6 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 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 used 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 literature 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 and accumulative process.

Because the validity of a test's results is relative and dependent on the purpose for which the test will be used, a variety of validity evidence should be accumulated. Most authors of current textbooks dealing with educational and psychological measurement (e.g., Aiken & Groth-Marnat, 2006; Anastasi & Urbina, 1997; Miller, Linn, & Gronlund, 2013) have suggested that those who develop tests should provide evidence of at least three types of validity: content-description validity, criterion-prediction validity, and construct-identification validity. The particular terms we use here are from Anastasi and Urbina (1997). Other authorities (e.g., American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014; Reynolds, Livingston, & Willson, 2009; Salvia, Ysseldyke, & Witmer, 2017) have referred to five categories of validity evidence that are related to test score interpretation (evidence based on test content, response processes, relations to other variables, test structure, and consequences of testing). Although the terms differ somewhat, the concepts they represent are more or less identical. We prefer Anastasi and Urbina's original terms—*content-description*, *criterion prediction*, and *construct-identification*—and describe the evidence of the validity for the SAGES-3 in those terms.

Content-Description Validity

"Content-description validation procedures involve essentially 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 type of validity is of prime importance because it relates to the basic constructs underlying the test and the selection of its items. The determination of content-description validity is a matter of judgment and is closely tied to the procedures used to construct the assessment tool. By determining the rationale underlying the selection of the testing formats and items and of the statistical procedures used to choose good items, test developers generate evidence of a test's content-description validity. Test developers usually deal with this kind of validity by showing that the abilities chosen to be measured are consistent with the current knowledge about a particular area and that the items hold up statistically.

In this section, we provide four demonstrations of content-description validity for the SAGES-3 subtests and composites. First, the rationale for selecting the test format and items is provided. Second, the validity of the items is ultimately supported by the results of conventional item analysis procedures used to choose items during the developmental stages of test construction. Third, the validity of the scores is demonstrated using analysis of floors, ceilings, and item gradients. Fourth, the validity of the items is reinforced by the results of test bias analyses, which show the absence of bias in the test's items.

Rationale Underlying the Selection of Test Formats and Items

During the development of the SAGES-3, current editions of cognitive and academic achievement tests were examined to guide the selection of test formats and items. The following tests were reviewed:

- *Iowa Test of Basic Skills,* Form C (Hoover, Dunbar, & Frisbie, 2007)
- *Kaufman Assessment Battery for Children–Second Edition* (Kaufman & Kaufman, 2004)
- *Kaufman Test of Educational Achievement–Third Edition* (Kaufman & Kaufman, 2014)
- *Naglieri Nonverbal Ability Test–Second Edition* (Naglieri, 2008)
- *Otis–Lennon School Ability Test–Eighth Edition* (Pearson, 2003)
- *Slosson Intelligence Test–Third Edition* (Slosson, Nicholson, & Hibpshman, 2002)
- *Test of Mathematical Abilities for Gifted Students* (Ryser & Johnsen, 1998)
- *Test of Reading Comprehension–Fourth Edition* (V. L. Brown, Wiederholt, & Hammill, 2009)
- *Test of Nonverbal Intelligence–Fourth Edition* (L. Brown, Sherbenou, & Johnsen, 2010)
- *Wechsler Intelligence Scale for Children–Fifth Edition* (Wechsler, 2014)
- *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition* (Wechsler, 2012)
- *Wide Range Achievement Test–Fourth Edition* (Wilkinson & Robertson, 2006)
- *Woodcock Reading Mastery Tests–Third Edition* (Woodcock, 2011)

After reviewing these assessments, we selected test formats that seemed suitable for measuring two major academic areas (language arts/social studies and mathematics/science) and reasoning (verbal and nonverbal). In the following sections we explain the rationales underlying the four SAGES-3 subtests and the procedures used to select the content of their respective items.

Subtest 1: Nonverbal Reasoning

Task Format. The Nonverbal Reasoning subtest measures reasoning (i.e., problem solving) through an analogies format. 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. The items are constructed to vary characteristics related to shading, function, size, shape, position, direction, movement, and mathematical concepts (i.e., number, addition, and part–whole).

Rationale. Reasoning relates to a student's potential to learn the kinds of information necessary to succeed in programs designed for gifted students, and reasoning with analogies is not related to abilities that are formally taught in school. Although a great number of items have been designed to measure reasoning, analogies have been extremely popular because of their strength in discriminating among abilities. Analogies are tasks that are found in most tests of intellectual ability. In fact, Spearman (1923) used analogies as the prototype for intelligent performance and a good measure of *g*, or general intelligence. Piagetian and information processing theorists of intelligence also use these tasks because they require the ability to see "second-order relations" (Sternberg, 1982, 1985; Sternberg & Rifkin, 1979).

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 although analogical reasoning is one of many behaviors associated with intelligence, it also reflects the level of intellectual functioning of the problem solver. Moreover, although knowledge or skills to solve problems that are unfamiliar or strange 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. Also, 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 31 for kindergarten through third-grade students (K–3), which is the same as Item 29 for fourth- through eighth-grade students (4–8), requires the student to identify a new relationship for a "sailboat" that is similar to the relationship between "flashlight" and "iPod." In this case, the relationship in common is the source of energy.

To ensure that this subtest is appropriate for screening giftedness in young school-aged children and demanding enough for older students, we developed an initial bank of 81 nonverbal reasoning items. Following item analysis of data from a study of 1,096 gifted K–3 students and a study of 928 gifted 4–8 students, we selected 33 items for K–3 students and 35 items for 4–8 students in the final version of the subtest.

Subtest 2: Language Arts/Social Studies

Task Format. Students answer a series of multiple-choice questions relating to language arts (e.g., literature, writing) and social studies.

Rationale. The content of this subtest included items from the SAGES-2, as well as new items drawn from current texts, professional literature, books, and the national standards for curriculum. The following texts aided item development:

- Afflerbach et al. (2011). *Reading Street Common Core* (Reading program, Grades K–6). Glenview, IL: Pearson.
- Banks et al. (2003). *Social Studies* (Social studies program, K–12). New York, NY: Macmillan/McGraw-Hill.
- Bauman et al. (2011). *Journeys* (English language arts program, Grades K–6). Orlando, FL: Houghton Mifflin Harcourt.
- Bednarz et al. (2003). *About My World* (Social studies program, K–12). Orlando, FL: Harcourt School Publishers/Holt, Rinehart, Winston.
- Bereiter et al. (2010). *Imagine It!* (Reading and writing program, Grades PreK–6). Columbus, OH: SRA/McGraw-Hill.

To ensure that this subtest is appropriate for screening giftedness in young school-aged children and demanding enough for older students, we developed an initial bank of 140 language arts and social study items. Following item analysis of data from a study of 1,096 gifted K–3 students and a study of 928 gifted 4–8 students, we selected 35 items for K–3 students and 40 items for 4–8 students in the final version of the subtest.

We decided to combine language arts and social studies into a single subtest because this integration is becoming increasingly common with elementary teachers supporting learning activities that incorporate both disciplines (Alleman & Brophy, 2010; Bogle & Ellis, 2009; Strachan, 2015; Whitlock & Fox, 2014). In addition, the *Common Core State Standards in English Language Arts* emphasizes the importance of reading informational texts so "students build a foundation of knowledge in the field that will also give them the background to be better readers in all content areas" (National Governors Association Center for Best Practices [NGA] & Council of Chief State School Officers [CCSSO], 2010b, p. 10). Elementary-grade standards also link language arts with content knowledge in social studies by expecting students to read and comprehend informational texts in history and social studies (NGA & CCSSO, 2010b, p. 13). For example, Item 32 for K–3 and Item 35 for 4–8 are good examples of drawing conclusions from reading historical texts.

As can be seen in Table 6.1, the language arts items are also aligned to specific strands outlined in the *Common Core State Standards in English Language Arts* (NGA & CCSSO, 2010b). For kindergarten through third grade, these strands include Foundational Skills and Speaking and Listening. For kindergarten through eighth grade, these strands include Literature, Writing, and Language. Because of the format of the assessments, we did not assess Speaking in the K–3 or 4–8 subtests, and we did not assess Listening in the 4–8 subtest because students were able to read all of the items.

Table 6.2 shows the alignment of the social studies items to the 10 themes identified in the *National Curriculum Standards for Social Studies: A Framework*

Table 6.1
Alignment of Language Arts Items to the Strands in the *Common Core State Standards in English Language Arts*

Common Core value	Item numbers
ELA strands K–3	
Literature	2, 3, 8, 15, 21, 23, 29, 34
Foundational Skills	1, 6, 35
Writing	17
Speaking and Listening	7, 13
Language	16, 26, 27, 31
ELA strands 4–8	
Literature	1, 2, 8, 20, 21, 31, 36
Writing	10, 11, 27, 28
Language	7, 12, 14, 15, 16, 22, 25, 26, 34, 37

for Teaching, Learning, and Assessment (National Council for the Social Studies [NCSS], 2010). Themes for kindergarten through eighth grade included Culture; Time, Continuity, and Change; People, Places, and Environments; Individual Development and Identity; Individuals, Groups, and Institutions; Power, Authority, and Governance; Production, Distribution, and Consumption; Science, Technology, and Society; Global Connections; and Civic Ideals and Practices.

Subtest 3: Verbal Reasoning

Task Format. The Verbal Reasoning subtest measures reasoning (i.e., problem solving) through an analogies format. This subtest requires the student to solve new problems by identifying relationships among words. For each analogy item, the student is shown three words, two of which are related, and a series of five words. The student is to point to or mark which of the five words relates to the third unrelated word in the same way that the first two words are related. Relationships may include common characteristics, synonyms or antonyms, examples of the other word, categories, functions, causes and effects, or time sequences.

Rationale. Given that analogies are excellent measures of intellectual ability in various formats (i.e., figural, pictorial, or verbal), we decided to add new content to ensure that the full range of abilities might be demonstrated (Lakin & Lohman, 2011). Examining evidence of reasoning in two different symbol systems (words vs. figures/pictures) might improve predictions of potential giftedness in students who have greater abilities with verbal rather than figural systems. Combined with the academic subtests, these two reasoning subtests provide multiple indicators of potential that have a greater likelihood of identifying students with

Table 6.2
Alignment of Social Studies Items to the Themes in the *National Curriculum Standards for Social Studies*

National Curriculum value	Item numbers
Social studies themes K–3	
Culture	14, 28
Time, Continuity, and Change	5, 10
People, Places, and Environments	19, 20, 33
Individual Development and Identity	9
Individuals, Groups, and Institutions	18
Power, Authority, and Governance	32
Production, Distribution, and Consumption	25, 30
Science, Technology, and Society	11
Global Connections	4, 22, 24
Civic Ideals and Practices	12
Social studies themes 4–8	
Culture	8
Time, Continuity, and Change	18, 24, 33, 40
People, Places, and Environments	8, 17
Individual Development and Identity	3, 38
Individuals, Groups, and Institutions	13, 35
Power, Authority, and Governance	19, 30, 39
Production, Distribution, and Consumption	29
Science, Technology, and Society	5, 32
Global Connections	23
Civic Ideals and Practices	4, 6, 9

potential, and predicting future success in gifted programs (Lakin & Lohman, 2011).

On the SAGES-3, special care was taken to include words that younger students were able to read (as indicated by the Dolch word list) and to include items for the older students that were challenging, such as those included on the SAT. We also included relationships requiring flexible and novel kinds of thinking while maintaining an emphasis on convergent skills. For example, Item 25 for K–3 requires the student to identify a new relationship for "crab" that is similar to a relationship between a "woodpecker" and a "hammer."

To ensure that this subtest is appropriate for screening giftedness in young school-aged children and demanding enough for older students, we developed an initial bank of 86 verbal reasoning items. Following item analysis of data from a study of 1,096 gifted K–3 students and a study of 928 gifted 4–8 students, we selected 30 items for K–3 students and 26 items for 4–8 students in the final version of the subtest.

Subtest 4: Mathematics/Science

Task Format. Students answer a series of multiple-choice questions relating to mathematics and science.

Rationale. The content of this subtest included items from the SAGES-2, as well as new items drawn from current texts, professional literature, books, and the national standards for curricula. The following texts aided item development:

- Badders, Bethel, Fu, Peck, Sumners, & Valentino. (2000). *Discovery Works* (Science program, Grades K–6). Boston, MA: Houghton Mifflin.
- Charles et al. (2009). *enVision Math* (Math program, Grades K–6). Glenview, IL: Pearson Scott Foresman.
- Goldenberg, Goldsmith, & Shteingold. (2009). *Think Math* (Math program, Grades K–5). Orlando, FL: Harcourt.
- Moyer, Daniel, Hackett, Baptiste, Stryker, & Vasquez. (2002). *Science* (Science program, Grades K–6). New York, NY: McGraw-Hill.
- Wright Group, University of Chicago STEM Education. (2007). *Everyday Mathematics* (Math program, Grades PreK–6). Columbus, OH: McGraw-Hill.

To ensure that this subtest is appropriate for screening giftedness in young school-aged children and demanding enough for older students, we developed an initial bank of 154 mathematics and science items. Following item analysis of data from a study of 1,096 gifted K–3 students and a study of 928 gifted 4–8 students, we selected 36 items for K–3 students and 34 items for 4–8 students in the final version of the subtest.

We decided to combine science and math into a single subtest because new standards emphasize STEM education and commonalities across disciplines (NGA & CCSSO, 2010a, 2010c; NGSS Lead States, 2013). In fact, within the *Next Generation Science Standards*, connections to *Common Core State Standards in Mathematics* are listed for each disciplinary core idea (NGSS Lead States, 2013). Researchers have suggested that STEM education has evolved into a metadiscipline that removes the traditional barriers between subjects and focuses on applied processes (Frykholm & Glasson, 2005; Kennedy & Odell, 2014). This seamlessness may occur by integrating two or more different branches of mathematics or science (e.g., algebra and geometry or biology and chemistry) or by using a standard from math to solve a problem in science (e.g., measurement, states of matter, and the water cycle) (Adamson, Secada, Maerten-Rivera, & Lee, 2011). For example, Item 22 for K–3 and Item 7 for 4–8 demonstrate this integration by requiring students to use math to solve science problems. We also examined *Progressions for the Common Core State Standards in Mathematics* (Common Core Standards Writing Team, 2013) and sample test items in core content areas from these states: California, Georgia, Illinois, Minnesota, New York, Ohio, Oregon, Tennessee, Texas, and Virginia.

As depicted in Table 6.3, mathematics items within the Mathematics/Science subtest are also closely related to the specific domains identified in the *Common Core State Standards in Mathematics* (NGA & CCSSO, 2010c). The domains for kindergarten through third grade include Counting and Cardinality, and the domains for kindergarten through eighth grade include Operations and Algebraic Thinking, Number and Operations in Base 10, Number and Operations—Fractions, Measurement and Data, and Geometry. The following domains are addressed in fourth through eighth grades only: Ratios and Proportional Relationships, The Number System, Expressions and Equations, and Statistics and Probability.

Table 6.4 shows how the science items within the Mathematics/Science subtest relate to specific domains identified in the *Next Generation Science Standards* (NGSS Lead States, 2013). The items relate to Physical Science, Life Science, Earth and Space Science, and the interdisciplinary area of Engineering, Technology, and Applications of Science.

Table 6.3
Alignment of Mathematics Items to the Domains in the *Common Core State Standards in Mathematics*

Common Core value	Item numbers
CCSSM domains K–3	
Counting and Cardinality	3, 5
Operations and Algebraic Thinking	4, 8, 10, 11, 15, 19, 35
Number and Operations in Base 10	17, 21
Number and Operations—Fractions	16, 25, 27, 32
Measurement and Data	13, 18, 22, 36
Geometry	28
CCSSM domains 4–8	
Operations and Algebraic Thinking	2, 4, 20, 22
Number and Operations in Base 10	31
Number and Operations—Fractions	5, 11
Measurement and Data	6, 7, 8, 19, 23
Geometry	34
Ratios and Proportional Relationships	15, 17
The Number System	9
Expressions and Equations	10, 16, 25, 28
Statistics and Probability	27

Table 6.4
Alignment of Science Items to the Domains in the *Next Generation Science Standards*

Next Generation value	Item numbers
NGSS domains K–3	
Physical Science	6, 26, 30, 31, 34
Life Science	1, 2, 9, 14, 24
Earth and Space Science	7, 12, 20
Engineering, Technology, and Applications of Science	18, 22, 23, 29, 33
NGSS domains 4–8	
Physical Science	30, 32
Life Science	1, 3, 26
Earth and Space Science	14, 18, 24
Engineering, Technology, and Applications of Science	7, 8, 12, 13, 21, 29, 33

Conventional Item Analysis

In previous sections, we provided qualitative evidence for the SAGES-3's content-description validity. In this section, we provide quantitative evidence for this type of 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 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 point-biserial correlation technique, in which each item is correlated with the total test score, was used to determine the item's discriminating power or item validity. Nunnally and Bernstein (1994) noted that items with a discriminating power of .20 or more will likely be satisfactory if the test is long, but in a short test, larger item values are needed. Because our intention is to build relatively short tests that have high reliability, we arbitrarily selected the more conservative value of .30 to serve as the minimum level of acceptability for items on the SAGES-3 subtests. As can be seen in Tables 6.5 and 6.6, all of the median item discrimination coefficients for the SAGES-3: K–3 and SAGES-3: 4–8 were above .40.

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 them in an easy-to-difficult order. Anastasi and Urbina (1997) wrote that an average difficulty should approximate 50% and have a large dispersion. Items distributed between 15% and 85% are generally considered acceptable. However, for a test such as the SAGES-3, which is designed to identify gifted and talented students, items should be more difficult for the average population. As can be

Table 6.5
Median Item Discrimination Coefficients for SAGES-3: K–3 Scores at Five Age Intervals (Decimals Omitted)

	Age (in years)				
Subtest	**5**	**6**	**7**	**8**	**9**
Nonverbal Reasoning	55	53	53	54	59
Language Arts/Social Studies	57	45	50	53	53
Verbal Reasoning	60	58	56	59	60
Mathematics/Science	62	47	53	53	42

Table 6.6
Median Item Discrimination Coefficients for SAGES-3: 4–8 Scores at Six Age Intervals (Decimals Omitted)

	Age (in years)					
Subtest	**9**	**10**	**11**	**12**	**13**	**14**
Nonverbal Reasoning	48	47	49	44	48	52
Language Arts/Social Studies	50	55	58	62	61	65
Verbal Reasoning	54	52	55	49	64	58
Mathematics/Science	48	55	56	61	61	64

seen in Tables 6.7 and 6.8, only 7 out of 44 of the median item difficulties for the SAGES-3: K–3 and SAGES-3: 4–8 were above .50.

On the basis of the item discrimination and item difficulty statistics, unsatisfactory items (i.e., those that did not satisfy the criteria described previously) were deleted from the experimental version of the test before norming the SAGES-3. The items that satisfied the item discrimination and item difficulty criteria were placed in easy-to-difficult order, and the test was normed. For the SAGES-3, in which a few new items were added to every subtest to eliminate ceiling effects, the item analysis procedures were repeated, and the acceptable items of each subtest were arranged in the easy-to-difficult order. As seen in Tables 6.5, 6.6, 6.7, and 6.8, the test items satisfy the requirements previously described and provide evidence of content-description validity.

Floors, Ceilings, and Item Gradients

Experts (e.g., Alfonso & Flanagan, 1999; Bracken, 1987; Rathvon, 2004) have agreed that to be clinically useful, a test's standard scores must have adequate

Table 6.7
Median Item Difficulty Coefficients for SAGES-3: K–3 Scores at Five Age Intervals (Decimals Omitted)

	Age (in years)				
Subtest	5	6	7	8	9
Nonverbal Reasoning	10	11	52	55	58
Language Arts/Social Studies	3	12	26	43	35
Verbal Reasoning	12	11	40	51	56
Mathematics/Science	6	25	43	46	30

Table 6.8
Median Item Difficulty Coefficients for SAGES-3: 4–8 Scores at Six Age Intervals (Decimals Omitted)

	Age (in years)					
Subtest	9	10	11	12	13	14
Nonverbal Reasoning	29	32	41	38	44	41
Language Arts/Social Studies	11	15	24	24	35	42
Verbal Reasoning	19	29	32	39	51	54
Mathematics/Science	16	13	24	25	39	41

floors, ceilings, and item gradients. Test publishers can take steps to ensure the adequacy of these characteristics during development. First, items should be developed, reviewed, and selected to include those with an average difficulty of 50% and a distribution of difficulty between 15% and 85%. Second, examinees who vary widely in ability should be included in the normative sample to extend the range of scores and confirm that the test adequately measures their abilities. Finally, standard scores in the normative tables should be smoothed to ensure that an increase in 1 raw score point (or age level) does not result in an increase of more than 5 standard score points. In the following sections, we review the SAGES-3 floors, ceilings, and item gradients.

Floors

A *test floor* refers to the lowest obtainable standard score when only one or fewer items are answered correctly. Tests that do not have sufficiently low floors cannot accurately identify individuals with very low ability or differentiate among those who function at that level. Bracken (1987) suggested that to be considered adequate, the average floor has to be at or below a standard score of 70. In this section, we discuss the SAGES-3 subtest and composite floors. Because the SAGES-3 is designed to assess students with high ability rather than very low ability, however, we would neither expect nor require its floors to be adequate across all ages.

The adequacy of the SAGES-3 floors was evaluated according to standards originally suggested by Bracken (1987). According to these standards, a floor or ceiling that fails to assess the functioning of 10.01% or more of the population at either end of the distribution is considered poor; omission of the extreme from 7.01% to 10.00% of the population is considered fair; omission from 5.01% to 7.00% is considered good; omission from 3.01% to 5.00% is considered very good; and omission from .01% to 3.00% is considered excellent.

The results from analyses of average subtest floors are reported in Tables 6.9 and 6.10. The subtest floors were evaluated across the entire age range of the SAGES-3: K–3 and the SAGES-3: 4–8. As expected, the average subtest floors for the SAGES-3: K–3 ranged from poor to excellent. The average subtest floors for the SAGES-3: 4–8 ranged from very good to excellent.

Flanagan and Alfonso (1995) suggested that for tests intended to identify a disability, the floors of composite measures are more important than those of subtests because recommendations for additional services are primarily based on these total scores. It stands to reason, therefore, that in a test intended to identify giftedness, the floors of the composites are less relevant than the ceilings. Nonetheless, we examined the floors for the composites. Composite floors represent the lowest possible standard scores (derived from the sum of subtest standard scores) for individuals who obtained the lowest possible raw score on all contributing subtests. The results of this analysis appear in Tables 6.11 through 6.16. As expected, the SAGES-3: K–3 floors for the Reasoning Ability, Academic Ability, and General Ability composites ranged from poor to excellent. The average composite floors for the SAGES-3: 4–8 were all excellent for all three composites.

Ceilings

A *test ceiling* refers to the highest obtainable standard score when all items are answered correctly. Tests that do not have sufficiently high ceilings cannot accurately identify individuals with very high ability or differentiate among those who function at that level. To be considered adequate, the average ceiling has to be at or above a standard score of 130. The adequacy of the SAGES-3 ceilings was evaluated according to the same criteria used to evaluate floors. In this section, we review the SAGES-3 subtest and composite ceilings.

The results from analyses of average subtest ceilings are reported in Tables 6.9 and 6.10. Because the SAGES-3 is intended to identify giftedness, the average subtest ceilings were evaluated across the entire age range. The ceiling of a test is determined by the extent to which there are sufficient difficult items to distinguish between examinees of average ability and examinees of above-average ability. As can be seen in Tables 6.9 and 6.10, the average SAGES-3 subtest ceilings (i.e., the subtest standard scores associated with a perfect raw score for each subtest) were all excellent.

Composite ceilings are the highest possible standard scores on scales, given the sum of possible subtest standard scores. According to Flanagan and Alfonso (1995), ceilings of composite measures are more important than those of subtests

Table 6.9
Average SAGES-3: K–3 Subtest Floors and Ceilings for Each Age Level

	Floor				Ceiling				Total test
Age level	Average subtest standard score associated with a raw score of 1	Number of standard deviations below the mean	Will assess all but the lowest ____ percent	Adequacy of average floor	Average subtest standard score associated with highest obtained raw score	Number of standard deviations above the mean	Will assess all but the highest ____ percent	Adequacy of average ceiling	Percentage of total population served
5-0 to 5-2	92.75	0.48	31.44	Poor	160.00	4.00	.00	Excellent	68.55
5-3 to 5-5	92.08	0.53	29.88	Poor	160.00	4.00	.00	Excellent	70.11
5-6 to 5-8	88.50	0.77	22.16	Poor	160.00	4.00	.00	Excellent	77.83
5-9 to 5-11	88.50	0.77	22.16	Poor	160.00	4.00	.00	Excellent	77.83
6-0 to 6-2	88.25	0.78	21.67	Poor	158.75	3.92	.00	Excellent	78.32
6-3 to 6-5	88.08	0.79	21.35	Poor	158.75	3.92	.00	Excellent	78.65
6-6 to 6-8	84.33	1.04	14.81	Poor	156.25	3.75	.01	Excellent	85.18
6-9 to 6-11	82.00	1.20	11.51	Poor	156.25	3.75	.01	Excellent	88.48
7-0 to 7-2	78.50	1.43	7.59	Fair	153.25	3.55	.02	Excellent	92.39
7-3 to 7-5	77.42	1.51	6.61	Good	152.00	3.47	.03	Excellent	93.36
7-6 to 7-8	74.08	1.73	4.20	Very good	150.75	3.38	.04	Excellent	95.76
7-9 to 7-11	72.00	1.87	3.10	Very good	150.00	3.33	.04	Excellent	96.86
8-0 to 8-2	70.00	2.00	2.28	Excellent	148.25	3.22	.06	Excellent	97.66
8-3 to 8-5	69.50	2.03	2.10	Excellent	147.50	3.17	.08	Excellent	97.82
8-6 to 8-8	68.17	2.12	1.69	Excellent	147.50	3.17	.08	Excellent	98.23
8-9 to 8-11	66.75	2.22	1.33	Excellent	147.50	3.17	.08	Excellent	98.59
9-0 to 9-11	66.50	2.23	1.28	Excellent	146.25	3.08	.10	Excellent	98.62

Table 6.10
Average SAGES-3: 4–8 Subtest Floors and Ceilings for Each Age Level

	Floor				Ceiling				Total test
Age level	Average subtest standard score associated with a raw score of 1	Number of standard deviations below the mean	Will assess all but the lowest ______percent	Adequacy of average floor	Average subtest standard score associated with highest obtained raw score	Number of standard deviations above the mean	Will assess all but the highest ______percent	Adequacy of average ceiling	Percentage of total population served
9-0 to 9-3	75.00	1.67	4.78	Very good	155.00	3.67	.01	Excellent	95.21
9-4 to 9-7	74.38	1.71	4.38	Very good	155.00	3.67	.01	Excellent	95.61
9-8 to 9-11	73.75	1.75	4.01	Very good	154.75	3.65	.01	Excellent	95.98
10-0 to 10-3	72.50	1.83	3.34	Very good	152.50	3.50	.02	Excellent	96.64
10-4 to 10-7	71.88	1.88	3.04	Very good	152.50	3.50	.02	Excellent	96.94
10-8 to 10-11	70.25	1.98	2.37	Excellent	151.25	3.42	.03	Excellent	97.60
11-0 to 11-11	70.25	1.98	2.37	Excellent	148.25	3.22	.06	Excellent	97.57
12-0 to 12-11	69.00	2.07	1.94	Excellent	146.50	3.10	.10	Excellent	97.96
13-0 to 13-11	68.25	2.12	1.71	Excellent	143.25	2.88	.20	Excellent	98.09
14-0 to 14-11	67.75	2.15	1.58	Excellent	142.00	2.80	.26	Excellent	98.17

because important clinical and academic decisions are primarily based on these total scores. The results of this analysis appear in Tables 6.11 through 6.16. As can be seen in these tables, the ceilings for the three composite indexes were all excellent for the SAGES-3: K–3 and the SAGES-3: 4–8. These results indicate that the SAGES-3 has substantial and consistently excellent ceilings, even for the oldest and brightest examinees whose abilities may be assessed by the instrument.

Item Gradients

Item gradients refers to how rapidly standard scores increase as a function of examinees' success or failure on a single test item (Bracken, 1987). Tests and subtests with smaller increments in standard scores relative to single raw score points are more effective, sensitive, and finely tuned as measures of an examinee's true ability. A test's item gradient should not be so steep that an increase or decrease in a single raw score point results in a subtest standard score change of more than 1/3 standard deviation (.33 *SD*). Likewise, an increase or decrease in 1 sum-of-standard-score point should not result in an index change of more than 1/3 standard deviation. Item gradients that are steeper than this criterion result in little differentiation of ability.

The SAGES-3 normative tables were all smoothed to conform to the recommended standard that an increase or decrease in a single raw-score point did not result in a standard score change of more than 1/3 standard deviation (i.e., 5 standard score points). This procedure, in conjunction with the adequacy of the test's ceilings, resulted in SAGES-3 subtest and composite difficulty gradients that are consistently adequate for detecting minor fluctuations in examinees' abilities.

Analyses of Test Bias

We provide two studies of test bias. The first of these uses differential item functioning (DIF) analysis to detect possible bias at the item level. The second examines subgroup performance to detect possible bias at the subtest and composite index levels.

Differential Item Functioning Analysis

The two item analysis techniques described in the previous section (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 giftedness, they are still incomplete. Camilli and Shepard (1994) recommended that test developers go further and perform statistical tests for item bias. *Item bias*, also known as *differential item functioning*, is said to exist when examinees from different racial or gender groups who have the same ability level perform differently on the same item (i.e., evidence indicates that one group has an advantage over another on that item). The procedures used to identify biased items are described in this section.

The logistic regression procedure developed by Swaminathan and Rogers (1990) is used for detecting DIF. This procedure compares the adequacy of two

(*text continues on p. 72*)

Table 6.11
SAGES-3: K–3 Composite Floors and Ceilings for Each Age Level: Reasoning Ability Index

	Floor				Ceiling				Total test
Age level	Lowest possible standard score	Number of standard deviations below the mean	Will assess all but the lowest ____ percent	Adequacy of average floor	Highest possible standard score	Number of standard deviations above the mean	Will assess all but the highest ____ percent	Adequacy of average ceiling	Percentage of total population served
5-0 to 5-2	84	1.07	14.31	Poor	160	4.00	.00	Excellent	85.69
5-3 to 5-5	84	1.07	14.31	Poor	160	4.00	.00	Excellent	85.69
5-6 to 5-8	79	1.40	8.08	Fair	160	4.00	.00	Excellent	91.92
5-9 to 5-11	79	1.40	8.08	Fair	160	4.00	.00	Excellent	91.92
6-0 to 6-2	79	1.40	8.08	Fair	159	3.93	.00	Excellent	91.92
6-3 to 6-5	79	1.40	8.08	Fair	159	3.93	.00	Excellent	91.92
6-6 to 6-8	75	1.67	4.78	Very good	157	3.80	.01	Excellent	95.21
6-9 to 6-11	75	1.67	4.78	Very good	157	3.80	.01	Excellent	95.21
7-0 to 7-2	69	2.07	1.94	Excellent	152	3.47	.03	Excellent	98.04
7-3 to 7-5	69	2.07	1.94	Excellent	150	3.33	.04	Excellent	98.02
7-6 to 7-8	64	2.40	0.82	Excellent	147	3.13	.09	Excellent	99.09
7-9 to 7-11	64	2.40	0.82	Excellent	147	3.13	.09	Excellent	99.09
8-0 to 8-2	53	3.13	0.09	Excellent	144	2.93	.17	Excellent	99.75
8-3 to 8-5	53	3.13	0.09	Excellent	144	2.93	.17	Excellent	99.75
8-6 to 8-8	53	3.13	0.09	Excellent	144	2.93	.17	Excellent	99.75
8-9 to 8-11	53	3.13	0.09	Excellent	144	2.93	.17	Excellent	99.75
9-0 to 9-11	53	3.13	0.09	Excellent	141	2.73	.31	Excellent	99.60

Table 6.12
SAGES-3: K–3 Composite Floors and Ceilings for Each Age Level: Academic Ability Index

	Floor				Ceiling				Total test
Age level	Lowest possible standard score	Number of standard deviations below the mean	Will assess all but the lowest ____ percent	Adequacy of average floor	Highest possible standard score	Number of standard deviations above the mean	Will assess all but the highest ____ percent	Adequacy of average ceiling	Percentage of total population served
5-0 to 5-2	88	0.80	21.19	Poor	160	4.00	.00	Excellent	78.81
5-3 to 5-5	88	0.80	21.19	Poor	160	4.00	.00	Excellent	78.81
5-6 to 5-8	85	1.00	15.87	Poor	160	4.00	.00	Excellent	84.13
5-9 to 5-11	85	1.00	15.87	Poor	160	4.00	.00	Excellent	84.13
6-0 to 6-2	85	1.00	15.87	Poor	160	4.00	.00	Excellent	84.13
6-3 to 6-5	84	1.09	13.81	Poor	160	4.00	.00	Excellent	86.19
6-6 to 6-8	79	1.40	8.08	Fair	160	4.00	.00	Excellent	91.92
6-9 to 6-11	76	1.60	5.48	Good	160	4.00	.00	Excellent	94.52
7-0 to 7-2	76	1.60	5.48	Good	160	4.00	.00	Excellent	94.52
7-3 to 7-5	74	1.73	4.15	Very good	160	4.00	.00	Excellent	95.85
7-6 to 7-8	69	2.07	1.94	Excellent	160	3.98	.00	Excellent	98.06
7-9 to 7-11	64	2.40	0.82	Excellent	159	3.93	.00	Excellent	99.18
8-0 to 8-2	61	2.60	0.47	Excellent	159	3.93	.00	Excellent	99.53
8-3 to 8-5	59	2.73	0.31	Excellent	159	3.93	.00	Excellent	99.68
8-6 to 8-8	56	2.91	0.18	Excellent	159	3.93	.00	Excellent	99.82
8-9 to 8-11	53	3.13	0.09	Excellent	159	3.93	.00	Excellent	99.91
9-0 to 9-11	51	3.27	0.05	Excellent	159	3.93	.00	Excellent	99.94

Table 6.13
SAGES-3: K–3 Composite Floors and Ceilings for Each Age Level: General Ability Index

	Floor				Ceiling				Total test
Age level	Lowest possible standard score	Number of standard deviations below the mean	Will assess all but the lowest _____ percent	Adequacy of average floor	Highest possible standard score	Number of standard deviations above the mean	Will assess all but the highest _____ percent	Adequacy of average ceiling	Percentage of total population served
5-0 to 5-2	85	1.00	15.87	Poor	160	4.00	.00	Excellent	84.13
5-3 to 5-5	84	1.04	14.81	Poor	160	4.00	.00	Excellent	85.18
5-6 to 5-8	80	1.33	9.12	Fair	160	4.00	.00	Excellent	90.88
5-9 to 5-11	80	1.33	9.12	Fair	160	4.00	.00	Excellent	90.88
6-0 to 6-2	80	1.33	9.12	Fair	160	4.00	.00	Excellent	90.88
6-3 to 6-5	79	1.38	8.41	Fair	160	4.00	.00	Excellent	91.58
6-6 to 6-8	75	1.64	5.00	Very good	160	4.00	.00	Excellent	94.99
6-9 to 6-11	74	1.73	4.15	Very good	160	4.00	.00	Excellent	95.85
7-0 to 7-2	71	1.93	2.66	Excellent	160	4.00	.00	Excellent	97.34
7-3 to 7-5	70	1.98	2.40	Excellent	160	4.00	.00	Excellent	97.60
7-6 to 7-8	66	2.29	1.10	Excellent	160	4.00	.00	Excellent	98.89
7-9 to 7-11	63	2.47	0.68	Excellent	160	4.00	.00	Excellent	99.31
8-0 to 8-2	59	2.73	0.31	Excellent	160	4.00	.00	Excellent	99.68
8-3 to 8-5	58	2.78	0.27	Excellent	159	3.93	.00	Excellent	99.72
8-6 to 8-8	57	2.89	0.19	Excellent	159	3.93	.00	Excellent	99.80
8-9 to 8-11	55	3.00	0.13	Excellent	159	3.93	.00	Excellent	99.86
9-0 to 9-11	55	3.00	0.13	Excellent	158	3.87	.01	Excellent	99.86

Table 6.14
SAGES-3: 4–8 Composite Floors and Ceilings for Each Age Level: Reasoning Ability Index

	Floor				Ceiling				Total test
Age level	Lowest possible standard score	Number of standard deviations below the mean	Will assess all but the lowest ______percent	Adequacy of average floor	Highest possible standard score	Number of standard deviations above the mean	Will assess all but the highest ______percent	Adequacy of average ceiling	Percentage of total population served
9-0 to 9-3	65	2.33	.98	Excellent	160	4.00	.00	Excellent	99.02
9-4 to 9-7	64	2.43	.75	Excellent	160	4.00	.00	Excellent	99.25
9-8 to 9-11	62	2.53	.56	Excellent	160	4.00	.00	Excellent	99.43
10-0 to 10-3	62	2.53	.56	Excellent	156	3.73	.01	Excellent	99.43
10-4 to 10-7	61	2.63	.42	Excellent	156	3.73	.01	Excellent	99.57
10-8 to 10-11	59	2.73	.31	Excellent	156	3.73	.01	Excellent	99.68
11-0 to 11-11	59	2.73	.31	Excellent	155	3.67	.01	Excellent	99.67
12-0 to 12-11	57	2.87	.21	Excellent	155	3.67	.01	Excellent	99.78
13-0 to 13-11	57	2.87	.21	Excellent	153	3.53	.02	Excellent	99.77
14-0 to 14-11	57	2.87	.21	Excellent	153	3.53	.02	Excellent	99.77

Table 6.15
SAGES-3: 4–8 Composite Floors and Ceilings for Each Age Level: Academic Ability Index

	Floor				Ceiling				Total test
Age level	Lowest possible standard score	Number of standard deviations below the mean	Will assess all but the lowest ____ percent	Adequacy of average floor	Highest possible standard score	Number of standard deviations above the mean	Will assess all but the highest ____ percent	Adequacy of average ceiling	Percentage of total population served
9-0 to 9-3	64	2.40	.82	Excellent	160	4.00	.00	Excellent	99.18
9-4 to 9-7	64	2.40	.82	Excellent	160	4.00	.00	Excellent	99.18
9-8 to 9-11	64	2.40	.82	Excellent	160	4.00	.00	Excellent	99.18
10-0 to 10-3	64	2.40	.82	Excellent	158	3.87	.01	Excellent	99.17
10-4 to 10-7	61	2.60	.47	Excellent	158	3.87	.01	Excellent	99.53
10-8 to 10-11	58	2.80	.26	Excellent	156	3.73	.01	Excellent	99.74
11-0 to 11-11	58	2.80	.26	Excellent	151	3.40	.03	Excellent	99.71
12-0 to 12-11	53	3.13	.09	Excellent	148	3.20	.07	Excellent	99.84
13-0 to 13-11	50	3.33	.04	Excellent	143	2.87	.21	Excellent	99.75
14-0 to 14-11	42	3.87	.01	Excellent	141	2.73	.31	Excellent	99.68

Table 6.16
SAGES-3: 4–8 Composite Floors and Ceilings for Each Age Level: General Ability Index

	Floor				Ceiling				Total test
Age level	Lowest possible standard score	Number of standard deviations below the mean	Will assess all but the lowest _____ percent	Adequacy of average floor	Highest possible standard score	Number of standard deviations above the mean	Will assess all but the highest _____ percent	Adequacy of average ceiling	Percentage of total population served
9-0 to 9-3	64	2.40	.82	Excellent	160	4.00	.00	Excellent	99.18
9-4 to 9-7	63	2.47	.68	Excellent	160	4.00	.00	Excellent	99.31
9-8 to 9-11	62	2.53	.56	Excellent	160	4.00	.00	Excellent	99.43
10-0 to 10-3	62	2.53	.56	Excellent	160	4.00	.00	Excellent	99.43
10-4 to 10-7	60	2.67	.38	Excellent	160	4.00	.00	Excellent	99.61
10-8 to 10-11	59	2.73	.31	Excellent	160	4.00	.00	Excellent	99.68
11-0 to 11-11	59	2.73	.31	Excellent	160	4.00	.00	Excellent	99.68
12-0 to 12-11	57	2.87	.21	Excellent	158	3.87	.01	Excellent	99.79
13-0 to 13-11	56	2.93	.17	Excellent	154	3.60	.02	Excellent	99.82
14-0 to 14-11	55	3.00	.13	Excellent	152	3.47	.03	Excellent	99.84

different logistic regression models to account for the ability being measured; the first model uses ability (i.e., the subtest score) alone to predict item performance (restricted model), and the second model uses ability and group membership to predict item performance (full model). This technique compares the full model with the restricted model to determine whether the full model provides a significantly better solution. If the full model does not provide a significantly better solution than the restricted model, then the differences between groups on the item are best explained by ability alone. In other words, if the full model is not significantly better than the restricted model at predicting item performance, 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). Stated another way, if the full model is significantly better than the restricted model at predicting item performance, the item is said to exhibit uniform DIF. Uniform DIF occurs when one group consistently performs better on the item than does the other group at all levels of ability.

To distinguish statistical significance from practical significance, we had to establish criteria for significance and magnitude. All items on both forms of the SAGES-3 were analyzed, and comparisons were made for each of the focus groups compared to the reference groups (female vs. male, Black/African American vs. non–Black/African American, and Hispanic vs. non-Hispanic). Because 807 comparisons were made for these analyses, a significance level of .001 was adopted to prevent the overidentification of potentially biased items that might occur when large numbers of comparisons are made. Although strict Bonferroni correction (.05/807 number of comparisons) would have resulted in a significance level of .00006, we opted for .001, because the more strict adjustment to the alpha level might have prevented the detection of any biased items.

Next, for those items that were flagged as statistically significant, an effect size was used to evaluate the magnitude or amount of DIF. Zumbo (1999) suggested using the R^2 difference (ΔR^2, a weighted least-squared effect size) between the restricted model and the full model to determine the degree of an item's DIF. Using R. J. Cohen, Swerdlik, and Smith's (1992) conventions for small, medium, and large effects, Jodoin and Gierl (2001) suggested that an R^2 difference less than .035 indicates negligible DIF, R^2 greater than .034 but less than .070 indicates moderate DIF, and R^2 greater than .069 indicates large DIF. Because we are interested only in items that may be meaningfully biased, items with moderate or large effect sizes were targeted for possible removal from the test.

Using the entire normative sample as participants, we applied the logistic regression procedure to all items contained in each SAGES-3 subtest and made comparisons between three dichotomous groups: male versus female, Black/African American versus non–Black/African American, and Hispanic versus non-Hispanic. Comparisons found to be statistically significant at the .001 level are reported in Table 6.17.

Four SAGES-3 item comparisons were found to be statistically significant at the .001 level. Two SAGES-3: K–3 Nonverbal Reasoning items were significant; one favored males, and one favored females. One SAGES-3: K–3 Verbal Reasoning item was significant and favored males. One SAGES-3: 4–8 Verbal Reasoning item was significant and favored non-Hispanic examinees. Further investigation of the meaningfulness of these results revealed that all of the statistically significant comparisons had negligible effect sizes according to Jodoin and Gierl's (2001) criteria. All significant items were further examined for content. It was

Table 6.17
Number of SAGES-3 Items With Significant Effect Sizes for Selected Subgroups

		Dichotomous groups		
SAGES-3 subtest	Number of items	Male/female	Black/African American/ non–Black/African American	Hispanic/ non-Hispanic
Nonverbal Reasoning K–3	33	0(2)	0(0)	0(0)
Nonverbal Reasoning 4–8	35	0(0)	0(0)	0(0)
Language Arts/Social Studies K–3	35	0(0)	0(0)	0(0)
Language Arts/Social Studies 4–8	40	0(0)	0(0)	0(0)
Verbal Reasoning K–3	30	0(1)	0(0)	0(0)
Verbal Reasoning 4–8	26	0(0)	0(0)	0(1)
Mathematics/Science K–3	36	0(0)	0(0)	0(0)
Mathematics/Science 4–8	34	0(0)	0(0)	0(0)

Note. Numbers inside parentheses represent the number of statistically significant items for each subgroup; numbers outside parentheses represent the number of moderate or large effect sizes detected for each group.

determined that regardless of statistical significance, the differences were not meaningful. Therefore, one may conclude that the SAGES-3 items possess little or no systematic bias in regard to gender, race, and ethnicity.

Demographic Subgroup Comparisons

Two studies are described in this section. In the first study, we present the mean subtest and composite indexes for selected demographic subgroups in the normative sample. In the second study, we present the results of mean difference analyses between selected demographic subgroups and a demographically matched comparison sample from the entire pool of SAGES-3 examinees.

First, we examined the mean subtest and composite indexes for three mainstream subgroups (males, females, Whites) and four minority subgroups (Black/African American, Asian/Pacific Islander, Hispanic, and two or more races) from the SAGES-3: K–3 and SAGES-3: 4–8 normative samples. Because special attention was devoted to controlling racial and gender bias during item development, one would expect that all subgroups would score in the *unlikely* probability of giftedness range (i.e., between 90 and 109 points) on the SAGES-3: K–3 and the SAGES-3: 4–8. Tables 6.18 and 6.19 indicate that with the exception of the higher scoring Asian/Pacific Islander subgroup, subtest and composite indexes were within the *unlikely* range. This is consistent with studies examining measures of IQ (see Rushton & Jensen, 2005), in which Asian/Pacific Islander examinees scored higher than the other groups. Overall, these tables provide further evidence for the fairness of the test for both mainstream and minority subgroups.

Table 6.18
SAGES-3: K–3 Index Means (and Standard Deviations) for Selected Subgroups in the Normative Sample

	Male	Female	White	Black/ African American	Asian/ Pacific Islander	Two or more races	Hispanic
	(*n* = 407)	(*n* = 401)	(*n* = 604)	(*n* = 140)	(*n* = 28)	(*n* = 33)	(*n* = 190)
SAGES-3: K–3 value	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Subtest							
Nonverbal Reasoning	99 (16)	101 (15)	100 (15)	93 (15)	109 (13)	104 (17)	98 (14)
Language Arts/Social Studies	101 (15)	100 (15)	102 (14)	93 (14)	109 (18)	100 (13)	97 (14)
Verbal Reasoning	100 (15)	99 (14)	101 (14)	94 (14)	110 (13)	102 (12)	96 (14)
Mathematics/Science	101 (15)	99 (14)	101 (14)	92 (13)	109 (19)	100 (11)	95 (12)
Composite							
Reasoning Ability	99 (15)	100 (15)	101 (15)	92 (15)	110 (13)	103 (13)	96 (14)
Academic Ability	101 (15)	99 (15)	102 (14)	91 (13)	110 (20)	100 (12)	95 (13)
General Ability	100 (15)	100 (15)	101 (15)	91 (13)	112 (17)	102 (12)	95 (13)

Table 6.19
SAGES-3: 4–8 Index Means (and Standard Deviations) for Selected Subgroups in the Normative Sample

	Male	Female	White	Black/ African American	Asian/ Pacific Islander	Two or more races	Hispanic
	(*n* = 513)	(*n* = 510)	(*n* = 765)	(*n* = 153)	(*n* = 46)	(*n* = 45)	(*n* = 251)
SAGES-3: 4–8 value	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
Subtest							
Nonverbal Reasoning	99 (15)	100 (14)	100 (14)	95 (16)	107 (13)	96 (15)	96 (14)
Language Arts/Social Studies	99 (15)	101 (15)	100 (14)	96 (17)	108 (15)	98 (15)	96 (14)
Verbal Reasoning	99 (15)	101 (15)	101 (15)	95 (14)	106 (15)	96 (15)	96 (14)
Mathematics/Science	100 (15)	100 (14)	101 (15)	95 (15)	107 (12)	93 (11)	96 (15)
Composite							
Reasoning Ability	99 (15)	101 (15)	101 (15)	95 (15)	108 (13)	96 (15)	95 (14)
Academic Ability	99 (15)	101 (14)	101 (14)	95 (17)	109 (13)	95 (13)	96 (15)
General Ability	99 (15)	101 (15)	101 (15)	95 (15)	109 (13)	95 (15)	95 (14)

Next, we examined the mean differences between selected subgroups and a control sample matched on key demographic variables (age, gender, race, parent education) on the SAGES-3: K–3 and the SAGES-3: 4–8. Some test users consider mean score differences between subgroups an index of bias. The underlying assumption is that groups should perform approximately equally (e.g., within 1 *SEM*), and if they do not, the test is biased against the subgroup obtaining the lower scores. The subgroup and comparison group were matched (where appropriate) on gender, race, Hispanic status, and socioeconomic status (indicated by parent education level). Examinees with exceptionalities or disabilities were excluded from the study. Specifically, the following groups were compared: male and female examinees, Black/African American and White examinees, Asian/Pacific Islander and White examinees, and Hispanic and non-Hispanic examinees. The demographics for these samples (i.e., the selected subgroups and their matched counterparts) are presented in Tables 6.20 and 6.21.

Subgroup mean scores, standard deviations, score differences, and effect sizes are presented for each of the comparisons, which are discussed next. Both Cohen's *d* and effect size *r* are presented in these studies. Hopkins (2002)

Table 6.20
Demographic Characteristics of the Samples Used in the SAGES-3: K–3 Demographic Subgroup Comparison Studies

	Demographic subgroup			
Sample characteristic	Gender	Black/African American	Asian/Pacific Islander	Hispanic
Total number of participants	658	290	46	384
Age (in years)	5–9	5–9	5–9	5–9
Gender				
Male	329	128	20	182
Female	329	162	26	202
Race				
White	508	145	23	352
Black/African American	116	145	0	12
Asian/Pacific Islander	14	0	23	4
Two or more races	18	0	0	16
Hispanic status				
Yes	176	12	4	192
No	482	278	42	192
Parent education				
Less than Bachelor's degree	474	212	30	280
Bachelor's degree	184	78	16	104

Table 6.21
Demographic Characteristics of the Samples Used in the SAGES-3: 4–8 Demographic Subgroup Comparison Studies

	Demographic subgroup			
Sample characteristic	Gender	Black/African American	Asian/Pacific Islander	Hispanic
Total number of participants	630	246	60	382
Age (in years)	9–14	9–14	9–14	9–14
Gender				
Male	315	110	32	204
Female	315	136	28	178
Race				
White	146	123	30	344
Black/African American	92	123	0	16
Asian/Pacific Islander	24	0	30	0
American Indian/Alaska Native	2	0	0	4
Two or more races	16	0	0	18
Hispanic status				
Yes	148	16	0	191
No	482	230	60	191
Parent education				
Less than Bachelor's degree	448	176	40	276
Bachelor's degree	182	70	20	106

described effect size *r* in six categories: *r*s less than .10 are considered very small or trivial, between .10 and .29 are considered small, between .30 and .49 are considered moderate, between .50 and .69 are considered large, between .70 and .89 are considered very large, and .90 and above are considered nearly perfect. Hopkins also described Cohen's *d* in six categories: *d*s less than .20 are very small or trivial, between .20 and .59 are small, between .60 and 1.19 are moderate, between 1.20 and 1.99 are large, between 2.00 and 3.99 are very large, and 4.00 and above are nearly perfect. All of the studies of group differences are discussed in the following sections.

Gender. As can be seen in Tables 6.22 and 6.23, scores for male and female examinees were very similar across the SAGES-3: K–3 and SAGES-3: 4–8 subtests and composites. On the SAGES-3: K–3, the values of the mean difference scores ranged from −1.84 to 2.19; on the SAGES-3: 4–8, they ranged from −2.02 to .51. The magnitude of the effect sizes for the difference scores were all trivial and within 1 *SEM* (4–5 points for the subtests and 3–4 points for the composites) of each other on the SAGES-3: K–3 and the SAGES-3: 4–8. These results indicate little or no gender bias on the SAGES-3.

Table 6.22
Comparison of SAGES-3: K–3 Scores for Male and Female Examinees

	Male (*n* = 329)	Female (*n* = 329)	Difference		Effect	Effect	
SAGES-3: K–3 value	*M* (*SD*)	*M* (*SD*)	score	*t*	size *d*	size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	99.74 (16.21)	101.58 (15.33)	−1.84	−1.49 *ns*	−.12	−.06	Trivial
Language Arts/Social Studies	100.81 (14.67)	100.29 (13.20)	.52	.47 *ns*	.04	.02	Trivial
Verbal Reasoning	100.14 (14.47)	99.51 (13.76)	.63	.57 *ns*	.04	.02	Trivial
Mathematics/Science	101.26 (15.44)	99.07 (13.07)	2.19	1.97 *ns*	.15	.08	Trivial
Composite							
Reasoning Ability	99.80 (15.07)	100.48 (14.31)	−.68	−.59 *ns*	−.05	−.02	Trivial
Academic Ability	100.95 (15.36)	99.46 (13.26)	1.49	1.33 *ns*	.10	.05	Trivial
General Ability	100.79 (14.89)	100.29 (13.51)	.50	.46 *ns*	.04	.02	Trivial

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.

[a] Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

Table 6.23
Comparison of SAGES-3: 4–8 Scores for Male and Female Examinees

	Male (*n* = 315)	Female (*n* = 315)	Difference		Effect	Effect	
SAGES-3: 4–8 value	*M* (*SD*)	*M* (*SD*)	score	*t*	size *d*	size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	98.70 (13.61)	99.51 (12.92)	−.81	−.76 *ns*	−.06	−.03	Trivial
Language Arts/Social Studies	97.93 (13.75)	99.95 (13.65)	−2.02	−1.85 *ns*	−.15	−.07	Trivial
Verbal Reasoning	97.83 (13.95)	99.59 (13.79)	−1.76	−1.59 *ns*	−.13	−.06	Trivial
Mathematics/Science	99.08 (14.09)	98.57 (13.41)	.51	.46 *ns*	−.04	−.02	Trivial
Composite							
Reasoning Ability	98.15 (13.72)	99.64 (12.95)	−1.49	−1.40 *ns*	−.11	−.06	Trivial
Academic Ability	98.75 (13.85)	99.56 (13.06)	−.81	−.76 *ns*	−.06	−.03	Trivial
General Ability	98.22 (13.64)	99.48 (12.76)	−1.26	−1.20 *ns*	−.10	−.05	Trivial

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.

[a] Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

Black/African American. As can be seen in Tables 6.24 and 6.25, the values of the SAGES-3: K–3 mean difference scores ranged from −7.41 to −2.83; on the SAGES-3: 4–8, they ranged from −6.91 to −4.80. Although the mean difference scores were approximately 1 or more *SEM*s and favor White examinees,

Table 6.24
Comparison of SAGES-3: K–3 Scores for Black/African American Examinees and a Matched Sample

	Black/African American (*n* = 145)	White (*n* = 145)					
SAGES-3: K–3 value	*M* (*SD*)	*M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	95.50 (15.98)	98.33 (14.01)	−2.83	1.61 *ns*	−.19	−.09	Trivial
Language Arts/Social Studies	93.67 (13.23)	99.99 (12.97)	−6.32	4.11 ***	−.48	−.23	Small
Verbal Reasoning	94.97 (14.06)	98.92 (12.01)	−3.95	2.57 *	−.30	−.15	Small
Mathematics/Science	93.57 (13.62)	100.19 (14.26)	−6.62	4.04 ***	−.47	−.23	Small
Composite							
Reasoning Ability	94.52 (14.91)	98.34 (12.93)	−3.82	2.33 *	−.27	−.14	Small
Academic Ability	92.47 (13.33)	99.88 (13.78)	−7.41	4.66 ***	−.55	−.26	Small
General Ability	92.93 (13.59)	99.12 (13.08)	−6.19	3.95 ***	−.46	−.23	Small

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

**p* < .05. ****p* < .001.

Table 6.25
Comparison of SAGES-3: 4–8 Scores for Black/African American Examinees and a Matched Sample

	Black/African American (*n* = 123)	White (*n* = 123)					
SAGES-3: 4–8 value	*M* (*SD*)	*M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	93.92 (15.01)	100.49 (13.92)	−6.57	3.56 ***	−.45	−.22	Small
Language Arts/Social Studies	94.02 (15.22)	99.67 (12.28)	−5.65	3.21 **	−.41	−.20	Small
Verbal Reasoning	93.49 (13.19)	98.29 (13.96)	−4.80	2.78 **	−.35	−.17	Small
Mathematics/Science	93.68 (13.56)	99.05 (13.10)	−5.37	3.16 **	−.40	−.20	Small
Composite							
Reasoning Ability	92.83 (13.90)	99.43 (13.01)	−6.60	3.85 ***	−.49	−.24	Small
Academic Ability	93.28 (15.18)	99.76 (12.20)	−6.48	3.69 ***	−.47	−.23	Small
General Ability	92.52 (13.66)	99.43 (12.09)	−6.91	4.20 ***	−.54	−.26	Small

Note. Samples were matched according to age, race, gender, and parent education.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

***p* < .01. ****p* < .001.

the differences between the composite indexes are much smaller than the range of 7.5 to 15 points often reported in the literature (see Rushton & Jensen, 2005; Suzuki & Valencia, 1997). Moreover, the magnitudes for the differences ranged from trivial to small. One may conclude, therefore, that the SAGES-3 scores possess little bias against Black/African American examinees.

Asian/Pacific Islander. As can be seen in Tables 6.26 and 6.27, the subtest and composite means for Asian/Pacific Islander examinees on the SAGES-3: K–3 and the SAGES-3: 4–8 were mostly higher than those for the comparison sample. On the SAGES-3: K–3, the values of the mean difference scores ranged from 2.18 to 4.78; on the SAGES-3: 4–8, they ranged from −1.37 to 4.64. Although a few of the mean difference scores were approximately 1 or more *SEM*s, the magnitudes for the differences ranged from trivial to small. One may conclude, therefore, that the SAGES-3 scores possess little bias against Asian/Pacific Islander examinees.

Hispanic. As can be seen in Tables 6.28 and 6.29, the values of the SAGES-3: K–3 mean difference scores ranged from −5.24 to −.96; on the SAGES-3: 4–8, they ranged from –4.63 to –2.39. Although some of the mean difference scores favored White examinees and were approximately 1 or more *SEM*s, the magnitudes for the differences ranged from trivial to small. One may conclude, therefore, that the SAGES-3 scores possess little bias against Hispanic examinees.

Table 6.26
Comparison of SAGES-3: K–3 Scores for Asian/Pacific Islander Examinees and a Matched Sample

	Asian/Pacific Islander (n = 23)	White (n = 23)					
SAGES-3: K–3 value	*M* (*SD*)	*M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	108.83 (14.41)	105.91 (11.64)	2.92	–.75 *ns*	.22	.11	Small
Language Arts/Social Studies	108.91 (18.17)	104.13 (13.47)	4.78	–1.01 *ns*	.31	.15	Small
Verbal Reasoning	108.09 (12.77)	105.91 (12.16)	2.18	–.59 *ns*	.19	.09	Trivial
Mathematics/Science	108.87 (19.29)	105.61 (13.96)	3.26	–.66 *ns*	.17	.08	Trivial
Composite							
Reasoning Ability	109.35 (12.81)	106.39 (10.88)	2.96	–.84 *ns*	.25	.12	Small
Academic Ability	109.78 (20.36)	105.39 (13.65)	4.39	–.86 *ns*	.25	.13	Small
General Ability	110.83 (16.89)	106.70 (12.69)	4.13	–.94 *ns*	.28	.14	Small

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.

[a] Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

Table 6.27
Comparison of SAGES-3: 4–8 Scores for Asian/Pacific Islander Examinees and a Matched Sample

SAGES-3: 4–8 value	Asian/Pacific Islander ($n = 30$) M (SD)	White ($n = 30$) M (SD)	Difference score	t	Effect size d	Effect size r	Magnitude[a]
Subtest							
Nonverbal Reasoning	103.20 (10.99)	104.57 (14.84)	−1.37	.41 *ns*	−.10	−.05	Trivial
Language Arts/Social Studies	104.47 (13.89)	99.83 (12.93)	4.64	−1.34 *ns*	.35	.17	Small
Verbal Reasoning	103.77 (14.17)	102.13 (14.20)	1.64	−.45 *ns*	.12	.06	Trivial
Mathematics/Science	104.57 (11.40)	102.20 (11.93)	2.37	−.79 *ns*	.20	.10	Small
Composite							
Reasoning Ability	104.27 (12.23)	104.07 (14.48)	.20	−.06 *ns*	.01	.01	Trivial
Academic Ability	105.27 (11.84)	101.47 (12.07)	3.80	−1.23 *ns*	.32	.16	Small
General Ability	105.20 (11.65)	103.00 (12.77)	2.20	−.70 *ns*	.18	.09	Trivial

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.
[a] Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

Criterion-Prediction Validity

Anastasi and Urbina (1997) described criterion-related validity as "the effectiveness of a test in predicting an individual's performance in specific activities" (p. 118). They stated that performance on a test should be checked against a criterion that can be either a direct or an indirect measure of what the test is designed to predict. So to be valid, a test like the SAGES-3, which is built to measure reasoning and academic abilities, should (a) correlate strongly with established tests that measure the same abilities, (b) yield the same or similar means and standard deviations as those of the criterion tests, and (c) accurately differentiate between persons who are known to have better than average reasoning or academic abilities and those who are known to have poor or severely delayed reasoning and academic abilities.

Correlation With Criterion Measures

In this investigation, we correlated SAGES-3 scores with scores from the criterion tests. Most of the participants used in the study were members of the SAGES-3 normative sample, and the criterion test data came from examiners' case files or school or clinic records. Only current data from these sources were used. Additional participants were tested by PRO-ED professional staff or by other professionals under the direction of this staff; in these cases, the SAGES-3 and criterion tests were administered concurrently. Because the six samples used in

Table 6.28
Comparison of SAGES-3: K–3 Scores for Hispanic Examinees and a Matched Sample

SAGES-3: K–3 value	Hispanic (n = 192) *M* (*SD*)	Non-Hispanic (n = 192) *M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	98.73 (14.93)	99.69 (14.41)	–.96	–.64 *ns*	–.07	–.03	Trivial
Language Arts/Social Studies	96.45 (13.27)	98.74 (13.97)	–2.29	–1.65 *ns*	–.17	–.08	Trivial
Verbal Reasoning	95.61 (14.06)	99.21 (12.76)	–3.60	–2.63 **	–.27	–.13	Small
Mathematics/Science	94.55 (12.08)	99.79 (12.84)	–5.24	–4.12 ***	–.42	–.21	Small
Composite							
Reasoning Ability	96.65 (14.51)	99.22 (13.47)	–2.57	–1.80 *ns*	–.18	–.09	Trivial
Academic Ability	94.63 (13.02)	98.98 (13.16)	–4.35	–3.26 **	–.33	–.16	Small
General Ability	95.38 (13.31)	99.15 (12.62)	–3.77	–2.85 **	–.29	–.14	Small

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

**p < .01. ***p < .001.

Table 6.29
Comparison of SAGES-3: 4–8 Scores for Hispanic Examinees and a Matched Sample

SAGES-3: 4–8 value	Hispanic (n = 191) *M* (*SD*)	Non-Hispanic (n = 191) *M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	95.98 (13.69)	99.29 (11.54)	–3.31	–2.56 **	–.26	–.13	Small
Language Arts/Social Studies	95.71 (13.95)	99.02 (13.26)	–3.31	–2.38 **	–.24	–.12	Small
Verbal Reasoning	95.71 (13.23)	100.34 (13.14)	–4.63	–3.43 **	–.35	–.17	Small
Mathematics/Science	96.27 (14.28)	98.66 (12.93)	–2.39	–1.72 *ns*	–.18	–.09	Trivial
Composite							
Reasoning Ability	95.33 (13.29)	99.91 (11.85)	–4.58	–3.56 ***	–.36	–.18	Small
Academic Ability	95.87 (13.96)	99.19 (12.55)	–3.32	–2.45 **	–.25	–.12	Small
General Ability	95.14 (13.47)	99.38 (11.67)	–4.24	–3.29 **	–.34	–.17	Small

Note. Samples were matched according to age, race, gender, and parent education. *ns* = not significant.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

**p < .01. ***p < .001.

these studies were demographically different, their specific characteristics are described in Table 6.30 for the SAGES-3: K–3 and Table 6.31 for the SAGES-3: 4–8.

The criterion measures used in our studies were taken from six cognitive or academic achievement batteries. The internal consistency of both the SAGES-3 and criterion test scores reaches or exceeds .85. The criterion measures are as follows:

- The *Cognitive Abilities Test, Form 6* (CogAT; Lohman & Hagen, 2001) is an assessment designed to measure the verbal, quantitative, and

Table 6.30
Demographic Characteristics of the Samples Used in the SAGES-3: K–3 Criterion-Prediction Validity Studies

	Criterion				
Sample characteristic	CogAT	WPPSI-IV	YCAT-2	Gifted and talented sample	IQ sample
Total number of participants	73	43	30	218	169
Age range (in years)	6–9	5–7	5–7	5–9	5–9
Location	CO, MO	MO, NY, TX, WV	AZ, MN, NY, TX, WV	AR, AZ, IL, KS, MD, MO, TX, WV	AR, AZ, CA, CO, IL, MN, MO, NY, TX, WV
Gender					
Male	38	21	15	107	79
Female	35	22	15	111	90
Race					
White	57	38	26	190	139
Black/African American	14	1	3	4	23
Asian/Pacific Islander	0	0	0	13	1
Two or more races	2	4	1	11	6
Hispanic status					
Yes	17	1	4	14	32
No	56	42	26	204	137
Exceptionality status					
None	43	13	29	0	98
Gifted and talented	30	29	1	218	67
Physical or health impairment	0	0	0	0	1
Learning disability	0	0	0	0	1
Language impairment	0	0	0	0	1
Behavioral disorder	0	1	0	0	1
Other	0	0	0	0	1

Note. CogAT = *Cognitive Abilities Test, Form 6* (Lohman & Hagen, 2001); WPPSI-IV = *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition* (Wechsler, 2012); YCAT-2 = *Young Children's Achievement Test–Second Edition* (Hresko, Peak, Herron, & Hicks, 2018).

nonverbal reasoning abilities in students in kindergarten through 12th grade.

- The *Detroit Tests of Learning Abilities–Fifth Edition* (DTLA-5; Hammill, McGhee, & Ehrler, 2018) is an assessment that measures a wide variety of cognitive abilities in individuals ages 6 through 17 years.
- The *Universal Nonverbal Intelligence Test–Group Abilities Test* (UNIT-GAT; Bracken & McCallum, 2019) is a nonverbal assessment of cognitive functioning in individuals ages 5 through 21 years.

Table 6.31
Demographic Characteristics of the Samples Used in the SAGES-3: 4–8 Criterion-Prediction Validity Studies

	Criterion				
Sample characteristic	DTLA-5	WJ IV ACH	UNIT-GAT	Gifted and talented sample	IQ sample
Total number of participants	30	40	54	678	121
Age range (in years)	9–14	9–14	9–14	9–14	9–14
Location	KS, NE, NY, TX	MN, NY	AZ, MN, NE, NY, TX	U.S.	AZ, KS, MN, MO, NE, NY, TX
Gender					
Male	11	20	23	328	57
Female	19	20	31	350	64
Race					
White	29	32	51	597	115
Black/African American	0	4	2	17	1
Asian/Pacific Islander	1	3	1	30	2
American Indian/Alaska Native	0	0	0	3	0
Two or more races	0	1	0	31	3
Hispanic status					
Yes	7	4	7	86	11
No	23	36	47	592	110
Exceptionality status					
None	26	35	47	0	70
Gifted and talented	1	3	5	678	46
Visual impairment	0	0	0	1	0
Learning disability	1	0	1	0	2
Attention-deficit/hyperactivity disorder	2	2	2	1	3

Note. DTLA-5 = *Detroit Tests of Learning Abilities–Fifth Edition* (Hammill, McGhee, & Ehrler, 2018); WJ IV ACH = *Woodcock–Johnson IV Tests of Achievement* (Schrank, Mather, & McGrew, 2014); UNIT-GAT = *Universal Nonverbal Intelligence Test–Group Abilities Test* (Bracken & McCallum, 2019).

- The *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition* (WPPSI-IV; Wechsler, 2012) is a widely used measure of global intellectual functioning in children ages 2 through 7 years.
- The *Woodcock–Johnson IV Tests of Achievement*, Brief Achievement (WJ IV ACH; Schrank, Mather, & McGrew, 2014) is an abbreviated version of the WJ IV ACH, a multisubtest measure of academic achievement in children and adults ages 2 through 19 years. Brief Achievement comprises three subtests (Letter-Word Identification, Applied Problems, and Spelling).
- The *Young Children's Achievement Test–Second Edition* (YCAT-2; Hresko, Peak, Herron, & Hicks, 2018) is an assessment designed to measure the achievement abilities of preschool, kindergarten, and first-grade children with respect to those skills that ensure success in school.

The results of this study are presented in Tables 6.32 and 6.33. The coefficients for the criterion tests are organized in a way corresponding to the constructs of the SAGES-3: Nonverbal Reasoning, Language Arts/Social Studies, Verbal Reasoning, Mathematics/Science, General Cognitive Ability, and General Achievement. In this study, we are asking a theoretical question: Does the SAGES-3 measure reasoning and academic abilities? Because the question is theoretical, we attenuated the coefficients for any lack of reliability in the criterion test and corrected for any range effects that might artificially reduce or inflate the size of the coefficients. Both the corrected and uncorrected coefficients are reported in the tables (uncorrected coefficients appear in parentheses).

In interpreting the magnitude of coefficients in this study, we are guided by Hopkins (2002). As previously stated, he suggested that coefficients between .00 and .09 are very small or trivial, coefficients between .10 and .29 are small, coefficients between .30 and .49 are moderate, coefficients between .50 and .69 are large, coefficients between .70 and .89 are very large, and coefficients between .90 and 1.00 are nearly perfect.

Interpretation of these tables is fairly straightforward. Of the 49 SAGES-3: K–3 coefficients, 84% range from large to very large in magnitude. Similarly, of the 49 SAGES-3: 4–8 coefficients, 84% range from large to nearly perfect in magnitude. The majority of coefficients are large enough to provide convincing evidence that the SAGES-3 and the criterion tests are measuring the same constructs.

Interestingly, the correlation between the SAGES-3: K–3 Nonverbal Reasoning subtest and the CogAT Nonverbal Battery was lower than expected (.62 corrected), as were the correlations between the SAGES-3: 4–8 Nonverbal Reasoning subtest and the DTLA-5 Nonverbal Problem Solving (.68 corrected) and the UNIT-GAT Full Scale Index (.57 corrected). On closer examination of the CogAT Nonverbal Battery items, we noted that they were very similar to the DTLA-5 Nonverbal Problem Solving items but dissimilar to the SAGES-3 Nonverbal Reasoning items. In fact, the SAGES-3 Nonverbal Reasoning items were more similar to those found on the CogAT Verbal Battery (Picture/Verbal Analogies). The CogAT Picture/Verbal Analogies items for younger students display two images that go together and a third image. The examinee must determine which picture in the answer choices goes with the third image. The SAGES-3 Nonverbal Reasoning items are formatted the same way, and this could help explain why the

Table 6.32
Corrected (and Uncorrected) Correlation Coefficients Showing the Relationship Between SAGES-3: K–3 and Criterion Measures (Decimals Omitted)

		SAGES-3: K–3 value									
		Subtest				Composite					
Content/criterion test	*N*	Nonverbal Reasoning	Language Arts/Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Magnitude[a]	Academic Ability	Magnitude[a]	General Ability	Magnitude[a]
Nonverbal Reasoning											
CogAT Nonverbal Battery	73	62 (63)	75 (76)	85 (76)	86 (81)	80 (76)	Very large	84 (85)	Very large	87 (87)	Very large
Language Arts/Social Studies											
YCAT-2 Reading	30	20 (19)	60 (60)	45 (50)	73 (69)	45 (44)	Moderate	65 (68)	Large	64 (65)	Large
Verbal Reasoning											
CogAT Verbal Battery	73	63 (66)	72 (76)	81 (74)	85 (82)	79 (76)	Very large	83 (85)	Very large	85 (86)	Very large
Mathematics/Science											
YCAT-2 Mathematics	87	21 (18)	60 (57)	45 (47)	56 (52)	47 (43)	Moderate	58 (58)	Large	58 (58)	Large
General Cognitive Ability											
CogAT Composite	73	68 (69)	79 (81)	86 (77)	92 (86)	84 (80)	Very large	91 (90)	Very large	94 (91)	Very large
WPPSI-IV Full Scale IQ	43	64 (71)	70 (70)	61 (57)	61 (63)	72 (73)	Very large	73 (74)	Very large	81 (80)	Very large
General Achievement											
YCAT-2 Early Achievement Index	30	40 (39)	63 (65)	43 (50)	71 (70)	56 (56)	Large	72 (67)	Very large	67 (71)	Large

Note. Coefficients inside the parentheses are uncorrected; coefficients outside the parentheses are corrected for range effects and reliability of the criterion measures. CogAT = *Cognitive Abilities Test*, Form 6 (Lohman & Hagen, 2001); YCAT-2 = *Young Children's Achievement Test–Second Edition* (Hresko, Peak, Herron, & Hicks, 2018); WPPSI-IV = *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition* (Wechsler, 2012).

[a] Magnitude of the corrected coefficient for the composite indexes based on Hopkins's (2002) criteria.

Table 6.33
Corrected (and Uncorrected) Correlation Coefficients Showing the Relationship Between SAGES-3: 4–8 and Criterion Measures (Decimals Omitted)

		SAGES-3: 4–8 value									
		Subtest				Composite					
Content/criterion test	*N*	Nonverbal Reasoning	Language Arts/Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Magnitude[a]	Academic Ability	Magnitude[a]	General Ability	Magnitude[a]
Nonverbal Reasoning											
DTLA-5 Nonverbal Problem Solving	30	68 (56)	66 (57)	51 (42)	51 (45)	65 (55)	Large	61 (56)	Large	68 (61)	Large
UNIT-GAT Full Scale Index	54	57 (49)	55 (51)	39 (39)	46 (45)	55 (51)	Large	55 (53)	Large	66 (62)	Large
Language Arts/Social Studies											
WJ IV ACH Letter-Word Identification	40	38 (23)	85 (61)	75 (54)	70 (46)	70 (48)	Very large	84 (59)	Very large	78 (58)	Very large
Verbal Reasoning											
DTLA-5 Verbal Comprehension	30	60 (41)	79 (63)	77 (57)	70 (55)	73 (55)	Very large	75 (62)	Very large	80 (65)	Very large
Mathematics/Science											
WJ IV ACH Applied Problems	40	47 (40)	58 (47)	21 (19)	70 (60)	43 (36)	Moderate	71 (59)	Very large	58 (48)	Large
General Cognitive Ability											
DLTA-5 General Cognitive Ability	30	64 (43)	85 (69)	83 (63)	73 (56)	80 (61)	Very large	80 (66)	Very large	85 (70)	Very large
General Achievement											
WJ IV ACH Brief Achievement	40	43 (31)	88 (71)	61 (47)	80 (64)	63 (48)	Large	91 (75)	Nearly perfect	85 (69)	Very large

Note. Coefficients inside the parentheses are uncorrected; coefficients outside the parentheses are corrected for range effects and reliability of the criterion measures. DTLA-5 = *Detroit Tests of Learning Abilities–Fifth Edition* (Hammill, McGhee, & Ehrler, 2018); UNIT-GAT = *Universal Nonverbal Intelligence Test–Group Abilities Test* (Bracken & McCallum, 2019); WJ IV ACH = *Woodcock–Johnson IV Tests of Achievement* (Schrank, Mather, & McGrew, 2014).

[a] Magnitude of the corrected coefficient for the composite indexes based on Hopkins's (2002) criteria.

correlation between the SAGES-3: K–3 and CogAT Verbal Battery is higher than the correlations between the SAGES-3 and the nonverbal criterion measures. The UNIT-GAT has analogic reasoning items like the SAGES-3 Nonverbal Reasoning items, but it also has quantitative reasoning items, which require students to solve math problems depicted by math symbols/numbers or an array of white and black domino-like objects.

Comparison of the SAGES-3 and Criterion Test Means and Standard Deviations

When two tests are highly correlated, they are likely to be measuring the same or a similar ability. They may not, however, yield the same test results. For example, one test may score consistently higher than another test even though they correlate highly with each other. The validity of both tests is supported when the two tests produce similar means and correlate highly with each other.

The composite standard score means, standard deviations, and comparative information for the SAGES-3, CogAT, DTLA-5, WJ IV ACH, WPPSI-IV, UNIT-GAT, and YCAT-2 are presented in Tables 6.34 and 6.35. The probabilities of giftedness used to describe the means are listed in Table 3.1. The differences between the SAGES-3 means and the corresponding criterion test score means were analyzed using the dependent samples *t* test (Guilford & Fruchter, 1978), effect size *d* for correlated designs (Formula #3; Dunlap, Cortina, Vaslow, & Burke, 1996), and effect size *r* from *d* (Borenstein, 2009) estimates.

As expected, the mean standard score differences between SAGES-3 scores and those of the criterion tests are all small or trivial in magnitude. The findings reported in Tables 6.34 and 6.35 support the idea that, for all practical purposes, the standard scores that result from giving the SAGES-3 will likely be similar to those obtained from giving other reasoning or academic achievement tests.

Diagnostic Accuracy Analyses

The studies just reported show that the scores of the SAGES-3 are highly related to the scores of current well-established tests of cognitive and academic achievement. This provides a type of apostolic, theoretical evidence for the SAGES-3's criterion-predictive validity (i.e., if the criterion tests are indeed valid, then the SAGES-3 is valid as well). The studies about to be discussed provide practical evidence for the SAGES-3 criterion-predictive validity using statistical procedures referred to in the literature as *diagnostic accuracy analyses*. These analyses demonstrate the precision with which the SAGES-3 scores can differentiate students with a high IQ (i.e., a score of 130 or higher on a cognitive abilities test) from students who do not have a high IQ (i.e., a score lower than 130 on a cognitive abilities test), and they can do so without excessive false positives (i.e., the misclassification of typical students as exceptional, which leads directly to overreferrals).

Researchers such as Swets (1996); Betz, Eickhoff, and Sullivan (2013); Dollaghan (2004); Gray, Plante, Vance, and Henrichsen (1999); and Pepe (2003) have long suggested that diagnostic accuracy is the preferred method of assessing the usefulness of a diagnostic measure. Dollaghan (2004) went so far as to proclaim it "the most important criterion for evaluating a diagnostic measure" (p. 395).

Table 6.34
Index Means (and Standard Deviations) and Related Statistics and *t* Values for the SAGES-3: K–3 and Criterion Tests

SAGES-3: K–3/criterion test	*N*	*M* (*SD*)	Probability of giftedness	*t*[a]	Effect size *r*[b]	Effect size *d*	Magnitude[c]
Nonverbal Reasoning							
SAGES-3: K–3 Nonverbal Reasoning	73	110 (16)	Possibly	1.87 *ns*	.09	.19	Trivial
CogAT Nonverbal Battery		107 (16)	Unlikely				
Language Arts/Social Studies							
SAGES-3: K–3 Language Arts/Social Studies	30	99 (17)	Unlikely	–.95 *ns*	–.08	–.16	Trivial
YCAT-2 Reading		101 (15)	Unlikely				
Verbal Reasoning							
SAGES-3: K–3 Verbal Reasoning	73	107 (13)	Unlikely	–.95 *ns*	–.04	–.07	Trivial
CogAT Verbal Battery		108 (18)	Unlikely				
Mathematics/Science							
SAGES-3: K–3 Mathematics/Science	30	103 (15)	Average	–.25 *ns*	–.02	–.04	Trivial
YCAT-2 Math		103 (16)	Average				
General Cognitive Ability							
SAGES-3: K–3 Reasoning Ability Index	73	109 (16)	Unlikely	.77 *ns*	.04	.07	Trivial
CogAT Composite		108 (14)	Unlikely				
SAGES-3: K–3 Reasoning Ability Index	43	118 (15)	Possibly	–2.63 *	–.15	–.30	Small
WPPSI-IV Full Scale IQ		123 (16)	Likely				
General Achievement							
SAGES-3: K–3 Academic Ability Index	30	102 (16)	Unlikely	.52 *ns*	.04	.07	Trivial
YCAT-2 Early Achievement Index		101 (17)	Unlikely				

Note. ns = not significant. CogAT = *Cognitive Abilities Test*, Form 6 (Lohman & Hagen, 2001); YCAT-2 = *Young Children's Achievement Test–Second Edition* (Hresko, Peak, Herron, & Hicks, 2018); WPPSI-IV = *Wechsler Preschool and Primary Scale of Intelligence–Fourth Edition* (Wechsler, 2012).

[a] Values of *t* were computed by the dependent samples method (Guilford & Fruchter, 1978). [b] Effect size was calculated using Dunlap, Cortina, Vaslow, and Burke's (1996) Formula #3, which corrects for inflated effect size due to correlated design *t* tests. [c] Values of magnitude of the effect size correlation between the SAGES-3: K–3 score and the criterion tests according to Hopkins's (2002) criteria.

$^*p < .05$.

Methods for establishing diagnostic accuracy involve the computation of a test's sensitivity and specificity indexes and receiver operating characteristic/area under the curve (ROC/AUC). In the current context, the *sensitivity index* reflects the SAGES-3's ability to correctly identify students who are likely to be gifted. The *specificity index* refers to the ability of a test to correctly identify examinees who are not likely to be gifted. ROC/AUC "is a measure of the overall performance of a diagnostic test and is interpreted as the average value of sensitivity for all possible values of specificity" (Park, Goo, & Jo, 2004, p. 13).

Table 6.35
Index Means (and Standard Deviations) and Related Statistics and *t* Values for the SAGES-3: 4–8 and Criterion Tests

SAGES-3: 4–8/criterion test	*N*	*M* (*SD*)	Probability of giftedness	t^a	Effect size r^b	Effect size *d*	Magnitude[c]
Nonverbal Reasoning							
SAGES-3: 4–8 Nonverbal Reasoning	30	100 (14)	Unlikely	–3.61 **	–.26	–.53	Small
DTLA-5 Nonverbal Problem Solving		108 (12)	Unlikely				
SAGES-3: 4–8 Nonverbal Reasoning	54	100 (14)	Unlikely	–1.94 *ns*	–.14	–.28	Small
UNIT-GAT Full Scale Index		104 (15)	Unlikely				
Language Arts/Social Studies							
SAGES-3: 4–8 Language Arts/Social Studies	40	104 (13)	Unlikely	.94 *ns*	.04	.08	Trivial
WJ IV ACH Letter-Word Identification		102 (10)	Unlikely				
Verbal Reasoning							
SAGES-3: 4–8 Verbal Reasoning	30	101 (13)	Unlikely	–1.00 *ns*	–.06	–.12	Trivial
DTLA-5 Verbal Comprehension		103 (12)	Unlikely				
Mathematics/Science							
SAGES-3: 4–8 Mathematics/Science	40	106 (14)	Unlikely	–3.67 ***	–.22	–.45	Small
WJ IV ACH Applied Problems		113 (14)	Possibly				
General Cognitive Ability							
SAGES-3: 4–8 Reasoning Ability Index	30	101 (14)	Unlikely	–1.78 *ns*	–.1	–.21	Small
DTLA-5 General Cognitive Ability		104 (11)	Unlikely				
General Achievement							
SAGES-3: 4–8 Academic Ability Index	40	106 (13)	Average	–1.55 *ns*	–.05	–.10	Trivial
WJ IV ACH Brief Achievement		108 (12)	Average				

Note. ns = not significant. DTLA-5 = *Detroit Tests of Learning Abilities–Fifth Edition* (Hammill, McGhee, & Ehrler, 2018); UNIT-GAT = *Universal Nonverbal Intelligence Test–Group Abilities Test* (Bracken & McCallum, 2019); WJ IV ACH = *Woodcock–Johnson IV Tests of Achievement* (Schrank, Mather, & McGrew, 2014).

[a] Values of *t* were computed by the dependent samples method (Guilford & Fruchter, 1978). [b] Effect size was calculated using Dunlap, Cortina, Vaslow, and Burke's (1996) Formula #3, which corrects for inflated effect size due to correlated design *t* tests. [c] Values of magnitude of the effect size correlation between the SAGES-3: 4–8 score and the criterion tests according to Hopkins's (2002) criteria.

$^{**}p < .01$. $^{***}p < .001$.

Sensitivity and specificity indexes are reported as proportions (i.e., percentages). The size of the proportions necessary to be considered acceptable varies depending on the purpose of the analysis (e.g., when screening for cancer, a relatively high number of false positives is tolerable in order to ensure that the number of true positives identified is high). ROC/AUC, however, is a more comprehensive index of the overall accuracy of a measure and ranges from 0 (representing *no predictive ability*) to 1 (representing *perfect predictive ability*). ROC/AUC values closer to 1 are always preferred. Of the multiple measures of diagnostic accuracy, ROC/AUC has become the preferred statistic for evaluating the overall diagnostic accuracy of a measure (Dollaghan, 2004; Gray et al., 1999; Pepe, 2003; Swets, 1996), whereas specificity and sensitivity are more useful for evaluating the diagnostic accuracy of a measure at a particular cut score.

Educational researchers vary in their opinions about the minimum acceptable levels for sensitivity, specificity, and ROC/AUC. Wood, Flowers, Meyer, and Hill (2002) recommended that sensitivity and specificity indexes should be at least .70. Jansky (1978), Gredler (2000), and Kingslake (1983) preferred .75 for both indexes. Carran and Scott (1992) and Plante and Vance (1994) recommended a more rigorous standard of .80 or higher. Jenkins and others (Jenkins, 2003; Jenkins, Hudson, & Johnson, 2007; Johnson, Jenkins, Petscher, & Catts, 2009) recommended that sensitivities be high—perhaps as high as .90—and that specificity levels be relatively high as well. Similarly, Compton, Fuchs, Fuchs, and Bryant (2006) suggested that ROC/AUCs of .90 and above are excellent, .80 to .89 are good, .70 to .79 are fair, and .69 or below are poor. Swets (1988) suggested that ROC/AUCs of .96 and above are excellent, .85 to .95 are very good, .75 to .84 are reasonable, and less than .75 are relatively poor.

Because the SAGES-3 is a measure of reasoning and academic abilities, a series of analyses was conducted to examine its ability to accurately differentiate students who had been tested with a measure of cognitive ability (e.g., DTLA-5, CogAT, WPPSI-IV, UNIT-GAT, *Naglieri Nonverbal Ability Test–Second Edition* [NNAT2; Naglieri, 2008], *Wechsler Intelligence Scale for Children–Fourth Edition* [WISC-IV; Weschler, 2003]) and obtained a score of 130 or higher from students who scored lower than 130 on one of those measures. For the SAGES-3: K–3, the analyses included 34 high-IQ students and 135 lower IQ students. For SAGES-3: 4–8, the analyses included 20 high-IQ students and 101 lower IQ students.

Researchers (e.g., Dolloghan, 2004; Gray et al., 1999; Merrell & Plante, 1997; Plante & Vance, 1994, 1995; Rescorla, 1989; Rescorla & Alley, 2001; Rice & Wexler, 2001; Spaulding, Plante, & Farinella, 2006) have advocated for empirically based cutoff scores that maximize sensitivity and specificity (i.e., equalizing the rates of false positives and false negatives). The diagnostic accuracy of the SAGES-3 was examined at seven different cutoff scores—composite indexes of 108 (.5 *SD*), 110 (.7 *SD*), 115 (1 *SD*), 120 (1.33 *SD*), 122 (1.5 *SD*), 126 (1.75 *SD*), and 130 (2 *SD*).

Using the two dichotomous groups that are created based on the selected cutoff scores, forty-two 2 × 2 frequency matrices were created. An example of a matrix used to examine the diagnostic accuracy of the SAGES-3: K–3 when using a Reasoning Ability index cutoff score of 122 to predict high IQ is presented in Table 6.36. In this table, the number of students correctly identified by the SAGES-3: K–3 Reasoning Ability index is represented by cells a and d. Cell a represents true positives, and cell d represents true negatives. The number of students who were not correctly identified is represented by cells b and c. Cell b represents false positives (overreferrals). Cell c represents false negatives (underreferrals). The sensitivity index is calculated by dividing the number of true positives (cell a) by the sum of true positives and false negatives (cell a + cell c). The specificity index is calculated by dividing the number of true negatives (cell d) by the sum of true negatives and false positives (cell d + cell b). The bolded quotients in the table note correspond to the values found in Table 6.37.

Tables 6.37 and 6.38 report the results of the diagnostic accuracy analyses for the SAGES-3: K–3 and the SAGES-3: 4–8 composite indexes (i.e., Reasoning Ability, Academic Ability, and General Ability) in differentiating students who

Table 6.36
Diagnostic Accuracy Matrix Demonstrating SAGES-3: K–3 Reasoning Ability Index's Ability to Predict High IQ

SAGES-3: K–3 Reasoning Ability index	IQ Greater than 120	 Less than 120	 Total
Greater than 121	24[a]	6[b]	30
Less than 122	10[c]	129[d]	139
Total	34	135	169

Note. Sensitivity index = 24 / 34 = **.71**; specificity index = 129 / 135 = **.96**.
[a]True positives. [b]False positives. [c]False negatives. [d]True negatives.

have a high IQ from students who do not have a high IQ. Interestingly, the cut scores of 120 and 122 met acceptable minimal criteria for both the indexes on the SAGES-3: K–3 and SAGES-3: 4–8. Although these cut scores are lower than the commonly used criterion of an index of 130, they are consistent with tests of cognitive ability that report a mean below 130 for their gifted and talented samples (see DTLA-5, UNIT2, SB5, and WJ IV COG).

In summary, the SAGES-3 met and exceeded the minimum standards for diagnostic accuracy recommended by the authorities mentioned earlier in this section when used to differentiate students who have high IQs from students who do not have high IQs. Nonetheless, sensitivity and specificity comparisons of SAGES-3 diagnostic accuracy must be interpreted in light of several methodological factors. First, the accuracy of the IQ scores must be verified, as some of them were obtained from previous assessments (e.g., examiner's case files, school records). Moreover, students may have been misclassified by their scores, meaning there could have been students in the lower IQ group who should have obtained scores of 130 or higher but did not. The rate of false positives would appear to be large if unidentified high-IQ students were included in the sample but were inaccurately identified as having an IQ lower than 130. Based on these considerations, the results of this diagnostic accuracy study of the SAGES-3 should be considered an underestimate of its true abilities to discriminate between students who have high IQs from students who do not have high IQs.

Construct-Identification Validity

Construct-identification validity, the final type of validity to be examined, relates to the degree to which underlying traits of a test can be identified and the extent to which these traits reflect the theoretical model on which the test is based. For the SAGES-3, we used a three-step procedure to demonstrate this kind of validity. First, we identified several constructs presumed to account for test performance. Second, we generated hypotheses based on the identified constructs. Third, we verified the hypotheses by logical or empirical methods. The following

Table 6.37
Diagnostic Accuracy of the SAGES-3: K–3 in Predicting High IQ (*N* = 169)

SAGES-3: K–3 value	Cutoff index score	*SD*	Percentile rank	Sensitivity index	Specificity index	ROC/ AUC	True positives	False positives	True negatives	False negatives
Reasoning Ability index	108	.50	70	.91	.58	.92	31	57	78	3
	110	.70	75	.91	.63		31	50	85	3
	115	1.00	84	.88	.84		30	22	113	4
	120	1.33	91	.79	.95		27	7	128	7
	122	1.50	93	.71	.96		24	6	129	10
	126	1.75	96	.59	.98		20	3	132	14
	130	2.00	98	.44	.99		15	2	133	19
Academic Ability index	108	.50	70	.94	.64	.92	32	49	86	2
	110	.70	75	.94	.66		32	46	89	2
	115	1.00	84	.88	.77		30	31	104	4
	120	1.33	91	.76	.87		26	17	118	8
	122	1.50	93	.71	.91		24	12	123	10
	126	1.75	96	.59	.96		20	5	130	14
	130	2.00	98	.38	.99		13	2	133	21
General Ability index	108	.50	70	.97	.59	.94	33	55	80	1
	110	.70	75	.94	.61		32	52	83	2
	115	1.00	84	.94	.78		32	30	105	2
	120	1.33	91	.85	.89		29	15	120	5
	122	1.50	93	.85	.89		29	15	120	5
	126	1.75	96	.74	.96		25	6	129	9
	130	2.00	98	.59	.99		20	2	133	14

Note. SD = standard deviation of the normal curve; ROC/AUC = receiver operating characteristic/area under the curve.

basic constructs thought to underlie the SAGES-3 are discussed in the remainder of this chapter:

1. Because cognitive ability (i.e., reasoning and academic abilities) is known to be developmental in nature, one might expect that the raw scores of the SAGES-3 subtests would be strongly related to age.
2. Because the SAGES-3 subtests and composites measure aspects of cognitive ability, the test results should differentiate between groups of students known to possess above- or below-average cognitive ability.
3. Because cognitive ability is thought to be related to spoken language, SAGES-3 results should correlate strongly with measures of spoken language.

Table 6.38
Diagnostic Accuracy of the SAGES-3: 4–8 in Predicting High IQ (*N* = 121)

SAGES-3: 4–8 value	Cutoff index score	*SD*	Percentile rank	Sensitivity index	Specificity index	ROC/AUC	True positives	False positives	True negatives	False negatives
Reasoning Ability index	108	.50	70	.95	.61	.90	19	39	62	1
	110	.70	75	.95	.68		19	32	69	1
	115	1.00	84	.85	.80		17	20	81	3
	120	1.33	91	.70	.91		14	9	92	6
	122	1.50	93	.70	.93		14	7	94	6
	126	1.75	96	.50	.96		10	4	97	10
	130	2.00	98	.20	.98		4	2	99	16
Academic Ability index	108	.50	70	.95	.56	.90	19	44	57	1
	110	.70	75	.90	.63		18	37	64	2
	115	1.00	84	.90	.80		18	20	81	2
	120	1.33	91	.70	.94		14	6	95	6
	122	1.50	93	.60	.95		12	5	96	8
	126	1.75	96	.35	.96		7	4	97	13
	130	2.00	98	.30	.98		6	2	99	14
General Ability index	108	.50	70	1.00	.63	.93	20	37	64	0
	110	.70	75	1.00	.65		20	35	66	0
	115	1.00	84	.90	.82		18	18	83	2
	120	1.33	91	.80	.88		16	12	89	4
	122	1.50	93	.65	.90		13	10	91	7
	126	1.75	96	.40	.95		8	5	96	12
	130	2.00	98	.30	.99		6	1	100	14

Note. SD = standard deviation of the normal curve; ROC/AUC = receiver operating characteristic/area under the curve.

4. Because the SAGES-3 subtests and composites measure different aspects of cognitive ability, they should be significantly intercorrelated.
5. Because the test was built to conform to a particular model of cognitive ability, a factor analysis of the subtests should confirm the relationship of the subtests to the constructs in the model (i.e., the subtests should load on factors that are consistent with that model).

Relationship to Age

The means and standard deviations for the SAGES-3: K–3 subtests at five age intervals and the SAGES-3: 4–8 subtests at six age intervals are reported in Tables 6.39 and 6.40. Coefficients and magnitudes showing the relationship of age to test performance on the subtests are reported in the bottom two rows of the tables.

Table 6.39
Raw Score Means (and Standard Deviations) and Correlations With Age for the SAGES-3: K–3 at Five Age Intervals

Age		Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science
(in years)	*N*	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
5	154	5 (4)	3 (2)	4 (4)	3 (3)
6	151	6 (4)	4 (3)	5 (5)	4 (3)
7	165	11 (6)	9 (5)	11 (7)	9 (5)
8	184	15 (8)	13 (7)	14 (8)	12 (5)
9	154	16 (8)	14 (7)	16 (8)	13 (6)
Correlation with age		.56	.61	.56	.62
Magnitude[a]		Large	Large	Large	Large

[a] Magnitude of the effect sizes based on Hopkins's (2002) criteria for interpreting correlation coefficients.

Table 6.40
Raw Score Means (and Standard Deviations) and Correlations With Age for the SAGES-3: 4–8 at Six Age Intervals

Age		Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science
(in years)	*N*	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)	*M* (*SD*)
9	157	11 (5)	8 (6)	7 (5)	7 (4)
10	179	13 (6)	11 (7)	9 (6)	9 (5)
11	156	14 (7)	13 (8)	10 (6)	11 (7)
12	189	14 (6)	13 (9)	11 (6)	11 (8)
13	167	15 (7)	15 (10)	13 (7)	14 (9)
14	175	15 (7)	18 (11)	13 (7)	15 (9)
Correlation with age		.21	.34	.30	.35
Magnitude[a]		Small	Moderate	Moderate	Moderate

[a] Magnitude of the effect sizes based on Hopkins's (2002) criteria for interpreting correlation coefficients.

The SAGES-3: K–3 subtest raw score means become larger as the students grow older, an observation that demonstrates that the content of the subtests is in fact developmental in nature. The conclusion is verified by the size of the correlation coefficients at the bottom of the table, which are all large in magnitude. On the SAGES-3: 4–8, the subtest raw score means also become larger as the students grow older. With one exception, the correlations are moderate in magnitude, suggesting that the relationship between age and test performance decreases somewhat as students get older.

Differences Among Groups

One way of establishing a test's validity is to study the performance of different diagnostic groups of students on the test. Each group's test results should be consistent with what is known or expected relative to the group's cognitive makeup. In the case of the SAGES-3, a test of reasoning and academic abilities, one would expect that students with disabilities that adversely affect those skills would do less well on the test than students without such disabilities. For example, students who are diagnosed as having intellectual impairment would be expected to do poorly on the test compared to other students. Conversely, one would expect students who are identified as having a high IQ to do very well on the test.

Two studies are described in this section. In the first study, we present the mean subtest and composite indexes for three selected exceptionality subgroups in the total SAGES-3 sample (i.e., the normative and validity study samples). In the second study, we present the results of mean difference analyses between selected exceptionality subgroups and a demographically matched comparison sample from the entire pool of SAGES-3 examinees.

First, we examined the mean subtest and composite indexes for students formally identified in three exceptionality subgroups (high IQ, gifted and talented, and learning disability) from the entire pool of SAGES-3 examinees. This sample includes additional cases collected during the standardization phase but not included in the normative sample. The demographic characteristics of these subgroups are presented in Table 6.41.

We differentiated gifted and talented from the high-IQ (>129) subgroup because schools have more broadly defined *giftedness* to include students who meet criteria in three of four areas: mental ability, achievement, creativity, and motivation (Florida Department of Education, Division of Public Schools, Bureau of Curriculum and Instruction, 2013; Georgia Department of Education, 2008). Although, historically, giftedness has been synonymous with an IQ in the top 2% (e.g., IQ = 130) (McIntosh, Dixon, & Pierson, 2012), it is now possible for students who have IQ scores of less than 130 to be identified as gifted and talented. In light of these changes to gifted and talented identification, we concluded it was important to examine the performance of both subgroups.

We would expect the high-IQ and gifted and talented subgroups to exhibit reasoning and academic abilities in the *possibly* gifted to *very likely* gifted range, while we would expect the learning disability subgroup to exhibit reasoning abilities in the *unlikely* range and academic abilities in the *very unlikely* range. Indeed, Table 6.42 indicates that the exceptionality subgroups performed as expected. For the SAGES-3: K–3, the Reasoning Ability index was 126 (*likely* gifted), the Academic Ability index was 127 (*likely* gifted), and the General Ability index was 130 (*very likely* gifted) for the high-IQ subgroup. The Reasoning Ability index was 121 (*likely* gifted), the Academic Ability index was 124 (*likely* gifted), and the General Ability index was 126 (*likely* gifted) for the gifted and talented subgroup. For the SAGES-3: 4–8, the Reasoning Ability index was 123 (*likely* gifted), the Academic Ability index was 123 (*likely* gifted), and the General Ability index was 125 (*likely* gifted) for the high IQ subgroup. The Reasoning Ability

Table 6.41
Demographic Characteristics of the Samples Used in the SAGES-3 Construct-Identification Validity Studies

	Study				
Sample characteristic	SAGES-3: K–3 high-IQ sample	SAGES-3: 4–8 high-IQ sample	SAGES-3: K–3 gifted and talented sample	SAGES-3: 4–8 gifted and talented sample	SAGES-3: 4–8 learning disability sample
Total number of participants	34	20	218	678	39
Age range (in years)	6–8	10–13	5–9	9–14	9–14
Location	AR, MO, TX, WV	MO, NY	AR, AZ, IL, KS, MD, MO, TX, WV	U.S.	CA, CO, MI, MS, NJ, NY, TX
Gender					
Male	17	12	107	328	21
Female	17	8	111	350	18
Race					
White	31	19	190	597	33
Black/African American	0	0	4	17	2
Asian/Pacific Islander	1	0	13	30	0
American Indian/Alaska Native	0	0	0	3	0
Two or more races	2	1	11	31	4
Hispanic status					
Yes	1	0	14	86	15
No	33	20	204	592	24
Exceptionality status					
None	0	0	0	0	0
Gifted and talented	34	20	218	678	0
Language impairment	0	0	0	0	3
Learning disability	0	0	0	0	38
Attention-deficit/hyperactivity disorder	0	0	0	1	2
Visual impairment	0	0	0	1	0
Other	0	0	0	0	27
Parent education					
Less than Bachelor's degree	19	14	122	414	28
Bachelor's degree	15	6	96	264	11

index was 118 (*possibly* gifted), the Academic Ability index was 118 (*possibly* gifted), and the General Ability index was 120 (likely gifted) for the gifted and talented subgroup. For the learning disability subgroup, the Reasoning Ability index was 83 (*very unlikely* gifted), the Academic Ability index was 82 (*very unlikely* gifted), and the General Ability index was 82 (*very unlikely* gifted).

Table 6.42
SAGES-3 Subtest and Composite Means for Selected Exceptionality Subgroups (Decimals Omitted)

		SAGES-3 value									
		Subtest				Composite					
Subgroup	*N*	Nonverbal Reasoning	Language Arts/Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Probability of giftedness	Academic Ability	Probability of giftedness	General Ability	Probability of giftedness
SAGES-3: K–3 High IQ	34	128 (14)	123 (10)	120 (8)	124 (10)	126 (11)	Likely	127 (10)	Likely	130 (10)	Very likely
SAGES-3: 4–8 High IQ	20	119 (12)	121 (12)	120 (7)	122 (12)	123 (9)	Likely	123 (9)	Likely	125 (8)	Likely
SAGES-3: K–3 Gifted and talented	218	121 (15)	120 (12)	119 (11)	122 (14)	121 (11)	Likely	124 (12)	Likely	126 (12)	Likely
SAGES-3: 4–8 Gifted and talented	678	113 (12)	116 (13)	117 (10)	118 (13)	118 (10)	Possibly	118 (10)	Possibly	120 (10)	Likely
SAGES-3: 4–8 Learning disability	39	89 (15)	84 (10)	83 (9)	84 (11)	83 (11)	Very unlikely	82 (12)	Very unlikely	82 (11)	Very unlikely

Next, we examined the mean differences between selected exceptionality subgroups and a control sample matched on key demographic variables (age, gender, race, and ethnicity). Subgroup mean scores, standard deviations, score differences, and effect sizes are presented for each of the comparisons, which are discussed next. Both Cohen's *d* and effect size *r* are presented in these studies. As previously noted, Hopkins (2002) described effect size *r* in six categories: *r*s less than .10 are very small or trivial, between .10 and .29 are considered small, between .30 and .49 are considered moderate, between .50 and .69 are considered large, between .70 and .89 are considered very large, and .90 and above are considered nearly perfect. Hopkins also described Cohen's *d* in six categories: *d*s less than .20 are very small or trivial, between .20 and .59 are small, between .60 and 1.19 are moderate, between 1.20 and 1.99 are large, between 2.00 and 3.99 are very large, and 4.00 and higher are nearly perfect. Each of the studies of subgroup differences is discussed in the following sections.

High IQ

For SAGES-3: K–3, a sample of 34 students with IQs higher than 129 was compared to a sample of 34 students who were selected from the SAGES-3 pool of examinees and matched on age, gender, race, and ethnicity. For SAGES-3: 4–8, a sample of 20 students with IQs higher than 129 was compared to a sample of 20 students who were selected from the SAGES-3 pool of examinees and matched on age, gender, race, ethnicity, and parent education. The results were compared with those of a matched sample selected from the standardization sample. The demographic characteristics of the high-IQ sample are reported in Table 6.41. The performance results of the high-IQ sample and the matched sample are provided in Tables 6.43 and 6.44.

As expected, the high-IQ group functioned significantly better in reasoning and academic abilities than the matched sample on both the SAGES-3: K–3 and the SAGES-3: 4–8. In fact, the high-IQ group scored 4 or more *SEM*s higher than the matched sample on all of the SAGES-3: K–3 scores. The range of score differences for the subtests was 19.83 to 25.82; for the composites the range was 24.71 to 30.06. On the SAGES-3: 4–8, the high-IQ group scored 4 or more *SEM*s higher than the matched sample on all of the SAGES-3: 4–8 scores. The range of score differences for the subtests was 18.75 to 20.25; for the composites the range was 20.85 to 24.30. The magnitudes of the effect sizes for the differences across subtests and composites were all very large. These results provide strong support for the validity of the SAGES-3 as a measure of reasoning ability and academic achievement.

Gifted and Talented

For the SAGES-3: K–3, a sample of 218 students identified as gifted and talented was compared to a sample of 218 students who were selected from the SAGES-3 pool of examinees and matched on age, gender, race, and ethnicity. For the SAGES-3: 4–8, a sample of 678 examinees identified as gifted and talented was compared to a sample of 678 students who were selected from the SAGES-3 pool of examinees and matched on age, gender, race, ethnicity, and parent education. The results were compared with those of a matched sample selected from the standardization sample. The demographic characteristics of the gifted and

Table 6.43
Comparison of SAGES-3: K–3 Scores for a High-IQ Sample and a Matched Sample

	High IQ ($n = 34$)	Typical ($n = 34$)	Difference		Effect	Effect	
SAGES-3: K–3 value	*M* (*SD*)	*M* (*SD*)	score	*t*	size *d*	size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	127.82 (13.87)	102.00 (11.32)	25.82	8.41 ***	2.04	.71	Very large
Language Arts/Social Studies	122.94 (9.67)	98.06 (12.62)	24.88	9.13 ***	2.21	.74	Very large
Verbal Reasoning	120.24 (7.65)	100.41 (8.58)	19.83	10.06 ***	2.44	.77	Very large
Mathematics/Science	124.44 (10.38)	99.12 (8.98)	25.32	10.76 ***	2.49	.78	Very large
Composite							
Reasoning Ability	126.18 (10.56)	101.47 (9.11)	24.71	10.33 ***	2.51	.78	Very large
Academic Ability	126.62 (9.63)	98.21 (9.10)	28.41	12.51 ***	3.03	.83	Very large
General Ability	129.88 (10.00)	99.82 (7.19)	30.06	14.23 ***	3.45	.87	Very large

Note. Samples were matched according to age, gender, race, ethnicity, and parent education.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

***$p < .001$.

Table 6.44
Comparison of SAGES-3: 4–8 Scores for a High-IQ Sample and a Matched Sample

	High IQ ($n = 20$)	Typical ($n = 20$)	Difference		Effect	Effect	
SAGES-3: 4–8 value	*M* (*SD*)	*M* (*SD*)	score	*t*	size *d*	size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	119.25 (11.82)	99.50 (6.22)	19.75	6.61 ***	2.09	.72	Very large
Language Arts/Social Studies	120.70 (11.66)	100.45 (8.73)	20.25	6.22 ***	2.00	.70	Very large
Verbal Reasoning	119.80 (7.41)	101.05 (11.12)	18.75	6.28 ***	2.00	.70	Very large
Mathematics/Science	121.70 (11.52)	102.10 (6.67)	19.60	6.58 ***	2.08	.72	Very large
Composite							
Reasoning Ability	123.45 (8.94)	100.45 (7.52)	23.00	8.81 ***	2.78	.81	Very large
Academic Ability	122.95 (9.24)	102.10 (6.89)	20.85	8.09 ***	2.56	.79	Very large
General Ability	125.40 (8.41)	101.10 (5.59)	24.30	10.76 ***	3.40	.86	Very large

Note. Samples were matched according to age, gender, race, ethnicity, and parent education.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

***$p < .001$.

talented sample are reported in Table 6.41. The performance results of the gifted and talented sample and the matched sample are provided in Tables 6.45 and 6.46.

As expected, the gifted and talented group functioned significantly better in reasoning and academic abilities than the matched sample on both the SAGES-3: K–3 and the SAGES-3: 4–8. In fact, the gifted and talented group scored 3 or more *SEM*s higher than the matched sample on all of the SAGES-3: K–3 scores. The range of score differences for the subtests was 17.90 to 19.24; for the composites the range was 20.39 to 23.33. On the SAGES-3: 4–8, the gifted and talented group scored 4 or more *SEM*s higher than the matched sample on all of the SAGES-3: 4–8 scores. The range of score differences for the subtests was 16.10 to 19.91; for the composites the range was 19.85 to 22.34. The magnitudes of the effect sizes for the differences across subtests and composites were all large. These results provide strong support for the validity of the SAGES-3 as a measure of reasoning and academic abilities.

Learning Disability

On the SAGES-3: 4–8, a sample of 39 students diagnosed with a diverse collection of learning disabilities was compared to a sample of 39 examinees matched on age, gender, race, ethnicity, and parent education. The demographic characteristics of the learning disability sample are reported in Table 6.41. The performance results of the two samples are provided in Table 6.47.

Examinees with learning disabilities scored at least 2 *SEM*s lower than the control sample on both the SAGES-3 subtests and at least 5 *SEM*s lower on the composites. The difference scores for the subtests ranged from −15.77 to −11.47; for the composites the range was −17.18 to −15.77. With one exception (Nonverbal Reasoning), the effect sizes for the difference scores were large for the subtests and the composites.

Relationship to Spoken Language

Most professionals agree that language ability and general ability are related. If true, one way to demonstrate that a general ability test is valid would be to show that its scores are related to those of spoken language tests.

To investigate this kind of validity, we correlated the SAGES-3 with the following measures of spoken language:

- *Test of Early Language Development–Fourth Edition* (TELD-4; Hresko, Reid, & Hammill, 2018)
- *Test of Language Development–Primary: Fifth Edition* (TOLD-P: 5; Hammill & Newcomer, 2019b)
- *Test of Language Development–Intermediate: Fifth Edition* (TOLD-I: 5; Hammill & Newcomer, 2019a)

In all, we investigated the SAGES-3's relationship to three different spoken language tests using three different samples of students as participants. The demographics of these samples are described in Table 6.48. The results for these studies are presented in Table 6.49. As can be readily seen, the correlation

Table 6.45
Comparison of SAGES-3: K–3 Scores for a Gifted and Talented Sample and a Matched Sample

	Gifted and talented ($n = 218$)	Typical ($n = 218$)					
SAGES-3: K–3 value	*M* (*SD*)	*M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	120.82 (14.58)	101.58 (15.21)	19.24	13.48 ***	1.29	.54	Large
Language Arts/Social Studies	120.31 (11.30)	101.80 (14.43)	18.51	14.91 ***	1.43	.58	Large
Verbal Reasoning	118.61 (10.53)	100.71 (14.43)	17.90	14.80 ***	1.42	.58	Large
Mathematics/Science	122.01 (14.08)	103.03 (14.56)	18.98	13.83 ***	1.33	.55	Large
Composite							
Reasoning Ability	121.42 (11.23)	101.03 (14.71)	20.39	16.26 ***	1.56	.61	Large
Academic Ability	123.67 (12.33)	102.56 (14.62)	21.11	16.29 ***	1.56	.62	Large
General Ability	125.55 (11.78)	102.22 (14.33)	23.33	18.56 ***	1.78	.66	Large

Note. Samples were matched according to age, gender, race, ethnicity, and parent education.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

***$p < .001$.

Table 6.46
Comparison of SAGES-3: 4–8 Scores for a Gifted and Talented Sample and a Matched Sample

	Gifted and talented ($n = 678$)	Typical ($n = 678$)					
SAGES-3: 4–8 value	*M* (*SD*)	*M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	113.37 (12.49)	97.27 (14.04)	16.10	22.31 ***	1.21	.52	Large
Language Arts/Social Studies	115.53 (12.53)	98.11 (13.82)	17.42	24.33 ***	1.32	.55	Large
Verbal Reasoning	117.12 (9.94)	98.35 (14.72)	18.77	27.52 ***	1.49	.60	Large
Mathematics/Science	117.82 (12.92)	97.91 (13.56)	19.91	27.67 ***	1.50	.60	Large
Composite							
Reasoning Ability	118.07 (10.17)	97.61 (14.38)	20.46	30.25 ***	1.64	.63	Large
Academic Ability	118.02 (10.42)	98.17 (13.54)	19.85	30.27 ***	1.64	.63	Large
General Ability	119.96 (9.89)	97.62 (13.74)	22.34	34.37 ***	1.87	.68	Large

Note. Samples were matched according to age, gender, race, ethnicity, and parent education.

[a]Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

***$p < .001$.

Table 6.47
Comparison of SAGES-3: 4–8 Scores for a Sample With Learning Disabilities and a Matched Sample

	Learning disability ($n = 39$)	Typical ($n = 39$)					
SAGES-3: 4–8 value	*M* (*SD*)	*M* (*SD*)	Difference score	*t*	Effect size *d*	Effect size *r*	Magnitude[a]
Subtest							
Nonverbal Reasoning	88.56 (15.32)	100.03 (10.03)	–11.47	3.91 ***	–.89	–.40	Moderate
Language Arts/Social Studies	84.26 (10.04)	98.44 (11.40)	–14.18	5.83 ***	–1.32	–.55	Large
Verbal Reasoning	82.72 (9.33)	98.49 (12.19)	–15.77	6.42 ***	–1.45	–.59	Large
Mathematics/Science	84.23 (11.00)	98.44 (11.68)	–14.21	5.53 ***	–1.25	–.53	Large
Composite							
Reasoning Ability	83.46 (12.61)	99.23 (11.48)	–15.77	5.78 ***	–1.31	–.55	Large
Academic Ability	81.95 (12.02)	98.85 (11.00)	–16.90	6.48 ***	–1.47	–.59	Large
General Ability	81.51 (11.44)	98.69 (11.23)	–17.18	6.69 ***	–1.52	–.60	Large

Note. Samples were matched according to age, gender, race, ethnicity, and parent education.

[a] Values of the magnitude of the effect size *r* between the two groups are based on Hopkins's (2002) criteria.

***$p < .001$.

coefficients depicting the SAGES-3 composite indexes and the indexes of the spoken language tests range from .50 to .82 (large to very large in magnitude), which provides strong evidence of the construct-identification validity of the test.

Relationship Among Subtests

If the SAGES-3 subtests do in fact measure reasoning and academic abilities, they should correlate with each other to some moderate degree (i.e., .30 or higher). To investigate this kind of validity, we correlated the SAGES-3 subtest indexes using the entire normative sample as participants. As can be seen in Table 6.50, the correlation coefficients between the subtests range from .45 to .61 for the SAGES-3: K–3 and .49 to .67 for the SAGES-3: 4–8. The magnitude of the coefficients ranges from moderate to large. We also intercorrelated the Reasoning Ability and Academic Ability indexes, and the correlation coefficient was .65 (large) for the SAGES-3: K–3 and .69 (large) for the SAGES-3: 4–8.

Authorities are understandably reluctant to specify precisely how large a correlation coefficient should be 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-3 subtest coefficients are too low, it means that the subtests are measuring unrelated abilities rather than differing aspects of reasoning and academic abilities. If the

Table 6.48
Demographic Characteristics of the Samples Used in the SAGES-3 Construct-Identification Validity Studies With Language Measures

	Study		
Sample characteristic	**TELD-4**	**TOLD-P: 5**	**TOLD-I: 5**
Total number of participants	35	46	30
Age range (in years)	5–7	5–8	9–14
Location	AZ, CO, IL, NY, TX, WV	AZ, CA, CO, IL, MI, MN, NJ, NY, TX	AZ, MI, MN, NE, NY, TX
Gender			
Male	16	24	14
Female	19	26	16
Race			
White	29	38	25
Black/African American	6	7	2
Asian/Pacific Islander	0	1	3
Hispanic status			
Yes	5	6	3
No	30	40	27
Exceptionality status			
None	32	41	28
Gifted and talented	2	2	1
Language impairment	0	1	0
Learning disability	0	1	1
Attention-deficit/hyperactivity disorder	1	1	1
Other	0	1	0

Note TELD-4 = *Test of Early Language Development–Fourth Edition* (Hresko, Reid, & Hammill, 2018); TOLD-P: 5 = *Test of Language Development–Primary: Fifth Edition* (Hammill & Newcomer, 2019b); TOLD-I: 5 = *Test of Language Development–Intermediate: Fifth Edition* (Hammill & Newcomer, 2019a).

coefficients are too high, it means that the subtests are measuring the same ability to 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 based on a single predictor rarely exceed .30 or .40. Taking these figures as guides, one can see that all 12 coefficients reported in Table 6.50 exceed the .30 criterion of Anastasi and Urbina (1997), as well as the .40 criterion of Nunnally and Bernstein (1994), providing more evidence supporting the validity of the SAGES-3 subtests.

Table 6.49
Correlations Between SAGES-3 and Measures of Spoken Language (Decimals Omitted)

	SAGES-3 value									
	Subtest				Composite					
Measures of language	Nonverbal Reasoning	Language Arts/Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Magnitude[a]	Academic Ability	Magnitude[a]	General Ability	Magnitude[a]
TELD-4 (*N* = 35)	47 (42)	54 (51)	48 (48)	63 (61)	52 (51)	Large	59 (60)	Large	58 (60)	Large
TOLD-P: 5 (*N* = 46)	40 (39)	51 (48)	52 (55)	47 (50)	51 (53)	Large	50 (53)	Large	54 (57)	Large
TOLD-I: 5 (*N* = 30)	28 (14)	82 (53)	81 (56)	79 (54)	74 (46)	Very large	82 (56)	Very large	82 (57)	Very large

Note. TELD-4 = *Test of Early Language Development–Fourth Edition* (Hresko, Reid, & Hammill, 2018); TOLD-P: 5 = *Test of Language Development–Primary: Fifth Edition* (Hammill & Newcomer, 2019b); TOLD-I: 5 = *Test of Language Development–Intermediate: Fifth Edition* (Hammill & Newcomer, 2019a). Coefficients outside parentheses are corrected for range effects and reliability of the criterion; coefficients inside parentheses are uncorrected correlation coefficients.

[a] Magnitude of the corrected correlation between the SAGES-3: K–3 indexes and the spoken language measure based on Hopkins's (2002) criteria.

Table 6.50
Intercorrelation of SAGES-3 Subtests for Entire Normative Sample (Decimals Omitted)

Subtest	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science
Nonverbal Reasoning	—	50	49	49
Language Arts/Social Studies	45	—	60	67
Verbal Reasoning	57	55	—	57
Mathematics/Science	50	61	58	—

Note. SAGES-3: K–3 (*N* = 808) values appear below the diagonal. SAGES-3: 4–8 (*N* = 1,023) values appear above the diagonal.

Confirmatory Factor Analysis

One way to investigate construct-identification validity is to examine the degree to which a test's underlying traits can be identified and the extent to which those traits reflect the theoretical model on which the test is based. Because the SAGES-3 is founded on a specific model that describes general ability, confirmatory factor analysis (CFA) can be used to confirm that the factor structure of the SAGES-3 matches the model on which it is based. When a specified model exists, CFA provides a more rigorous test of construct validity than is provided by exploratory factor analysis (EFA). For example, in CFA, each subtest is permitted to load only on the factor that it represents. In EFA, subtests are permitted to load on all factors. CFA also provides guidelines to determine the extent to which the model fits the data. In EFA, no comparable guidelines are available.

To empirically investigate the structural validity of the SAGES-3, we tested a one-factor model of the SAGES-3: K–3 and the SAGES-3: 4–8 using maximum-likelihood CFA. Analyses included the entire normative sample. The results of these models were assessed using five indexes of fit: (a) Wheaton, Muthén, Alwin, and Summers's (1977) relative chi-square (chi-square divided by degrees of freedom); (b) Tucker and Lewis's (1973) index of fit (TLI); (c) Bentler's (1990) comparative fit index (CFI); (d) Bentler and Bonett's (1980) normed fit index (NFI); and (e) Browne and Cudeck's (1993) root mean square error of approximation (RMSEA). The criterion for an acceptable fit varies among different types of indexes. Marsh and Hocevar (1985) suggested that relative chi-square values can be as low as 2 or as high as 5 to indicate a reasonable fit. The TLI, CFI, and NFI values should be at or above .90 to indicate a satisfactory model fit, with values close to 1 indicating a very good fit on any of these indexes. An RMSEA of less than .11 indicates a reasonable fit, and an RMSEA of .05 or less indicates a close fit of the model in relation to the degrees of freedom (Browne & Cudeck, 1993).

The graphic results of the one-factor CFA models for the SAGES-3: K–3 and SAGES-3: 4–8 are presented in Figures 6.1 and 6.2, respectively. The values on the arrows between each subtest (rectangles) and the latent factors (large circles)

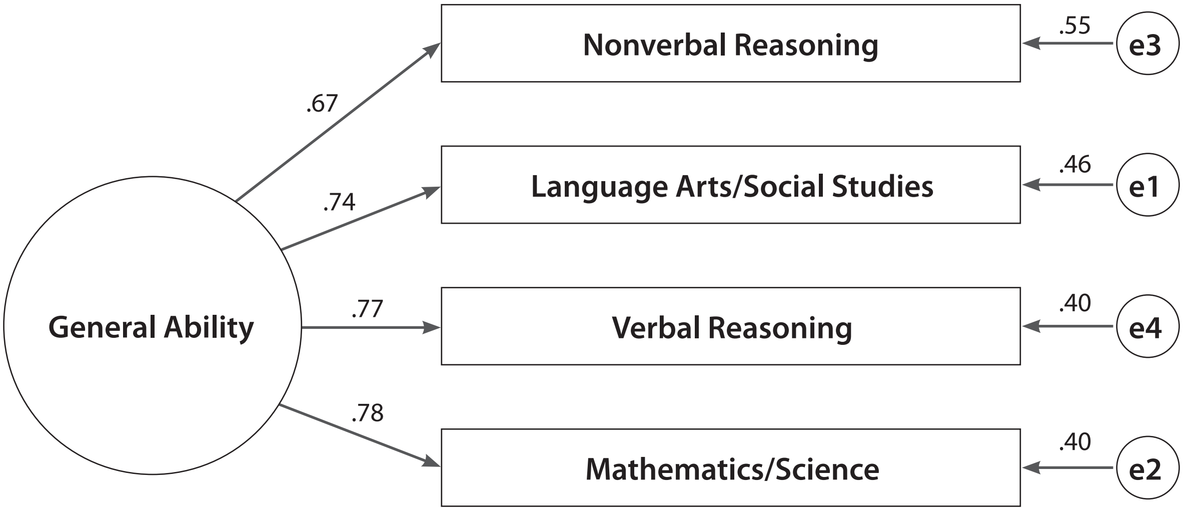

Figure 6.1. SAGES-3: K–3 confirmatory factor analysis.

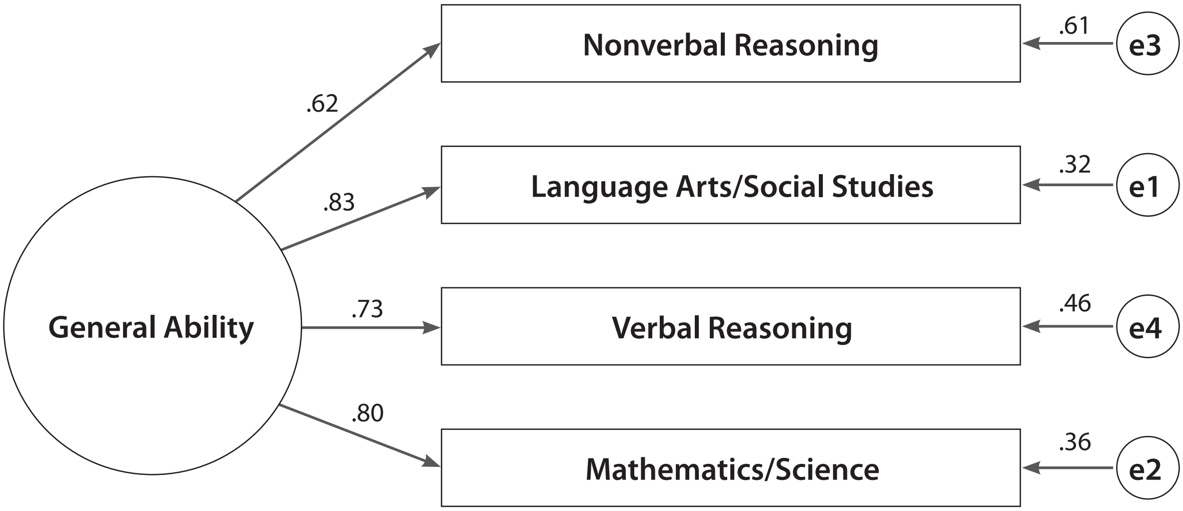

Figure 6.2. SAGES-3: 4–8 confirmatory factor analysis.

are factor loadings. The factor loadings are regression coefficients that represent the influence of these factors on the scales and other factors. The small circles labeled *e1* through *e4* represent unique variance and systematic variance of each subtest that is unrelated to the variances of the other subtests.

The fit statistics for the SAGES-3: K–3 and SAGES-3: 4–8 models for the normative sample are provided in Table 6.51. The results indicate that the SAGES-3 structure is highly plausible and supports interpreting the test as a measure of general ability. Analysis of the SAGES-3 data produced model TLI, CFI, and

Table 6.51
Fit Indexes for SAGES-3 Confirmatory Factor Analysis

	Fit index						
Model	**Chi-square**	***df***	**Chi-square/*df***	**TLI**	**CFI**	**NFI**	**RMSEA**
SAGES-3: K–3 one factor	28.46	2	14.23	.93	.98	.98	.13
SAGES-3: 4–8 one factor	7.01	2	3.51	.99	.99	.99	.05

Note. TLI = Tucker and Lewis's (1973) index of fit; CFI = Bentler's (1990) comparative fit index; NFI = Bentler and Bonett's (1980) normed fit index; RMSEA = Browne and Cudeck's (1993) root mean square error of approximation.

NFI indexes close to 1. The RMSEA was .13 for the SAGES-3: K–3 and .05 for the SAGES-3: 4–8. The RMSEA tends to favor more complex CFA models, so it is often higher in simple models like the ones for the SAGES-3. When we apply Hopkins's (2002) criteria, the sizes of the factor loadings in Figures 6.1 and 6.2 range from large to very large. When combined with the fit indexes for the one-factor model, these findings indicate support for the organization of subtests to composites on the SAGES-3.

Item Validity

The final assumption deals with item–test correlation. Guilford and Fruchter (1978) pointed out that information about a test's construct validity can be gained by examining the correlation between individual items and the total test results. We discussed this relationship (called *item discrimination*) in the section on item analysis of this manual. Item discrimination values of the SAGES-3 are reported in Tables 6.5 and 6.6. Values of this magnitude are consistent with the hypothesis that the SAGES-3 provides a valid assessment of reasoning and academic abilities and that these values are unlikely in a test having poor construct validity.

Summary of Validity Results

The information provided in this chapter suggests that the SAGES-3, like its predecessor the SAGES-2, is a valid measure of reasoning and academic abilities. Examiners can use the test with confidence for the assessment of students' overall functioning and to determine whether examinees may be eligible for a gifted and talented program. We encourage professionals to continue to study the benefits of the test with different samples, using different statistical procedures and related criterion measures. We also encourage researchers to share their results with us so their findings can be included in subsequent printings of the manual. The accumulation of research data will further clarify the validity of the SAGES-3 and provide guidance for future revisions of the test.