six Subgroups

Subtests	Subgroups of Normal Sample							
	Normative Sample Normal (N = 1,547)	Normative Sample Gifted (N = 836)	Male (N = 795)	Female (N = 752)	European American (N = 1,001)	African American (N = 249)	Hispanic American (N = 263)	Learning Disabled (N = 15)
Mathematics/Science	100	115	100	99	101	96	100	88
Language Arts/Social Studies	100	118	99	101	101	97	98	86
Reasoning	100	117	98	101	99	96	104	82

Table 6.13
 Standard Score Means for the Entire SAGES-2:4-8 Normative Sample and Six Subgroups

Subtests	Subgroups of Normal Sample							
	Normative Sample Normal (N = 1,476)	Normative Sample Gifted (N = 1,454)	Male (N = 792)	Female (N = 684)	European American (N = 1,012)	African American (N = 225)	Hispanic American (N = 177)	Learning Disabled (N = 44)
Mathematics/Science	100	115	101	99	102	93	99	90
Language Arts/Social Studies	101	118	101	101	103	95	101	93
Reasoning	100	117	99	101	101	93	102	86

Interpretative Manual for the *Kaufman Assessment Battery for Children*; Hammill, Pearson, and Wiederholt (1997) in their *Comprehensive Test of Nonverbal Intelligence*; and Newcomer and Hammill (1997) in their *Test of Language Development—Primary: Third Edition*. The authors of these three tests were particularly concerned with sociocultural issues and incorporated many bias-limiting procedures into their tests at the time of construction.

Unfortunately, most authors of achievement and aptitude tests do not report differences in mean standard scores among demographic subgroups in their normative samples. Given the current emphasis on ethnic diversity in the United States and the rising concern about possible test bias, omission of studies showing the performance of various demographic subgroups is a real limitation because test users are denied information they need to help evaluate the appropriateness of a test when given to certain subgroups in the U.S. population. The point is that, although subaverage standard scores made by a particular subgroup are not necessarily evidence that a test is biased against them, average or near-average standard scores are evidence that the test is unbiased.

Support for validity of the SAGES-2 is also seen in the mean standard scores for the learning disability sample. Because the mean standard scores for this disability subgroup are consistent with those reported by other test developers for this subgroup (e.g., Hammill et al., 1997; Kaufman & Kaufman, 1984; McGrew, Werder, & Woodcock, 1991; Naglieri & Das, 1997; Newcomer & Hammill, 1997; Wechsler, 1991), one may conclude that the SAGES-2 measures ability for this group in a valid manner.

Subtest Interrelationships

The SAGES-2:K-3 and SAGES-2:4-8 standard scores for both normative samples were intercorrelated. The resulting coefficients are presented in Tables 6.14 and 6.15. All coefficients are statistically significant at or beyond the .01 level. They range in size from .25 to .45, the median being .38.

Authorities are understandably reluctant to specify precisely how large a correlation coefficient should be in order to serve as evidence of a test's validity. In the case where coefficients representing relationships among subtests of a battery are being evaluated for validity purposes, one would want them all to be statistically significant and "acceptably" high (but not too high). If the SAGES-2 subtest coefficients are too low, it means that the subtests are measuring unrelated abilities rather than differing aspects of achievement and aptitude. If the coefficients are too high, it means that the subtests are measuring the same ability in the same degree and therefore are redundant.

In discussing validity coefficients, Anastasi and Urbina (1997) indicated that under certain circumstances validities as small as .20 or .30 may justify inclusion of a subtest on some battery. Nunnally and Bernstein (1994) observed that validity correlations

Table 6.14

Intercorrelation of SAGES–2:K–3 Subtests for Both Normative Samples
(Decimals Omitted)

Subtest	MS	LS	R
Mathematics/Science (MS)	—	36	35
Language Arts/Social Studies (LS)	31	—	37
Reasoning (R)	27	25	—

Note. Correlations above the diagonal reflect the normal sample; correlations below reflect the gifted sample.

Table 6.15

Intercorrelation of SAGES–2:K–3 Subtests for Both Normative Samples
(Decimals Omitted)

Subtest	MS	LS	R
Mathematics/Science (MS)	—	42	38
Language Arts/Social Studies (LS)	45	—	38
Reasoning (R)	41	38	—

Note. Correlations above the diagonal reflect the normal sample; correlations below reflect the gifted sample.

based on a single predictor rarely exceed .30 or .40. Taking the above figures as guides, one may see that all 12 coefficients reported in Tables 6.14 through 6.15 exceed the .20 criterion of Anastasi and Urbina. Moreover, the median of the 12 coefficients (.38) is within the .30 to .40 range mentioned by Nunnally and Bernstein. Therefore, the coefficients in Tables 6.14 and 6.15 can be accepted as evidence supporting the validity of the SAGES–2 subtests.

Item Validity

Guilford and Fruchter (1978) pointed out that information about a test's construct validity can be obtained by correlating performance on the items with the total score

made on the test. This procedure is also used in the early stages of test construction to select items that have good discriminating power. Strong evidence of the SAGES-2's validity is found in the discriminating powers reported in Tables 6.3 and 6.4. Tests having poor construct-identification validity would unlikely be composed of items having coefficients of the size reported in these tables.

Summary of Validity Results

Based on information provided in this chapter, one may conclude that the SAGES-2 is a valid measure of aptitude and intelligence. Examiners can use the SAGES-2 with confidence, especially when assessing individuals for whom most other tests might be biased.

We encourage professionals to continue to study the test using different samples, statistical procedures, and related measures. We also encourage these professionals to share their results with us so that their findings can be included in subsequent editions of the manual. The accumulation of research data will help further clarify the validity of the SAGES-2 and provide guidance for future revisions of the test.