



Test Reliability

THE CONCEPT OF RELIABILITY REFERS to the consistency with which any measuring instrument (e.g., a test, a scale, a clock) estimates various attributes of something. It is a key concept in measurement theory because it relates to the practical usefulness of all types and systems of measurement. Whether time, weight, height, distance, texture, achievement, feelings, or aptitude is being measured, ideas about reliability of measurement are important and must be considered.

With regard to psychometric measurement, tests that have adequate reliability will measure “true”; that is, they will yield more or less the same scores across periods of time and across different examiners. Tests that have poor reliability will yield markedly different scores when given at different times or when administered by different people. Obviously, reliability has considerable relevance when tests are used to identify students as gifted and talented.

The study of a test’s reliability centers on estimating the amount of error associated with its scores. When error variance is investigated, results are usually reported in terms of a reliability coefficient, a specific use of a common correlation coefficient. For a test such as the SAGES–2 to be considered minimally reliable, its reliability coefficients must approximate or exceed .80 in magnitude; coefficients of .90 or above are considered to be the most desirable (Aiken, 1994; Nunnally & Bernstein, 1994; Salvia & Ysseldyke, 1998). The status of the SAGES–2 subtest scores relative to three sources of error variance—content sampling, time sampling, and interscorer differences—is discussed in this chapter.

Content Sampling

Error associated with content sampling reflects the degree of homogeneity among items within a test or subtest. Because the purpose of the 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. If the items are unrelated to each other, they are most likely measuring different qualities, and the amount of test error due to content sampling would be great.

The internal consistency reliability of the items on the SAGES-2 subtests was investigated using Cronbach's (1951) coefficient alpha, a generalization of the Kuder-Richardson Formula #20 for dichotomously scored items. The scores of both normative samples (i.e., normal and gifted normative samples) were used for this analysis. Coefficient alphas for the subtests of SAGES-2:K-3 and SAGES:-2:4-8 are presented in Table 5.1 and Table 5.2, respectively. According to the tables, 97% of the alphas for the SAGES-2 subtests reach or exceed .80, the criterion for acceptable reliability; 74% attain .90 or above, the optimal level.

The alphas in Tables 5.1 and 5.2 were averaged using the z -transformation method for averaging correlation coefficients. The averaged alphas are found in the column at the extreme right of each table. The figures in the column show the overall reliability of the SAGES-2 subtests regardless of age. Inspection of the averaged alphas in the two tables indicates that all of the subtests are reliable in that they range from .85 to .94.

Table 5.1
Coefficient Alphas for SAGES-2:K-3 Subtests
at Five Age Intervals (Decimals Omitted)

SAGES-2 Sample	Subtest	Age					Average
		5	6	7	8	9	
Normal	Mathematics/Science	77	86	91	91	91	88
	Language Arts/Social Studies	77	87	88	89	89	87
	Reasoning	93	92	93	93	93	93
Gifted	Mathematics/Science	91	92	93	91	90	91
	Language Arts/Social Studies	88	93	92	91	90	91
	Reasoning	91	93	93	93	94	93

Table 5.2
Coefficient Alphas for SAGES-2:4-8 Subtests
at Six Age Intervals (Decimals Omitted)

SAGES-2 Sample	Subtest	Age						Average
		9	10	11	12	13	14	
Normal	Mathematics/Science	94	92	93	95	96	94	94
	Language Arts/Social Studies	91	93	94	95	95	95	94
	Reasoning	92	92	89	88	87	90	90
Gifted	Mathematics/Science	91	92	91	90	91	93	91
	Language Arts/Social Studies	90	92	92	92	92	93	92
	Reasoning	89	89	83	87	85	82	85

The standard errors of measurement (*SEMs*) listed in Tables 5.3 and 5.4 (for SAGES-2:K-3 and SAGES-2:4-8, respectively) provide a confidence interval that surrounds a particular test score. For example, consider Andrea's Mathematics/Science quotient of 127 (comparing the raw score to the normal norms). Because the associated *SEM* is 3 (see Table 5.4 for the *SEM* for 12-year-olds in the normal normative sample), we can say with 68% confidence that Andrea's true score lies in a range from 124 through 130, 95% confidence that it lies between 121 and 133 (1.963×3), and 99% confidence that it lies between 119 and 135 (2.583×3). The smaller the *SEM*, the more confidence one can have with the test results. Inspection of Tables 5.3 and 5.4 show that the *SEMs* for the SAGES-2 subtests are uniformly low, which supports the high degree of test reliability associated with the SAGES-2 scores.

The *SEMs* presented in Tables 5.3 and 5.4 are based on classical test theory. Classical test theory specifies the mathematical model underlying test scores and the trait being measured. With this model, the *SEM* for a given test or subtest is assumed to be the same for all ability levels. For instance, suppose Jason scored a Mathematics/Science quotient of 80. One would use the same *SEM* to create a confidence interval around his score as was done with Zach's score of 129. It is not unusual to assume, however, that the *SEM*, or the amount of error with which the examinee is being measured, differs given the ability level of the examinee. In other words, it is quite possible that Jason is being measured with more error than Zach is. Because the *SEM* may vary with ability level, it is important to assess the degree of measurement error present at the ability

Table 5.3
Standard Errors of Measurement for SAGES-2:K-3 Subtests
at Five Age Intervals (Decimals Omitted)

SAGES-2 Sample	Subtest	Age					Average
		5	6	7	8	9	
Normal	Mathematics/Science	7	6	5	5	5	5
	Language Arts/Social Studies	7	5	5	5	5	5
	Reasoning	4	4	4	4	4	4
Gifted	Mathematics/Science	5	4	4	5	5	5
	Language Arts/Social Studies	5	4	4	5	5	5
	Reasoning	5	4	4	4	4	4

Table 5.4
Standard Errors of Measurement for
SAGES-2:4-8 Subtests at Six Age Intervals
(Decimals Omitted)

SAGES-2 Sample	Subtest	Age					Average	
		9	10	11	12	13		14
Normal	Mathematics/Science	4	4	4	3	3	4	4
	Language Arts/Social Studies	5	4	4	3	3	3	4
	Reasoning	4	4	5	5	5	5	5
Gifted	Mathematics/Science	5	4	5	5	4	4	5
	Language Arts/Social Studies	5	4	4	4	4	4	4
	Reasoning	5	6	6	5	6	6	6

levels for which the test is intended. Because the SAGES-2 is intended to identify gifted students, it is important that examinees of above average ability are measured with the least amount of error.

To determine the *SEM* for above average ability levels, a second theory, item response theory (IRT), is used. IRT models have increasingly been used for test development (Hambleton & Swaminathan, 1985; Thisser& Wainer, 1990). IRT allows the *SEM* to differ with differing levels of examinee ability. Table 5.5 presents the range of the lowest *SEMs* for each subtest in the SAGES-2, as well as the corresponding ability levels for that range. It should be noted that in IRT, ability estimates are reported on a scale with a mean of 0 and a standard deviation of 1. Because all the ability estimates reported in Table 5.5 are at or above 0, examinees of above average ability levels are being measured by the SAGES-2 with the least amount of error.

To further illustrate the importance of examining *SEMs* for various ability levels, we included Figure 5.1. This figure displays the test information (solid line) and measurement error (dotted line) for the SAGES-2:4-8 Mathematics/Science subtest. The x-axis represents the scale score or ability level ($M = 0$, $SD = 1$). Two important relationships are illustrated by this figure: (1) It shows the relationship between test information (i.e., a measure of the test's effectiveness) and the ability level to be measured, and (2) it depicts the relationship between *SEM* (i.e., a measure of the test's accuracy) and the ability level to be measured. The test information curve peaks at an ability level between 1.0 and 1.3, which demonstrates that the Mathematics/Science subtest accurately measures students who perform one to two standard deviations above the mean. The *SEM*

Table 5.5
Range of Ability Levels for Which the SAGES-2
Measures with the Least Amount of Error

Level	SAGES-2 Subtest	Range of <i>SEM</i>	Range of Ability Levels (Θ)
K-3	Mathematics/Science	0.16-0.28	0.0-1.0
	Language Arts/Social Studies	0.20-0.25	0.0-1.0
	Reasoning	0.11-0.17	0.5-1.5
4-8	Mathematics/Science	0.16-0.22	1.0-2.0
	Language Arts/Social Studies	0.15-0.17	0.8-1.8
	Reasoning	0.19-0.25	1.5-2.5

curve is at its lowest point for an ability level between 1.0 and 1.3; therefore, this subtest has the least amount of error at these ability levels. This ability level is in the normal range for gifted students.

One cannot always assume that because a test is reliable for a general population, it will be equally reliable for every subgroup within that population. Therefore, those persons who build tests should demonstrate that their tests are indeed reliable for subgroups, especially those subgroups that are likely to be tested or that, because of ethnic and gender differences, might experience test bias. The alphas for five selected subgroups within the normal normative sample are reported in Tables 5.6 and 5.7 (for SAGES-2:K-3 and SAGES-2:4-8, respectively) The subgroups studied are males, females, European Americans, African Americans, and Hispanic Americans. The alphas for the gifted and normal samples were reported earlier in Tables 5.1 and 5.2. The subgroups represent a broad spectrum of “mainstream” and “minority” populations, embracing gender and ethnic categories. The large alphas in Tables 5.6 and 5.7 demonstrate that the SAGES-2 is about equally reliable for all the subgroups investigated and support the idea that the test contains little or no bias relative to those groups.

Time Sampling

Error due to time sampling refers to the extent to which a student’s test performance is constant over time and is usually estimated by the test–retest method. In this procedure, the test is given to a group of students, a period of time (generally 2 weeks or less) is allowed to pass, and the same students are tested again. Then the results of the two testings are compared. The SAGES-2’s stability over time reliability was investigated using the test–retest method. Sixty children were tested twice, with a 2-week period between testings. The examinees ranged in age from 6 through 14 and attended an elementary school (Austin, Texas) and a junior high school (Telham, Georgia). The elementary school was characterized by medium socioeconomic status and students were largely of European American ethnicity. Students in the elementary school were all identified as gifted in mathematics or language arts by the local school district. The junior high school was an alternative junior high school and was multicultural in student composition. Students in the junior high school were identified with behavioral disorders.

Raw scores for the two testings were converted into quotients to control for any effects of age in the sample. The values were then correlated, and the resulting coefficients are reported in Table 5.8, along with the means and standard deviations for each testing. The correlation coefficients in Table 5.8 were corrected for restricted and expanded range of the sample. As can be seen, these values are of sufficient magnitude to allow confidence in the test scores’ stability over time.

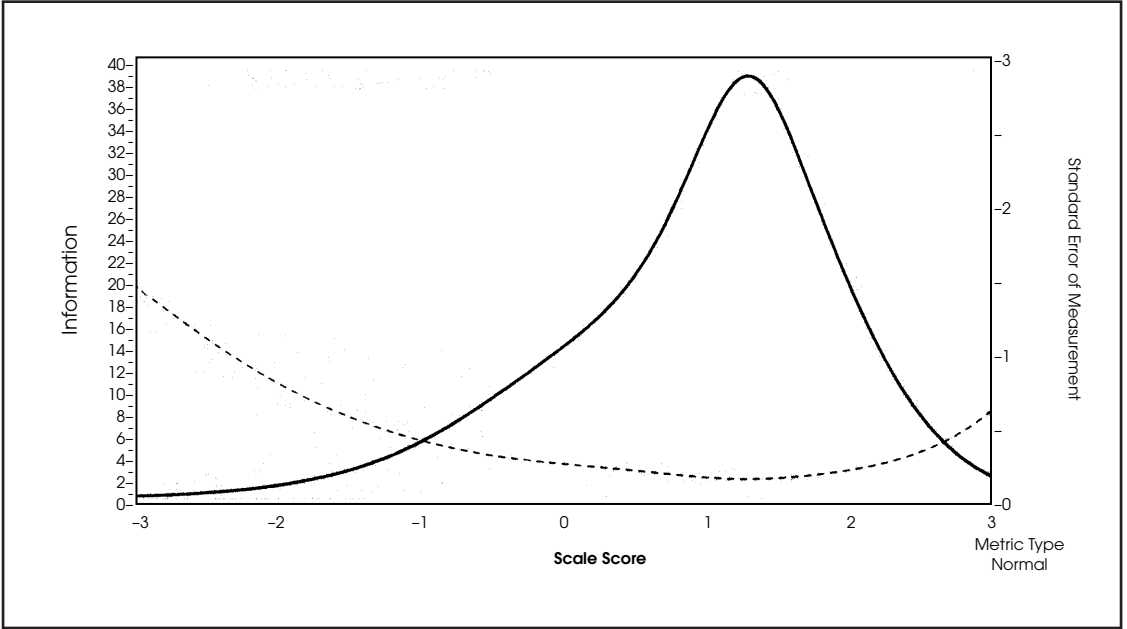


Figure 5.1. Test information for SAGES-2:4-8 Mathematics/Science subtest and measurement error.

Interscorer Differences

A third type of reliability refers to the amount of test error due to examiner variability in scoring. Unreliable scoring is usually the result of clerical error or improper application of standard scoring criteria on the part of an examiner. Scorer error can be reduced

Table 5.6
Coefficient Alphas for Selected Subgroups on SAGES-2:K-3
(Decimals Omitted)

Subtest	Subgroup				
	Male (N = 795)	Female (N = 7524)	European American (N = 1,001)	African American (N = 249)	Hispanic American (N = 263)
Mathematics/Science	93	92	92	92	92
Language Arts/Social Studies	93	94	93	94	93
Reasoning	96	96	96	96	97

Table 5.7
Coefficient Alphas for Selected Subgroups on SAGES-2:4-8
(Decimals Omitted)

Subtest	Subgroup				
	Male (N = 792)	Female (N = 684)	European American (N = 1,012)	African American (N = 225)	Hispanic American (N = 177)
Mathematics/Science	94	94	94	95	96
Language Arts/Social Studies	95	94	95	90	95
Reasoning	91	89	90	92	90

considerably by the availability of clear administration procedures, detailed guidelines governing scoring, and opportunities to practice scoring.

Nevertheless, test constructors should demonstrate statistically the amount of error in their tests due to different scorers. To do this, Anastasi and Urbina (1997) recommended that two trained individuals score a set of tests independently. The correlation between scorers is a relational index of agreement.

In the case of the SAGES-2, two staff persons in PRO-ED's research department independently scored a set of 72 completed protocols. The protocols were randomly selected from the normative sample. The sample ranged in age from 5 through 14 years. The raw scores were converted to standard scores, then correlated and reported by age intervals. The size of the resulting coefficients, listed in the Scorer column of Table 5.9, provides convincing evidence of the test's scorer reliability.

Summary of Reliability Results

The overall reliability of the SAGES-2 is summarized in Table 5.9. The content of this table shows the test's status relative to Anastasi and Urbina's (1997) three sources of test error: content, time, and scorer. The coefficients displayed are drawn from those reported in previous sections of this chapter.

As can be seen from viewing the figures listed in the table, the SAGES-2 evidences a consistently high degree of reliability across all three types. The magnitude of these coefficients strongly suggests that the SAGES-2 possesses little test error and that users can have confidence in its results.

Table 5.8
Test-Retest Reliability for the SAGES-2

SAGES-2 Level	Subtest	First Testing		Second Testing		<i>r</i>
		<i>M</i>	(<i>SD</i>)	<i>M</i>	(<i>SD</i>)	
K-3	Mathematics/Science	123	(10)	122	(11)	0.97
	Language Arts/Social Studies	123	(9)	125	(9)	0.97
	Reasoning	127	(6)	127	(6)	0.95
4-8	Mathematics/Science	109	(19)	110	(20)	0.92
	Language Arts/Social Studies	109	(21)	111	(22)	0.86
	Reasoning	98	(20)	100	(17)	0.78

Table 5.9
Summary of SAGES-2 Reliability Related to Three Sources of Test Error
(Decimals Omitted)

SAGES-2 Level	Subtest	Source of Test Error			
		Content Sampling		Time Sampling	Scorer
		Normal	Gifted		
K-3	Mathematics/Science	88	91	97	99
	Language Arts/Social Studies	87	91	97	92
	Reasoning	93	93	95	97
4-8	Mathematics/Science	94	91	92	97
	Language Arts/Social Studies	94	92	86	91
	Reasoning	90	85	78	95

