CHAPTER 4

UTAGS Reliability

Test scores are composed of two sources of variation: reliable variance and error variance. Reliable variance is the proportion of a test score that is true or consistent, while error variance is associated with that proportion of a test score that is random and not associated with the construct assessed. Only reliable variance contributes to our understanding of an examinee's performance on a test; error variance does not predict outcomes, it does not correlate with other measures, and it adds uncertainty about the meaningfulness of the examinee's test score. Because the portion of a test score that is not reliable is error, it is important for test developers to maximize a test's reliability so as to minimize error and maximize the confidence one has in the resulting test scores.

The study of test reliability centers on estimating the degree of variance that is known to be true or reliable, as opposed to the error associated with scores. When reliability is investigated, results are usually reported in terms of a coefficient that is a specific use of the common correlation coefficient and ranges from 0 (100% error) to 1.0 (i.e., perfect reliability). For instruments such as the UTAGS to be considered minimally reliable, their reliability coefficients must approximate or exceed .80 in magnitude (Bracken, 1987); coefficients of .90 or higher are considered the most desirable and most appropriate for tests designed for placement decisions (Aiken & Groth-Marnat, 2006; Bracken, 1987; Miller, Linn, & Gronlund, 2009; Reynolds, Livingston, & Willson, 2009; Salvia, Ysseldyke, & Bolt, 2010; Weiner & Craighead, 2010).

Anastasi and Urbina (1997) described three contributors to variance associated with reliability: content (i.e., internal consistency), time (stability), and scorers (interscorer). In our investigation of the UTAGS's reliability, we calculated three types of coefficients: coefficient alpha for internal consistency, test-retest correlation coefficients for stability, and interrater correlations to assess scorer consistency. Each of these types of reliability provides evidence relating to Anastasi and Urbina's (1997) sources of test error (i.e., content sampling error, content heterogeneity error, time sampling error, and scoring error).

INTERNAL CONSISTENCY

Coefficient alpha estimates internal consistency reliability and demonstrates the extent to which test items correlate with the total scale or subscale score. Alpha estimates the amount of test-item consistency associated with content sampling and content homogeneity. Coefficient alphas for

the UTAGS subscales were computed using Cronbach's (1951) method; coefficient alphas for the composite were computed by using Guilford's (1954, p. 393) formula designed for this purpose. Coefficient alphas for the UTAGS are reported in Table 6. Because the UTAGS had no significant age effects, and the normative tables are based on the entire normative sample, alphas were also computed on the entire normative sample.

Table 6 shows that the UTAGS is exceptionally reliable. For instruments used for diagnosis and placement decisions, the criterion for acceptable reliability is .90 (Bracken, 1987). As can be seen, the coefficient alphas exceed this criterion and provide strong support for the reliability of the UTAGS. Thus, examiners may place confidence in any of the subscales when interpreting UTAGS results and making diagnostic decisions.

The standard error of measurement (*SEM*) associated with coefficient alpha is an important statistic for examiners to use when interpreting scores. The *SEM* is the standard deviation of the error distribution associated with test scores. Attention to the meaning of this statistic is one way for an examiner to take into consideration the error variance that enters into the assessment situation. A test score is only an estimate of the student's true test performance; the extent to which that test score is accurate is due to reliability, and the extent to which it is inaccurate is due to test error. It is based on the formula $SEM = SD\ \sqrt{1 - r}$ (where *SD* = standard deviation; *r* = reliability), and establishes a zone within which an individual's true score probably lies. Table 6 provides the *SEM*s at three levels of confidence (i.e., 68%, 90%, and 95%) for all UTAGS scores.

The clinical value of *SEM* can be illustrated by the student's scores reported in the case study from Chapter 2. Demarcus earned a General Aptitude Index score of 128. Thus, the examiner knows with 68% confidence (i.e., +/- *SEM*) that his true score lies between 126.5 and 129.5 (128 +/- 1.5 points), with 90% confidence that the true score lies between 125.06 and 130.94 (+/- 2.94 points), and with 99% confidence that the true score lies between 121.13 and 131.87 (+/- 3.87 points). Obviously, the smaller the *SEM,* the more confidence one can have in the accuracy of the obtained test score. For practical purposes, obtained scores are rounded.

Because tests that are reliable for the general population may not be equally reliable for every subgroup within that population, evidence of test reliability must be provided for specific subgroups. The alphas for 12 selected subgroups within the normative sample are reported in Table 7. All of the alphas reported in this table exceed .90. These consistently high alphas demonstrate that the UTAGS is equally reliable for all subgroups investigated and support the premise that the test is equally accurate regardless of the examinee's gender, race, ethnicity, and exceptionality status.

STABILITY

Test-retest estimates of stability reflect the extent to which a student's test performance is consistent over time and are used to estimate time sampling error in a test. This reliability approach

involves administering the test and then readministering it generally 2 to 4 weeks later. The degree of similarity between the two test scores over time indicates the degree of stability of the construct assessed by the test.

Table 6

Coefficient Alpha and Standard Error of Measurement for UTAGS

	UTAGS Scores						
Alpha/CI	**Cognition**	**Creativity**	**Leadership**	**Literacy**	**Math**	**Science**	**General Aptitude**
α	.98	.98	.98	.98	.99	.98	.99
68% CI for *SEM*	2.12	2.12	2.12	2.12	1.50	2.12	1.50
90% CI for *SEM*	2.71	2.71	2.71	2.71	1.92	2.71	1.92
95% CI for *SEM*	3.48	3.48	3.48	3.48	2.46	3.48	2.46

Note. CI = confidence interval.

This kind of reliability was investigated for the UTAGS by having teachers rate 80 students between the ages of 5 and 17 years who were attending general public school classes in Tucson, AZ, Knoxville, TN, and Williamsburg, VA. Table 8 presents specific demographic information about the test-retest sample. After the testing was completed, the standard scores were computed for each form, and the scores for each testing were correlated. All coefficients were corrected for possible restriction or inflation of range. The resulting corrected coefficients, presented in Table 9, range from .84 to .96, demonstrating solid stability of the UTAGS results over time.

SCORER CONSISTENCY

Reliability among scorers of objective tests is understandably high. In such instances, the only possible type of error is clerical. Scorer error can be reduced considerably by the availability of clear administration procedures and detailed guidelines that govern scoring. Nevertheless, test authors should demonstrate statistically the amount of error in their test that is due to different scorers. In order to achieve this goal, authorities such as Anastasi and Urbina (1997) and Reynolds et al. (2009) recommended that two or more individuals score a set of tests independently. The mean correlation among the scorers yields an index of agreement.

In the case of the UTAGS, two members of the PRO-ED staff who were familiar with the test's scoring procedures independently scored 31 complete UTAGS protocols drawn at random from the normative sample. The results of the two scorings were then correlated and corrected for restriction of range. The resulting coefficients were all .99. This coefficient is further evidence of the UTAGS's suitable scorer reliability.

Table 7

Coefficients Alpha for Selected Subgroups on the UTAGS (Decimals Omitted)

	Subgroup											
UTAGS Score	Male (*n* = 1,283)	Female (*n* = 1,209)	White (*n* = 1,993	Black/ African American (*n* = 286)	Asian/ Pacific Islander (*n* = 122)	American/ Indian/ Eskimo/ Aleut (*n* = 31)	Hispanic (*n* = 413)	Other (*n* = 42)	Gifted and Talented (*n* = 163)	Intellectually Gifted/IQ > 120 (*n* = 27)	Learning Disabled (*n* = 56)	Intellectually Disabled/IQ < 80 (*n* = 25)
Subscale												
Cognition	98	98	98	99	97	99	99	99	98	93	99	98
Creativity	97	98	98	98	93	98	99	98	97	95	98	98
Leadership	98	98	98	97	96	98	98	98	97	95	98	97
Literacy	98	98	98	98	96	98	98	99	97	95	99	97
Math	98	98	98	99	97	99	99	99	98	95	99	99
Science	98	98	98	99	97	98	99	99	98	95	99	96
Composite												
General Aptitude	99	99	99	99	99	99	99	99	99	99	99	99

Table 8
Demographic Characteristics of the UTAGS Test-Retest Sample (N = 80)

Sample Characteristics	*n*
Gender	
Male	48
Female	32
Ethnicity	
White	62
African American	10
Asian/Pacific Islander	6
Other	2
Hispanic	
Yes	6
No	74

Table 9
Test-Retest Reliability Coefficients for UTAGS

	Time 1		Time 2			
UTAGS Score	***M***	***(SD)***	***M***	***(SD)***	r_c	(r_u)
Subscale						
Cognition	102	(15)	101	(15)	.91	(.91)
Creativity	102	(17)	102	(17)	.84	(.89)
Leadership	100	(17)	100	(15)	.85	(.87)
Literacy	101	(15)	101	(15)	.93	(.93)
Math	100	(12)	100	(13)	.96	(.93)
Science	99	(14)	100	(15)	.93	(.92)
Composite						
General Aptitude	100	(14)	101	(14)	0.96	(.95)

Note. N = 80. Coefficients corrected for range effects. *M* = mean; *SD*= standard deviation; r_c = corrected correlation of test to retest; r_u = uncorrected correlation of test to retest.

CONCLUSIONS

The overall reliability of the UTAGS is summarized in Table 10. The contents of this table show the test's status relative to three types of reliability coefficients: the coefficient alphas listed in the table are the median alphas reported at the top of Table 6; the test-retest coefficients are the corrected coefficients from Table 9; the scorer reliability coefficients were discussed in the previous section. The table also relates the three types of reliability coefficients to the different sources of possible test error described by Anastasi and Urbina (1997).

As the figures listed in the table show, the UTAGS evidences a high degree of reliability. This high level of reliability is consistent across all three types of reliability studied. The magnitude of the reported coefficients, whether at the subscale or composite level, satisfies the most demanding standards, including those of Bracken (1987), Nunnally and Bernstein (1994), Salvia et al. (2010), and Reynolds et al. (2009). These findings strongly suggest that the UTAGS possesses relatively little measurement error and that test users can have considerable confidence in the test's resulting subscale and total composite scores.

Table 10

Summary of UTAGS Reliability Relative to Three Types of Reliability (Decimals Omitted)

	Types of Reliability Coefficient		
UTAGS Score	**Coefficient Alpha**	**Test-Retest**	**Scorer**
Subscales			
Cognition	98	91	99
Creativity	98	84	99
Leadership	98	85	99
Literacy	98	93	99
Math	99	96	99
Science	98	93	99
Composite			
General Aptitude	99	96	99
Source of Test Error	Content sampling, Content heterogenity	Time Sampling	Interscorer Differences

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