5 Test Reliability

Reliability refers to the consistency with which a test measures an ability. Because unreliable tests yield inaccurate results, reliability coefficients for tests such as the SAGES-3 must be .80 or greater in magnitude to be considered minimally reliable; coefficients of .90 or above are considered the most desirable (Aiken & Groth-Marnat, 2006; Nunnally & Bernstein, 1994; Reynolds, Livingston, & Willson, 2009; Salvia, Ysseldyke, & Witmer, 2017). Anastasi and Urbina (1997) described three sources of error variance: content, time, and scorer. We calculated three types of correlation coefficients—coefficient alpha, test–retest, and scorer difference—to measure these sources of error.

Coefficient Alpha

Error associated with content sampling largely reflects the degree of homogeneity among items within a test or subtest. Because the purpose of a test is to measure a certain characteristic, ability, or content, the more the items relate to each other, the smaller the error in the test will be. Test items that are unrelated to each other are measuring different qualities; therefore, the amount of test error due to content sampling will be large.

Content sampling error (i.e., internal consistency reliability) for the SAGES-3 was investigated by applying Cronbach's (1951) coefficient alpha method. Coefficient alphas for the subtests and composites were calculated at five age intervals for SAGES-3: K–3 and six age intervals for SAGES-3: 4–8 using data from the entire normative sample. Coefficient alphas for the composites were calculated using Guilford's (1954, p. 393) formula. The results are reported in Tables 5.1 and 5.2. The coefficients were averaged using the Fisher *z*-transformation technique. The averaged coefficients are listed at the bottom of the tables. All but one of these average subtest reliability coefficients (Table 5.1, Mathematics/Science, .88) exceeded .90 for all age groups. All of the averaged composite coefficients are above .90, a most desirable level of reliability.

The standard errors of measurement (*SEM*s) for SAGES-3: K–3 values at five ages are reported in Table 5.3; *SEM*s for SAGES-3: 4–8 values at six ages are reported in Table 5.4. The *SEM* estimates the amount of error in an individual's test score due to less-than-perfect reliability of a test. The *SEM* is based on the formula $SEM = SD\sqrt{1 - r_{xx}}$ (SD = standard deviation for the score of interest; r_{xx} = reliability of the score of interest). The average *SEM*s for the subtests and composites are listed at the bottom of Tables 5.3 and 5.4.

An *SEM* can be used to estimate the precision of a score and provide a range of scores (i.e., a confidence interval) in which the student's hypothetical

Table 5.1
Coefficient Alphas for SAGES-3: K–3 Subtests at Five Age Intervals (Decimals Omitted)

	Subtest				Composite		
Age (in years)	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Academic Ability	General Ability
5	89	85	88	86	92	91	94
6	89	82	93	81	94	87	94
7	93	91	95	89	96	94	97
8	94	92	95	90	97	95	98
9	95	94	96	93	97	96	98
Average[a]	92	90	94	88	96	93	97

[a]Calculated using Fisher's average of alpha coefficients across all ages.

Table 5.2
Coefficient Alphas for SAGES-3: 4–8 Subtests at Six Age Intervals (Decimals Omitted)

	Subtest				Composite		
Age (in years)	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Academic Ability	General Ability
9	89	92	91	85	93	93	96
10	91	93	92	91	94	95	97
11	93	94	93	94	96	96	98
12	91	95	92	95	94	97	98
13	92	96	95	96	96	98	98
14	93	96	93	95	95	97	98
Average[a]	92	95	93	93	95	96	97

[a]Calculated using Fisher's average of alpha coefficients across all ages.

true score is likely to fall. *SEM*-based confidence intervals are calculated using the following formula:

$$\begin{aligned}&\text{The lower bound value} = \text{Obtained score} - z * (SEM)\\&\text{The upper bound value} = \text{Obtained score} + z * (SEM),\end{aligned}$$

where z is the area under the normal curve for the .05 or .01 level of probability.

The clinical value of *SEM*s is exemplified by a 12-year-old student, Aden, who earned a scaled score of 114 on the Nonverbal Reasoning subtest of the SAGES-3: 4–8. The *SEM* for this subtest is 4. Thus, the examiner knows with

Table 5.3
Standard Errors of Measurement (*SEMs*) for SAGES-3: K–3 Subtests at Five Age Intervals

	Subtest				Composite		
Age (in years)	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Academic Ability	General Ability
5	5	6	5	6	4	5	4
6	5	6	4	7	4	5	4
7	4	4	4	5	3	4	3
8	4	4	3	5	3	3	2
9	4	4	3	4	3	3	2
Average[a]	4	5	4	5	3	4	3

[a]Calculated using Fisher's average of alpha coefficients across all ages.

Table 5.4
Standard Errors of Measurement (*SEMs*) for SAGES-3: 4–8 Subtests at Six Age Intervals

	Subtest				Composite		
Age (in years)	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science	Reasoning Ability	Academic Ability	General Ability
9	5	4	5	6	4	4	3
10	5	4	4	5	4	3	3
11	4	4	4	4	3	3	2
12	4	3	4	3	4	3	2
13	4	3	3	3	3	2	2
14	4	3	4	3	3	3	2
Average[a]	4	3	4	4	3	3	3

[a]Calculated using Fisher's average of alpha coefficients across all ages.

68% probability (114 ± *SEM*) that her true score lies between 110 and 118, with 95% probability that her true score lies between 106 and 122 (114 ± *SEM* * 1.96), and with 99% probability that her true score lies between 104 and 124 (114 ± *SEM* * 2.58). Obviously, the smaller the *SEM*, the more confidence one can have in the test's results.

One cannot always assume that a test that is reliable for a general population will be equally reliable for every subgroup within that population. Therefore, the alphas for selected subgroups within the normative sample were calculated. They are reported in Tables 5.5 and 5.6. The subgroups represent a broad spectrum of populations, embracing gender, racial/ethnic, and two exceptionality

Table 5.5
Coefficient Alphas for Selected Subgroups of the SAGES-3: K–3 Normative Sample (Decimals Omitted)

		SAGES-3: K–3 scores			
Subgroup	*N*	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science
Gender					
Male	407	95	94	96	93
Female	401	94	94	95	91
Race/ethnicity					
White	604	95	94	96	92
Black/African American	140	94	91	95	89
Asian/Pacific Islander	28	93	96	96	95
Two or more races	33	94	93	94	89
Hispanic	190	94	93	95	91
Exceptionality status					
Gifted and talented	65	92	91	91	88
Learning disability	13	85	90	89	90

Table 5.6
Coefficient Alphas for Selected Subgroups of the SAGES-3: 4–8 Normative Sample (Decimals Omitted)

		SAGES-3: 4–8 scores			
Subgroup	*N*	Nonverbal Reasoning	Language Arts/ Social Studies	Verbal Reasoning	Mathematics/ Science
Gender					
Male	509	92	96	94	95
Female	506	92	95	93	95
Race/ethnicity					
White	757	92	95	93	95
Black/African American	153	92	96	94	95
Asian/Pacific Islander	46	90	95	94	94
Two or more races	45	92	96	94	90
Hispanic	229	91	95	93	94
Exceptionality status					
Gifted and talented	114	91	93	91	94
Learning disability	32	92	89	86	87

categories. The consistently large alphas in the table demonstrate that the SAGES-3 is equally reliable for all the subgroups investigated and support the idea that the test contains little or no bias relative to these groups.

Test–Retest

Time sampling error refers to the extent to which a person's test performance might change because of the passage of time between administrations. Time sampling error is usually estimated by the test–retest method. In this procedure, the test is given to groups of individuals, a period of time is allowed to pass, the same individuals are tested again, and the results of the two tests are compared. The degree of similarity between the two scores indicates the amount of stability reliability possessed by the test.

We investigated this type of reliability using a sample of 113 students ages 5 years 0 months through 14 years 11 months. The sample was divided into four age groups: 5-0 to 7-11 and 8-0 to 9-11 for SAGES-3: K–3 and 9-0 to 11-11 and 12-0 to 14-11 for SAGES-3: 4–8. The demographic characteristics of these samples are provided in Table 5.7. The SAGES-3 was administered twice to each student; the intervening time was approximately two weeks. After the testing was completed, the indexes for the SAGES-3 subtests and composites were correlated and corrected for range effects. The results of this analysis are presented in Tables 5.8 and 5.9. Uncorrected coefficients appear within parentheses. The resulting coefficients are large enough to support the ideas that the SAGES-3 scores contain little time sampling error and that the results are consistent over time.

Scorer Difference

Scorer difference reliability refers to the amount of test error due to examiner variability in scoring. Unreliable scoring is usually the result of clerical errors or improper application of standard scoring criteria on the part of an examiner. Scorer error can be reduced considerably by the availability of clear administration procedures, detailed guidelines governing scoring, and opportunities to practice scoring. Still, test constructors should demonstrate statistically the amount of error in their tests that is due to different scorers. To do this, two trained individuals should score a set of tests independently (Anastasi & Urbina, 1997). The correlation between scorers yields a relational index of agreement and is a measure of interscorer reliability.

To study interscorer reliability, two members of the PRO-ED research staff independently scored 50 SAGES-3 test protocols drawn at random from the normative sample. The scorings of the two scorers were correlated. The resulting coefficients for the subtests and composites were above .97. These coefficients are high enough to be accepted as evidence of SAGES-3 scorer reliability.

Summary of Reliability Results

The SAGES-3's overall reliability is summarized in Table 5.10. The contents of this table show the test's status relative to three types of reliability coefficients

Table 5.7
Demographic Characteristics of the Samples Used in the SAGES-3 Test–Retest Studies

	Sample			
Sample characteristic	**SAGES-3: K–3**		**SAGES-3: 4–8**	
Total number of participants	25	28	32	28
Age range	5-0 to 7-11	8-0 to 9-11	9-0 to 11-11	12-0 to 14-11
Location	CA, CO, ID, IL, MD, MN, MO, PA, TX, VA	AZ, CA, ID, IL, MI, MO, MS	ID, MO, NE	CA, ID, MN, MO, NJ, NY
Gender				
Male	14	16	15	13
Female	11	12	17	15
Race				
White	24	23	31	27
Black/African American	1	3	0	0
Two or more races	0	2	1	1
Hispanic status				
Yes	2	5	31	3
No	23	23	0	25
Exceptionality status				
None	11	12	6	19
Gifted and talented	14	16	26	9

and possible test error described by Anastasi and Urbina (1997), the coefficient alphas listed in the table are the averaged alphas reported in Tables 5.1 and 5.2, and the test–retest coefficients are from the "Combined Sample" sections in Tables 5.8 and 5.9.

The SAGES-3 composite indexes satisfy the most demanding standards for reliability, including those of Nunnally and Bernstein (1994), Salvia et al. (2017), and Reynolds et al. (2009). These authors recommended that when important decisions are to be made for individuals, the minimum standard for a reliability coefficient should be .90. The SAGES-3 composites meet this rigorous standard. These findings strongly suggest that the test possesses relatively little test error and that test users can have confidence in the SAGES-3's results.

Table 5.8
Corrected (and Uncorrected) Test–Retest Reliability for SAGES-3: K–3

	First testing	Second testing	
SAGES-3: K–3 score	***M* (*SD*)**	***M* (*SD*)**	**r_c (r_u)**
Ages 5-0 to 7-11 (*n* = 25)			
Subtest			
Nonverbal Reasoning	119 (17)	121 (18)	.86 (.92)
Language Arts/Social Studies	109 (19)	111 (17)	.86 (.93)
Verbal Reasoning	111 (12)	113 (12)	.87 (.80)
Mathematics/Science	112 (15)	114 (16)	.88 (.90)
Composite			
Reasoning Ability	116 (15)	118 (16)	.93 (.94)
Academic Ability	111 (18)	114 (17)	.87 (.93)
General Ability	116 (17)	118 (17)	.92 (.95)
Ages 8-0 to 9-11 (*n* = 28)			
Subtest			
Nonverbal Reasoning	114 (13)	115 (11)	.96 (.91)
Language Arts/Social Studies	114 (17)	117 (14)	.86 (.87)
Verbal Reasoning	116 (15)	117 (15)	.88 (.89)
Mathematics/Science	117 (17)	119 (17)	.85 (.90)
Composite			
Reasoning Ability	116 (15)	117 (14)	.95 (.95)
Academic Ability	117 (18)	120 (16)	.91 (.94)
General Ability	119 (17)	121 (17)	.94 (.96)
Combined sample (*n* = 53)			
Subtest			
Nonverbal Reasoning	116 (15)	118 (15)	.91 (.91)
Language Arts/Social Studies	112 (18)	114 (16)	.86 (.91)
Verbal Reasoning	114 (14)	115 (14)	.87 (.85)
Mathematics/Science	114 (16)	116 (17)	.86 (.90)
Composite			
Reasoning Ability	116 (15)	118 (14)	.94 (.94)
Academic Ability	114 (18)	117 (17)	.90 (.94)
General Ability	117 (17)	120 (16)	.93 (.95)

Note. *M* = mean; *SD* = standard deviation; r_c = corrected correlation coefficient; r_u = uncorrected correlation coefficient.

Table 5.9
Corrected (and Uncorrected) Test–Retest Reliability for SAGES-3: 4–8

SAGES-3: 4–8 score	First testing *M* (*SD*)	Second testing *M* (*SD*)	r_c (r_u)
Ages 9-0 to 11-11 (*n* = 32)			
Subtest			
Nonverbal Reasoning	114 (13)	116 (11)	.94 (.86)
Language Arts/Social Studies	112 (11)	113 (11)	.94 (.83)
Verbal Reasoning	112 (13)	115 (10)	.86 (.69)
Mathematics/Science	114 (12)	114 (11)	.94 (.90)
Composite			
Reasoning Ability	115 (13)	118 (10)	.94 (.85)
Academic Ability	114 (11)	114 (10)	.98 (.92)
General Ability	116 (12)	118 (11)	.97 (.91)
Ages 12-0 to 14-11 (*n* = 28)			
Subtest			
Nonverbal Reasoning	104 (12)	103 (15)	.93 (.90)
Language Arts/Social Studies	106 (14)	104 (15)	.83 (.82)
Verbal Reasoning	110 (15)	114 (12)	.80 (.72)
Mathematics/Science	107 (14)	110 (12)	.91 (.85)
Composite			
Reasoning Ability	108 (14)	110 (14)	.88 (.86)
Academic Ability	107 (13)	108 (13)	.94 (.89)
General Ability	108 (13)	110 (13)	.94 (.90)
Combined sample (*n* = 60)			
Subtest			
Nonverbal Reasoning	109 (14)	110 (15)	.90 (.89)
Language Arts/Social Studies	109 (13)	109 (14)	.88 (.83)
Verbal Reasoning	111 (14)	115 (11)	.83 (.70)
Mathematics/Science	111 (13)	112 (12)	.94 (.87)
Composite			
Reasoning Ability	112 (14)	115 (13)	.91 (.86)
Academic Ability	111 (12)	111 (12)	.96 (.91)
General Ability	112 (13)	114 (12)	.95 (.91)

Note. *M* = mean; *SD* = standard deviation; r_c = corrected correlation coefficient; r_u = uncorrected correlation coefficient.

Table 5.10
Summary of SAGES-3 Reliability Relative to Three Types of Reliability (Decimals Omitted)

	Type of reliability coefficient		
SAGES-3 value	**Coefficient alpha**	**Test–retest**	**Scorer**
Subtest			
Nonverbal Reasoning K–3	92	91	99
Nonverbal Reasoning 4–8	92	90	99
Language Arts/Social Studies K–3	90	86	99
Language Arts/Social Studies 4–8	95	88	99
Verbal Reasoning K–3	94	87	99
Verbal Reasoning 4–8	93	83	98
Mathematics/Science K–3	88	86	99
Mathematics/Science 4–8	93	94	99
Composite			
Reasoning Ability Index K–3	96	94	99
Reasoning Ability Index 4–8	95	91	99
Academic Ability Index K–3	93	90	99
Academic Ability Index 4–8	96	96	99
General Ability Index K–3	97	93	99
General Ability Index 4–8	97	95	99
Sources of test error	Content sampling, content heterogeneity	Time sampling	Interscorer differences

Note. The sources of error variance are from *Psychological Testing* (7th ed., p.101), by A. Anastasi and S. Urbina, 1997, Upper Saddle River, NJ: Prentice Hall.