CHAPTER 5

Validity of UTAGS Results

Most authors of current textbooks dealing with educational and psychological measurement—for example, Aiken and Groth-Marnat (2006), Anastasi and Urbina (1997), and Salvia et al. (2010)—encourage authors to provide evidence of at least three types of validity as a beginning point. These experts use slightly different terminology for the same concepts; we use Anastasi and Urbina's designations: content-description validity, criterion-prediction validity, and construct-identification validity. In this chapter we present data that are typically assumed to reflect these subtypes of validity, and thus begin the process of establishing validity for UTAGS.

CONTENT-DESCRIPTION VALIDITY

Content-description validity involves "the systematic examination of the assessment's content to determine whether it covers a representative sample of the behavior domain to be measured" (Anastasi & Urbina, 1997, p. 115). This kind of validity has to be built into the assessment at the time it is conceptualized and items are written. Test authors usually prove content validity by showing that the abilities chosen to be measured are consistent with the current knowledge about a particular area. They also demonstrate that the items hold up statistically.

Three aspects of content-description validity are addressed by responding to the following questions for the UTAGS items. First, does the literature support the value and accuracy of teacher ratings? Second, are the UTAGS items appropriate? Third, do results of classical item analyses support the validity of the UTAGS?

The Value of Teacher Judgments/Ratings

Because the UTAGS requires teacher ratings, discussion regarding the value of such ratings is appropriate. Frequently teachers are asked to rate their students' performance on a variety of behavioral, curricular, and learning attributes. Such ratings are valuable because they operationalize qualitative information based on direct observation by an experienced, professional who works on a regular basis with the person being rated. And in the case of the UTAGS, these ratings are even more valuable because they reflect performance of the examinee relative to performance of same-age peers. Consequently, the ratings and the resulting subscale scores are sensitive to localized expectations, or norms—a significant strength of this instrument. A large and growing body of research testifies to the

value of teacher ratings. An early representative example is the work of Morine-Dershimer (1978–1979a, 1978–1979b). According to data from these sources, teacher estimates of student ability can be accurate and highly useful. These findings are not surprising because teachers have a well-developed knowledge of the day-to-day behaviors of their students, and they develop a thorough understanding of students' general and specific abilities. Wiederholt and Bryant (1987) noted that experienced teachers appear to have an internalized scope-and-sequence chart that allows them to evaluate the learning abilities of students after a relatively short period of observation time. Teachers' perceptions can be translated into a more permanent form using items included on rating scales, as well as through written opinions and grades.

More current research has substantiated the accuracy of teacher judgments (e.g., Crosby & French, 2002; Gresham, MacMillan, & Bocian, 1997; Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Nelson, 1971; Perry & Meisels, 1996; Schafer, 1982). Some of the most compelling data have documented agreement between teacher ratings and the results of standardized academic achievement tests (Brophy & Good, 1974; Hammill & Hresko, 1994; Keogh & Smith, 1970; Meisels et al., 2001). Given the prominence of this research, Egan and Archer (1985) stated that teachers' ratings of student achievement have not been given significant credence:

> There is no compelling evidence that teachers' ratings are in fact inaccurate. Since the 1920s, there have been dozens of studies reporting correlations on the order of .50 to .60 between ratings and various standardized tests. These correlations may be considered as coefficients of concurrent validity, and as such they are quite large. (p. 26)

Finally, the accuracy of using teacher judgments has been verified by Gresham et al. (1997), who studied the use of judgments in differentiating among various groups of students referred to School Study Teams for evaluation. They found that the teachers correctly identified 91% of the learning disabilities group, 100% of the low-IQ group, and 95% of the low-achievement group. Other investigators noted that teachers use ratings to make important educational or instructional decisions, such as how to group students for instruction, whether students are responding appropriately to instructional material, and whether instructional groups should be changed (Borko, Cone, Russo, & Shavelson, 1979; Clark & Peterson, 1986; Harlen, 2005; Perry & Meisels, 1996; Peterson, 1988). Still other researchers emphasized the importance of teacher ratings in both general and special education settings (Beswick, Williams, & Sloat, 2005; Elliot, Gresham, Freeman, & McCloskey, 1988; Gerber & Semmel, 1984; Hoge, 1983; Ysseldyke, 1979).

According to Ysseldyke (1979), assessment occurs in all important aspects of education that require decision making, including (a) student screening and evaluation and (b) program evaluation. Any of these decision points should include some type of objective teacher input (Beringer, Stage, Smith, & Hildebrand, 2001). Ysseldyke and Algozzine (1982) reinforced the point that teacher evaluations and judgments are critical to determining a student's eligibility for special education services.

Given this literature, educators should ensure inclusion of teachers' observations of their students' performance, aptitudes, and interest when their students are under consideration for receipt of services for at-risk status (i.e., their needs cannot be met optimally within the "regular" classroom environment). These students may have an intellectual disability, one or more learning disabilities, or they may be gifted. Ultimately, judgments about these students' academic and related performance by education professionals who know them well should become a part of the decision-making process. Rating scales such as the UTAGS provide an efficient mechanism for obtaining this information.

Is the Theoretical Basis for UTAGS Items Sufficient?

The UTAGS is a behavior rating scale that addresses a broad range of behaviors across a number of aptitudes. Each item is designed to assess a unique behavior (or class of behaviors) and, in their aggregate form, to comprise a subscale, which reflects an operationalization of a larger construct (e.g., creativity, math aptitude). The format of the UTAGS requires teachers to rate a student's behavior based on their knowledge of the student's ability to perform specific behaviors. Using teachers to evaluate the performance of their students has a long history, and this rating format can yield strong psychometric characteristics, as mentioned previously (Bracken & Brown, 2008; Pfeiffer, & Jarosewich, 2003; Renzulli et al., 2013).

Authors of the UTAGS reviewed a broad literature base in order to identify behaviors relevant to areas that require cognitive, academic, creative and leadership skills. The relevant literature included research articles, textbooks, curricular guides, diagnostic tests, criterion-referenced tests, and theoretical papers. Specifically, the goal was to identify behaviors in the literature that teachers would consider authentic expressions of the UTAGS general constructs (i.e., cognition, creativity, and leadership) and UTAGS specific academic constructs (i.e., literacy, math, and science). That is, the behaviors referenced by the items were considered to be representative and frequently occurring (within-the-classroom) indicators of UTAGS constructs, as clearly defined and identified as critical by experts (e.g., Eckert, Dunn, Codding, Begeny, & Kleinmann, 2006).

Consistent with the recommendations of Shavelson, Cadwell, and Izu (1977); Shavelson and Stern (1981); and Perry and Meisels (1996), the authors provide raters with a reference group with which to compare a student's performance (i.e., average-performing students of like chronological ages within the local environment). This strategy provides raters with an anchor for their ratings and, more importantly, a point of reference that allows students to be compared with their local peers. This strategy creates a set of "built-in" norms and recognizes that behaviors considered gifted in one location might not be considered so in another.

UTAGS items were culled from the literature, as well as from the authors' experiences as practicing psychologists and teachers of psychologists, and from educators at all grade levels. Currently, each subscale contains 15 items; the original item pool was larger, ranging from 21 to 27 items per subscale. Items were eliminated from consideration based on preliminary item and reliability analyses, leaving the 15-item subscales that were used in norming the UTAGS. Items were selected in part because they reference behaviors characterized by educators as being highly desirable for success in U.S. public schools and the workforce. Students who are rated highly on these subscales demonstrate exceptional potential for academic success and for providing significant contributions to society; those who score poorly will likely be comparatively less successful in school and in the workplace.

Because the same sets of items are used across the age range from 5 to 17 years, examiners must consider how each item applies for students who are the age of the particular student being rated; that is, developmental considerations and expectations are essential for accurate assessment of students' behaviors. For example, on the Literacy subscale the first item is, "expresses ideas cogently in writing." Obviously, "average" performance in written expression varies enormously as a function of chronological age and cognitive sophistication. In other words, what is competent written expression for a 5-year-old differs considerably from that of an 8- or 17-year-old. Consider the following sentence, written by a 5-year-old student at midyear, "I like to crunch carrots." Most 5-year-old students are beginning to understand the structure and syntax of written language but still have trouble with grammatical nuances and spelling. This student not only understood the structure, but also spelled

every word correctly. A teacher of a 5-year-old might consider the sentence *above average*, or even *well above average* for a kindergarten student in a particular school, compared to his or her peers. But the same teacher might consider the sentence average by year's end. Also, a teacher at another school might consider the sentence to be average or even below average in a school of very high achievers. And, of course, elevated performance on this item would be very different for a middle school student. Middle school teachers expect their students to write and meaningfully connect lengthy passages, use topic sentences appropriately, develop ideas in an interesting manner, consider the background of the reading audience, and so on. In order to provide useful ratings, teachers must be developmentally aware of the typical performance of children of various ages within their local environment, and be able to characterize students' performance on behaviors that operationalize the item for same-age peers.

Finally, UTAGS authors created directions for raters that request sensitivity to performance without regard to the particular language or communication medium used by the examinee. The goal for raters is to focus exclusively on the ability of the examinee to communicate the skill(s) described by a particular item, not on the mechanism or vehicle used to communicate. Thus, those with limited English language skills should not be penalized just because their English is limited. Communication may be effective using sign, gestures, a second language, pantomime, and so on, and should be evaluated accordingly.

Evidence From Conventional Item Analysis

The previous sections provided qualitative evidence of the content-description validity of the UTAGS. In this section, quantitative evidence for this type of validity is presented. For example, results of traditional, time-tested procedures were used to select good (i.e., valid) items for a test. These procedures focus on the consideration 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 discrimination power or item validity. Nunnally and Bernstein (1994) noted that items with a discriminating power of .20 or more will likely be satisfactory but that higher values are preferred. As evidenced in Table 11, the median discriminating power for UTAGS subscales greatly exceed .20 and range from .81 to .92. These results show that the items of the UTAGS demonstrate highly desirable levels of discriminating power.

CRITERION-PREDICTION VALIDITY

Anastasi and Urbina (1997) described criterion-prediction validity as "the effectiveness of a test in predicting an individual's performance in specified activities" (p. 118). According to this definition, performance on an assessment is checked against a criterion that can be either a direct or indirect measure of what the assessment is designed to predict. To establish criterion-prediction validity of the UTAGS, its (a) scores were correlated with those of a criterion measure, (b) means and standard deviations were compared with those of the criterion measure, and (c) sensitivity and specificity, and ROC Area Under the Curve, were investigated through binary classification analyses.

Table 11
Median Discrimination Powers for UTAGS at 13 Age Intervals (Decimals Omitted)

Age (in years)	Cognition	Creativity	Leadership	Literacy	Math	Science
5	87	86	86	83	91	89
6	89	88	82	86	87	88
7	82	84	81	84	87	85
8	86	85	83	85	87	86
9	91	89	88	91	93	91
10	90	86	83	88	91	90
11	88	83	81	86	90	88
12	88	82	83	83	90	89
13	90	81	83	88	89	89
14	90	90	91	90	91	89
15	91	86	92	91	91	89
16	86	89	87	86	92	86
17	84	89	90	88	90	89

Criterion-prediction data were collected on 105 White, non-Hispanic students (57 males, 48 females) from Tennessee ranging in age from 8 to 13 years. For this study, the UTAGS and the *Gifted Rating Scales* (GRS; Pfeiffer & Jarosewich, 2003) were administered concurrently. Data were also collected from the students' records of their performance on the *Tennessee Comprehensive Assessment Program* (TCAP; Tennessee Department of Education, 2011).

Correlations With Criterion Measures

In this investigation of criterion-prediction validity, the UTAGS was correlated with the Intellect, Creativity, and Leadership Scales of the GRS and the Mathematics, Reading/Language Arts, and the Science tests from the TCAP. The resulting correlations between the UTAGS and the criterion measures answer a practical question: Does the UTAGS evaluate behaviors and aptitudes typically associated with cognitive ability? Because the question is practical (instead of theoretical), one should attenuate the coefficients for any lack of reliability in the criterion tests (but not in the UTAGS) and correct coefficients to account for any range effects that might artificially depress or inflate the coefficients.

Hopkins's (2002) criteria are accepted *a priori* to evaluate the resulting coefficients: 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. One would expect that the correlations between the UTAGS and the criterion test would be Large or Very Large.

The results of this study are reported in Table 12. The magnitude of the average corrected coefficients for the General Aptitude Index ranges from Large to Very Large (*r*'s ranging from .67 to .72). These magnitudes suggest that the UTAGS correlates well with the other measures of aptitude and provides compelling evidence for the UTAGS's criterion-prediction validity.

Table 12

Correlations Between UTAGS and Selected Criterion Test Scores (Decimals Omitted)

	GRS Score						TCAP Score					
	Creativity		Intellectual Ability		Leadership Ability		Mathematics		Reading/ Language Arts		Science	
UTAGS Score	(r_u)	r_c	(r_u)	r_c	(r_u)	r_c	(r_u)	r_c	(r_u)	r_c	(r_u)	r_c
Subscale												
Cognition	(64)	66	(73)	68	(53)	54	(59)	68	(63)	71	(59)	66
Creativity	(52)	52	(54)	51	(43)	43	(44)	56	(48)	59	(44)	53
Leadership	(35)	35	(41)	39	(65)	66	(19)	25	(22)	29	(20)	25
Literacy	(63)	63	(70)	66	(57)	58	(51)	61	(60)	70	(52)	60
Math	(66)	66	(72)	64	(54)	55	(60)	66	(61)	67	(59)	63
Science	(61)	61	(70)	59	(51)	51	(50)	54	(60)	63	(56)	57
Composite												
General Aptitude	(70)	70	(75)	72	(65)	65	(56)	67	(62)	72	(57)	65
Magnitude[1]	Very Large		Very Large		Large		Large		Very Large		Large	

Note. TCAP = *Tennessee Comprehensive Assessment Program* (Tennessee Department of Education, 2011). N = 105; Coefficients corrected for range effects and attenuation; r_u = uncorrected correlation coefficient, r_c = corrected correlation coefficient.

[1]Magnitude of the corrected General Aptitude correlation coefficient according to Hopkins's (2002) criteria.

Data from an independent study conducted after the original standardization/validity studies compared UTAGS scores to other measures of important school-based success (Kirkpatrick, McCallum, Bell, & Bracken, 2018); these results provide additional evidence of construct/concurrent validity and show relations between UTAGS and other important operationalizations of school success: critical thinking as assessed by The Test of Critical Thinking (TCT; Bracken et al., 2003); academic success as characterized by North Carolina End-of-Grade Tests (NCEOG; North Carolina Department of Education, 2017); and emotional intelligence (EI) as assessed by the Emotional Quotient Inventory: Youth Version (EQ-i:YV; Bar-on & Parker, 2003. Participants included 63 students, grades 3 through 5, from a public school in North Carolina. Special education students were included to the extent they comprised the student population. Means from the six UTAGS subscales range from 98.63 to 103.20; standard deviations range from 15.17 to 18.80. As is apparent from the UTAGS scores from this sample, their means are about average relative to the UTAGS population mean of 100. All of the correlation coefficients depicting the relation between UTAGS and critical thinking (TCT) are statistically significant ($p \leq .01$), and range from .37 (Leadership) to .66 (Literacy), with most above .46. These scores affirm the expected and moderate to strong relations between abilities assessed by the UTAGS and critical thinking skills. Correlation coefficients between UTAGS and end-of-year (NCEOG) high-stakes reading scores from the range from .26 (Creativity and Leadership) to .61 (Literacy), with most above .50; coefficients between UTAGS and end-of-year high-stakes math scores range from .48 (Creativity and Leadership) to .71 (Literacy), with most above .57. All coefficients are statistically

significant ($p \leq$ either .05 or .01; 12 of the 14 coefficients are significant at .01). Coefficients between UTAGS and the measure of EI were modest, although three were statistically significant. Apparently the relations between EI and the constructs assessed by UTAGS are somewhat independent. However, it is important to note that teachers provided the ratings on the UTAGS, and the ratings from the EI measure are self-reported (from students). Consequently, the relations are moderated by method variance differences. Finally, these data allow a mean-difference comparison between gifted and nongifted students and between at-risk special education students, those with learning or intellectual disabilities. The means were significantly different in the predicted directions; the mean of 117.91 for the gifted students is significantly higher than the mean of 96.49 for the nongifted peers, $t(62) = 4.55$, $p \leq .01$, Similarly, the mean of 82 for the students within the special education sample is significantly lower than the mean (103.94) obtained for their non-special education peers, $t(62) = -4.70$, $p \leq .01$. Importantly, because of regression to the mean, scores for the gifted students are lower than might be expected, and the means for the at-risk special education students are higher than might be expected. In general, results from this study indicate strong associations between UTAGS scores and important measures of school success.

Comparison of Means and Standard Deviations for UTAGS and Criterion Tests

When two tests are highly correlated, it usually means that they are likely to be measuring the same—or a similar—ability. However, high correlations do not necessarily mean that the tests will yield the same results. For example, one test may consistently score higher than another test even though they correlate well with each other. The validity of both tests is supported when the two tests produce similar means as well as correlate highly with each other.

The standard score means, standard deviations, and comparative information for UTAGS, GRS, and the TCAP are presented in Table 13. The descriptive terms used to describe the means are listed in Table 1. The differences between the UTAGS scores and the scores from the GRS and TCAP were analyzed using the *t* test results (Guilford & Fruchter, 1978) and effect size correlations (Dunlop, Cortina, Vaslow, & Burke, 1996; Hopkins, 2002).

The findings in the table demonstrate there are no meaningful differences between standard scores of the UTAGS and standard scores of the GRS, while the TCAP scores tended to be consistently higher than those from the UTAGS. The magnitude of the differences range from Very Small to Moderate. These summative statistics support the conclusion that the standard scores that result from giving the UTAGS likely will be similar to those obtained from administration of other related assessments of aptitude but may underestimate end-of-year measures of academic achievement.

Binary Classification and ROC Area Under the Curve (ROC/AUC)

Binary classification analysis involves the computation of a test's sensitivity and specificity indexes. In the current context, the *sensitivity index* reflects the ability of the UTAGS to correctly identify students demonstrating behaviors associated with high or low cognitive aptitude—the most important attribute of a screening scale. The *specificity index* reflects the ability of the UTAGS to correctly identify students who do not demonstrate behaviors associated with high or low cognitive aptitude.

Table 13

Comparison of the UTAGS Means (and Standard Deviations) With Those of Criterion Tests

Comparison	*n*	Mean (*SD*)	Descriptive Term	*t*[1]	Effect Size Correlation[2]	Magnitude[3]
UTAGS Cognitive	105	108 (15)	Average	-0.88 ns	0.09	Very Small
GRS Intellectual Ability Scale		109 (17)	Average			
UTAGS Creativity	105	104 (14)	Average	-3.42 **	0.32	Moderate
GRS Creativity Scale		108 (15)	Average			
UTAGS Leadership	105	105 (14)	Average	0.00 ns	0.00	Very Small
GRS Leadership Ability Scale		105 (14)	Average			
UTAGS Literacy	105	108 (14)	Average	-5.29 **	0.46	Moderate
TCAP Reading/ Language Arts		114 (12)	Above Average			
UTAGS Math	105	108 (16)	Average	-5.57 **	0.48	Moderate
TCAP Mathematics		115 (12)	Above Average			
UTAGS Science	105	108 (17)	Average	-2.99 *	0.28	Small
TCAP Science		112 (13)	Above Average			

Note. GRS = *Gifted Rating Scales* (Pfeiffer & Jarosewich, 2003); TCAP = *Tennessee Comprehensive Assessment Program* (Tennessee Department of Education, 2011). [1]Values of *t* were computed by the dependent samples method (Guilford & Fruchter, 1978). [2]Effect size was calculated using Dunlop et al.'s (1996) formula #3, which corrects for inflated effect size due to correlated design t-tests. [3]Values of magnitude of the effect size correlation between the UTAGS and the criterion measures according to Hopkins's (2002) criteria.

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

The results for sensitivity and specificity are reported as proportions (i.e., percentages). The size of the proportions necessary to be considered acceptable will vary 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 assure that the number of true positives is high). Authorities vary regarding how large a test's sensitivity and specificity indexes should be. Wood, Flowers, Meyer, and Hill (2002) recommended that the 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) recommended a more rigorous standard of .80 or higher. Others (Jenkins, 2003; Johnson, Jenkins, Petscher, & Catts, 2009) recommended that the sensitivities should be higher—perhaps as high as .90—and that specificity levels should be relatively high as well.

The receiver operating characteristic area under the curve (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). ROC/AUC values range from 0 (representing zero predictive ability) to 1 (representing perfect predictive ability). Compton, Fuchs, Fuchs, and Bryant (2006) suggested that ROC/AUCs of .90 and above are considered excellent; .80–.89 are good; .70–.79 are fair; and .69 or below are poor.

Because one purpose of the UTAGS is to screen for cognitive aptitude, test users must have confidence that any criterion measure used to investigate its validity accurately reflects this construct. With the sample of 105 students described in the earlier criterion-identification studies plus 52 additional students with known IQs, we estimated the UTAGS's ability to discriminate students with high cognitive ability from those with low to average cognitive ability (i.e., to identify the gifted and talented students from all others) and to discriminate students with low cognitive ability from those with average to high cognitive ability (i.e., to identify intellectually disabled students from all others). To do this, we explored the utility of three possible cut scores to predict membership in the high cognitive and low cognitive ability groups: UTAGS General Aptitude Index of 110, 115, and 120, and General Aptitude Index scores of 90, 80, and 70.

Table 14 reports the sensitivity, specificity, ROC/AUC, and related statistics for the UTAGS General Aptitude Index at the various cut scores. All but one of the cut scores examined satisfied the highest standards recommended by the authorities mentioned earlier in this section, and thereby provide strong evidence of criterion-prediction validity for the UTAGS. Only the sensitivity index reported for the highest cut score (120) fell short of the minimum criterion of .70 for sensitivity. The results of this analysis suggest that the UTAGS cut score of 110 and 70 resulted in the fewest false positives and false negatives when predicting high and low cognitive ability, respectively. Therefore, we recommend using these cut scores when screening students. If the goal is to identify gifted examinees, rather than screen, a higher score may be required, as determined by a particular school system or agency (e.g., 130, which corresponds to two standard deviations above the mean and represents the standard sometimes set by experts in the gifted literature). Of course, school personnel may choose to use other UTAGS subscale scores independently for screening as well, rather than the overall General Aptitude Index.

CONSTRUCT-IDENTIFICATION VALIDITY

Construct-identification validity addresses the theoretical framework on which a test is built by examining the relationship of test performance (i.e., scores) to the hypothetical and more global constructs that underlie or explain the test performance. Evidence for construct validity of an instrument is shown by delineating as fully as possible the variable (construct) that the instrument purports to measure. Evidence is obtained by formulating hypotheses about scores on the instrument in light of what is known about the constructs it assesses, then observing whether or not the scores fit the predicted pattern. The hypotheses are then accepted or rejected on the basis of the results. The following hypotheses were tested for the UTAGS:

1. Because UTAGS examinees are rated relative to same-age peers, item data obtained from norming should not correlate significantly with chronological age.
2. Because all of the UTAGS subscales were designed to measure behaviors undergirding several general aptitudes associated with school success (e.g., intellectual, academic, and social), they should be significantly intercorrelated, but only moderately.

Table 14
Binary Classification and ROC/AUC Curve Analyses

Criterion/ Cut Score	Sensitivity Index	Specificity Index	ROC/AUC	True Positives	False Positives	True Negatives	False Negatives	Classification Accuracy
High Cognitive Ability								
110	.81	.94	.983	65	3	74	15	.89
115	.75	100	.983	60	0	77	20	.87
120	.51	100	.983	41	0	77	39	.75
Low Cognitive Ability								
90	100	100	.986	18	18	121	0	.89
80	.94	.94	.986	17	8	131	1	.94
70	.83	.97	.986	15	4	135	3	.96

Note. N = 157. ROC/AUC = receiver operating characteristic/area under the curve.

3. Because the UTAGS assesses behaviors associated with cognitive ability, its results should differentiate between groups of students identified as gifted and those who are not and between groups of students identified as having intellectual disabilities and those who are not.
4. Because the UTAGS assesses behaviors associated with overall school success, the General Aptitude Composite is assumed to best reflect overall success; exploratory and confirmatory factor analyses should demonstrate that the items and subscales load significantly on one factor.

Relationship to Age

Because UTAGS raters use the examinee's same-age peers as the comparison group, one would expect the UTAGS subscales to have approximately equal raw score means across age and not be significantly related to age. The raw subscale means from the UTAGS standardization sample at 13 age intervals are listed in Table 15. Coefficients showing the relationship of age to performance on the subscales are also provided. This table clearly demonstrates that UTAGS performance is not significantly related to age.

Relationships Among the Subscales

If the UTAGS subscales do in fact measure some important aspects of school success, one could reasonably expect that they would correlate with each other to some degree. The subscales should not, however, correlate too highly with each other. For example, if two subscales correlate .90 or higher with each other, the percent of shared variance between the two equals 81. In this case, the two subscales may be considered so redundant that administration of both is inefficient. One the other hand, if they correlated .30–.70, both would be assumed to contribute unique variance to the battery's total score.

To investigate this hypothesis, we calculated the intercorrelations among the subscales using the entire normative sample as subjects. The resulting coefficients for the subscales are reported in Table 16. The median correlation in both tables is .76 (the range is .62 to .88). The correlations among the subscales are high enough to support the idea that they are measuring aspects of the same ability (i.e., school success) and small enough that they all make unique contributions to the rating scale. This finding provides additional support for the UTAGS's construct-identification validity.

Table 15
UTAGS Subscale Raw Score Means and Standard Deviations at 13 Age Intervals

Age (in years)	Cognition	Creativity	Leadership	Literacy	Math	Science
5	46 (11)	47 (10)	48 (11)	46 (11)	45 (11)	46 (10)
6	45 (12)	45 (11)	47 (11)	46 (12)	45 (11)	44 (10)
7	48 (11)	48 (10)	49 (11)	49 (12)	48 (11)	47 (10)
8	47 (11)	47 (10)	48 (11)	48 (11)	46 (10)	46 (9)
9	47 (15)	46 (12)	48 (14)	47 (15)	46 (14)	47 (13)
10	48 (13)	47 (10)	49 (11)	48 (12)	48 (12)	47 (11)
11	47 (12)	46 (10)	49 (10)	47 (12)	47 (11)	46 (10)
12	46 (12)	46 (10)	49 (11)	47 (12)	46 (12)	47 (12)
13	48 (13)	47 (10)	49 (11)	48 (12)	49 (12)	48 (12)
14	43 (14)	43 (14)	46 (15)	44 (15)	44 (13)	43 (13)
15	44 (13)	44 (11)	46 (14)	47 (13)	44 (11)	46 (11)
16	46 (12)	46 (12)	48 (13)	47 (14)	45 (12)	45 (12)
17	43 (11)	46 (12)	49 (14)	45 (13)	45 (12)	44 (11)
Correlation With Age	-.031	-.037	.001	-.012	-.004	.009
Magnitude[1]	Very Small	Very Small	Very Small	Very Small	Very Small	Very Small

[1]Magnitude of the correlation coefficient based on Hopkins's (2002) criteria.

Table 16
Intercorrelations of UTAGS Subscales for Entire Normative Sample (Decimals Omitted)

	Cognition	Creativity	Leadership	Literacy	Math	Science
Cognition	—	77	68	88	86	83
Creativity		—	62	76	71	76
Leadership			—	68	62	62
Literacy				—	85	81
Math					—	84
Science						—

Differences Between Groups

One way of establishing an instrument's validity is to study the performance of different groups of people on the test. Each group's results should make sense, given what is known about the relationship

of the test's content to the group. In the case of the UTAGS, an assessment of behaviors and aptitudes associated with intellectual, academic, and social success, one would expect that students identified as intellectually disabled (i.e., IQ < 80) would perform worse than those identified as average or gifted and talented or intellectually gifted (i.e., IQ > 120). Similarly, gifted and talented students should perform better than nongifted students and students with disabilities.

The mean standard scores for the entire normative sample, as well as for selected subgroups, are listed in Table 17. Data are presented for three "mainstream" subgroups (males, females, and Whites), five minority subgroups (Black/African Americans, Hispanics, Asian Americans, Native Americans, and Two or More/Other), and four "exceptionality" subgroups (gifted and talented, intellectually gifted [student with known IQs > 120], learning disabled, and intellectually disabled [students with known IQs < 80]).

The means in Table 17 support the construct-identification validity of the UTAGS. As expected, General Aptitude Index scores obtained by students with intellectual disabilities are well below average, and those from the gifted and talented subgroups are above average; General Aptitude Index scores for almost all other subgroups are in the average range, including those with learning disabilities, as expected. The only other subgroup to score above average was the Asian American subgroup, a finding consistent with the findings of other cognitive measures (e.g., the *Universal Nonverbal Intelligence Test*, Bracken & McCallum, 2016). Because mean scores for all subgroups are as expected, one can conclude that the UTAGS is a fair assessment for students who are identified by schools as deserving of special attention (e.g., those with intellectual and learning disabilities, those who are gifted, or those who are from culturally or linguistically different backgrounds).

Factor Analysis

Construct-identification validity also relates to the degree to which the underlying traits of a test can be identified and the extent to which those traits reflect the theoretical model on which the test is based. To empirically investigate the validity of the UTAGS model, both exploratory factor analyses (EFA) and confirmatory factor analyses (CFA) were conducted using individuals in the normative sample as subjects.

Exploratory Factor Analysis. One way to examine construct-identification validity is to analyze UTAGS subscale performance using EFA. In this analysis, we used the principal components of EFA with a promax rotation to analyze the data from the entire normative sample. We analyzed the data both at the item level and at the subscale level.

The item-level EFA is presented in Table 18. This analysis revealed both a strong one-factor solution and a strong six-factor solution. The one-factor item-level solution yielded an eigenvalue of 57.32, and this single factor accounted for 64% of the variance. The six-factor item-level solution yielded eigenvalues ranging from 57.32 to 1.47 and accounted for 80% of the variance. The subscale-level EFA is reported in Table 19. This analysis revealed a single eigenvalue of 4.78 and also accounted for 80% of the variance. The findings reported here provide strong evidence for the structure of the UTAGS and further strengthen the construct-identification validity of these rating scales.

Table 17

UTAGS Subscale and Composite Index Standard Score Means for the Total Sample and Selected Subgroups

	Subgroup												
UTAGS values	**Total Sample (*n* = 2,492)**	**Male (*n* = 1,283)**	**Female (*n* = 1,209)**	**White (*n* = 1,993)**	**Black/ African American (*n* = 286)**	**Hispanic (*n* = 413)**	**Asian American/ Pacific Islander (*n* = 122)**	**Native American (*n* = 31)**	**Two or More/ Other (*n* = 60)**	**Gifted and Talented (*n* = 163)**	**IQ > 120 (*n* = 27)**	**Learning Disabled (*n* = 56)**	**IQ < 80 (*n* = 25)**
Subscale													
Cognition	100	101	100	101	95	94	111	99	104	116	121	93	68
Creativity	100	100	101	101	98	94	106	96	104	112	124	94	66
Leadership	101	100	102	102	97	98	108	100	107	111	114	95	73
Literacy	100	100	101	101	96	93	110	98	103	115	119	91	67
Math	100	101	100	100	96	94	112	98	104	116	123	94	65
Science	100	101	100	100	95	94	112	96	105	116	123	95	63
Composite Index													
General Aptitude	100	100	100	100	95	93	110	97	105	115	122	92	62

Table 18

Item-Level Unrotated and Rotated Factor Pattern for UTAGS Using the Entire Normative Sample

	Unrotated Factor Pattern	Rotated Factor Pattern (Standardized Regression Coefficients)					
Statistics/Subscale	**Factor 1**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Eigenvalue	57.32	57.32	5.40	3.79	2.42	1.76	1.47
Percent Variance Accounted for	64	64	6	4	3	2	2
Subscale							
Leadership Item 1	0.61	0.92	-0.08	-0.02	-0.03	-0.01	0.07
Leadership Item 2	0.64	0.90	-0.03	0.04	-0.02	-0.07	0.05
Leadership Item 3	0.61	0.89	-0.06	-0.02	0.03	-0.03	0.03
Leadership Item 4	0.64	0.89	-0.12	0.01	0.05	0.03	0.03
Leadership Item 5	0.64	0.92	-0.03	-0.07	0.03	0.00	0.02
Leadership Item 6	0.65	0.83	-0.07	0.00	0.08	0.01	0.03
Leadership Item 7	0.67	0.84	0.00	0.00	0.08	-0.02	0.01
Leadership Item 8	0.66	0.88	0.03	-0.02	0.02	-0.04	0.04
Leadership Item 9	0.76	0.68	0.11	0.12	0.01	0.07	0.00
Leadership Item 10	0.72	0.78	0.12	0.07	0.02	-0.02	-0.01
Leadership Item 11	0.68	0.79	0.12	0.01	0.00	0.01	-0.02
Leadership Item 12	0.74	0.73	0.10	0.09	0.03	0.06	-0.05
Leadership Item 13	0.72	0.74	0.11	0.06	0.03	0.07	-0.05
Leadership Item 14	0.74	0.70	0.11	0.08	0.00	0.11	-0.04
Leadership Item 15	0.74	0.71	0.10	0.03	-0.05	0.19	-0.01
Math Item 1	0.83	0.03	0.79	0.00	0.07	0.03	0.06
Math Item 2	0.83	0.02	0.81	0.02	0.02	0.06	0.04
Math Item 3	0.81	0.02	0.83	-0.01	0.01	0.04	0.06
Math Item 4	0.85	0.07	0.77	0.03	0.05	0.03	0.05
Math Item 5	0.84	0.06	0.81	0.01	0.06	0.02	0.02
Math Item 6	0.84	0.04	0.79	0.02	-0.03	0.12	0.03
Math Item 7	0.85	0.01	0.82	0.03	0.04	0.07	0.01
Math Item 8	0.85	0.04	0.81	0.00	0.07	0.08	0.00
Math Item 9	0.84	0.02	0.69	0.07	0.12	0.05	0.04
Math Item 10	0.83	0.00	0.70	0.03	0.12	0.02	0.10
Math Item 11	0.86	-0.02	0.70	0.08	0.13	0.07	0.04
Math Item 12	0.85	0.01	0.73	0.03	0.17	0.02	0.04
Math Item 13	0.86	0.01	0.73	0.07	0.15	0.04	0.00
Math Item 14	0.81	-0.03	0.64	0.02	0.24	0.03	0.05
Math Item 15	0.80	0.00	0.61	-0.01	0.28	0.02	0.04
Creativity Item 1	0.77	0.00	0.03	0.60	0.12	0.05	0.15
Creativity Item 2	0.78	-0.04	0.00	0.61	0.14	0.06	0.17

Table 18, continued.

	Unrotated Factor Pattern	Rotated Factor Pattern (Standardized Regression Coefficients)					
Statistics/Subscale	**Factor 1**	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Creativity Item 3	0.73	0.06	0.10	0.66	0.08	0.00	0.01
Creativity Item 4	0.71	0.01	0.01	0.85	0.00	0.05	-0.03
Creativity Item 5	0.71	0.01	-0.06	0.70	0.15	-0.09	0.18
Creativity Item 6	0.69	-0.05	0.00	0.61	0.26	-0.14	0.16
Creativity Item 7	0.74	0.00	0.02	0.85	-0.04	0.12	-0.03
Creativity Item 8	0.69	0.06	-0.04	0.88	-0.01	0.07	-0.09
Creativity Item 9	0.74	0.03	0.00	0.90	-0.04	0.05	-0.02
Creativity Item 10	0.77	0.03	0.00	0.88	0.01	0.03	0.01
Creativity Item 11	0.75	0.01	0.03	0.88	-0.01	0.08	-0.05
Creativity Item 12	0.77	0.03	0.06	0.86	0.05	0.03	-0.06
Creativity Item 13	0.80	0.02	0.09	0.70	0.10	0.00	0.08
Creativity Item 14	0.78	0.05	0.05	0.80	-0.03	0.03	0.06
Creativity Item 15	0.77	0.05	0.06	0.76	0.08	-0.03	0.04
Science Item 1	0.80	0.05	0.03	0.07	0.77	-0.04	0.07
Science Item 2	0.83	0.05	0.10	-0.01	0.76	0.04	0.05
Science Item 3	0.80	0.06	-0.01	0.09	0.78	-0.01	0.05
Science Item 4	0.86	0.05	0.15	0.02	0.67	0.08	0.06
Science Item 5	0.86	0.06	0.11	0.02	0.66	0.12	0.06
Science Item 6	0.81	0.03	0.06	0.05	0.82	0.02	-0.01
Science Item 7	0.83	0.03	0.13	0.05	0.73	0.09	-0.05
Science Item 8	0.83	0.04	0.05	0.03	0.81	0.02	0.04
Science Item 9	0.85	0.05	0.10	0.03	0.72	0.05	0.05
Science Item 10	0.85	0.03	0.09	0.04	0.74	0.06	0.05
Science Item 11	0.75	-0.04	0.12	0.15	0.73	-0.03	-0.04
Science Item 12	0.83	-0.02	0.08	0.08	0.71	0.13	0.00
Science Item 13	0.84	-0.01	0.10	-0.01	0.71	0.16	0.03
Science Item 14	0.81	0.00	0.10	0.00	0.82	0.05	-0.02
Science Item 15	0.80	0.13	0.04	0.15	0.62	0.07	-0.06
Literacy Item 1	0.84	0.06	0.11	0.06	0.01	0.68	0.06
Literacy Item 2	0.86	0.03	0.20	0.01	0.11	0.65	0.01
Literacy Item 3	0.80	0.09	0.00	0.06	-0.03	0.73	0.10
Literacy Item 4	0.82	0.04	0.11	0.02	0.04	0.72	0.03
Literacy Item 5	0.83	0.06	0.11	0.05	0.04	0.70	0.03
Literacy Item 6	0.81	-0.03	0.05	-0.03	0.13	0.72	0.12
Literacy Item 7	0.84	-0.04	0.06	0.06	0.17	0.63	0.11
Literacy Item 8	0.85	-0.03	0.16	0.00	0.11	0.70	0.05
Literacy Item 9	0.83	0.09	0.16	0.04	0.01	0.70	-0.02

Table 18, continued.

	Unrotated Factor Pattern	Rotated Factor Pattern (Standardized Regression Coefficients)					
Statistics/Subscale	Factor 1	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Literacy Item 10	0.83	0.06	0.04	0.08	0.07	0.65	0.07
Literacy Item 11	0.83	-0.01	0.03	0.08	0.12	0.62	0.13
Literacy Item 12	0.85	0.03	0.05	0.13	0.11	0.63	0.06
Literacy Item 13	0.81	0.10	0.11	0.09	-0.02	0.67	0.01
Literacy Item 14	0.82	0.00	0.09	0.08	0.19	0.57	0.05
Literacy Item 15	0.84	0.08	0.06	0.09	0.08	0.68	0.00
Cognition Item 1	0.87	-0.02	0.28	0.06	0.07	0.10	0.53
Cognition Item 2	0.85	-0.05	0.21	0.02	0.10	0.10	0.61
Cognition Item 3	0.86	0.08	0.11	0.02	0.10	0.12	0.61
Cognition Item 4	0.86	0.05	0.12	0.06	0.05	0.11	0.64
Cognition Item 5	0.83	-0.05	0.12	0.09	0.13	0.03	0.66
Cognition Item 6	0.84	0.15	0.05	0.09	0.06	0.03	0.65
Cognition Item 7	0.84	0.20	0.10	0.04	0.01	0.12	0.55
Cognition Item 8	0.84	0.03	0.21	0.00	-0.06	0.27	0.56
Cognition Item 9	0.87	0.17	0.14	0.09	-0.01	0.08	0.57
Cognition Item 10	0.87	0.06	0.11	0.07	0.16	0.08	0.55
Cognition Item 11	0.86	-0.01	0.09	0.02	0.07	0.18	0.67
Cognition Item 12	0.85	0.02	0.08	0.04	0.13	0.09	0.66
Cognition Item 13	0.86	-0.03	0.19	0.15	0.11	0.05	0.55
Cognition Item 14	0.86	0.23	0.14	0.08	-0.03	0.11	0.52
Cognition Item 15	0.86	0.01	0.09	0.03	0.06	0.17	0.67

Table 19

Exploratory Factor Analysis of UTAGS Subscales Using the Entire Normative Sample

Subscale	Factor Loading
Eigenvalue	4.78
Percentage variance accounted for in Factor 1	.80
Cognition	.94
Literacy	.93
Math	.92
Science	.91
Creativity	.86
Leadership	.78

Confirmatory Factor Analysis. Another way to investigate construct-identification validity is to analyze the data from the normative sample using maximum-likelihood confirmatory factor analysis (CFA). To demonstrate the construct-identification validity of the UTAGS, a one-factor CFA model and a six-factor CFA model were tested. In testing these models, four indexes of fit were computed: Bentler's (1990) comparative fit index (CFI), Tucker and Lewis's (1973) index of fit (TLI), Bentler and Bonett's (1980) normed fit index (NFI), and Browne and Cudeck's (1993) root mean square error of approximation (RMSEA). The criterion for an acceptable fit varies among different types of 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 CFI, TLI, 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.

The first analysis examined the six-factor model using the item-level data. The items for each subscale were combined into 5-item sets (parcels) and allowed to load on their respective subscales. The subscales served as the latent variables in this model. The results of this analysis are presented in Figure 3. The values on the arrows between each subscale and the item parcels, which are represented by rectangles, are factor loadings. The factor loadings are regression coefficients that represent the influence of these factors on the scales. The e1 through e18 circles represent unique variance and systematic variance of each item parcel that is unrelated to the variances of the other parcels.

Applying Hopkins's (2002) criteria, the sizes of the factor loadings are all Very Large. The fit indices are also presented in Figure 3. These fit indices indicate the six-factor UTAGS model is a very strong fit (the CFI, TLI, and NFI exceed .97 and the RMSEA = .08), and supports the General Aptitude Index of the UTAGS.

The second CFA analysis examined a one-factor model with the UTAGS General Aptitude Index serving as the latent variable in this model, represented by an oval. The results of this analysis are presented in Figure 4. These fit indices indicate that the one-factor UTAGS model is also a very strong fit (the CFI, TLI, and NFI exceed .96 and the RMSEA = .10), and further supports the construct validity of the General Aptitude Index of the UTAGS.

CONCLUSIONS

Based on the information provided in the chapter, one may conclude that the UTAGS is a valid measure of behaviors and aptitudes related to school success, and particularly those behaviors at the extremes, and that examiners can use the scale with confidence. Validation is an ongoing process, however, and in all likelihood, additional validity studies will be forthcoming. We encourage professionals to continue to investigate the test using different samples, statistical procedures, and related measures. We also encourage these researchers to share their results with us so that their findings can be included in subsequent editions of the manual.

The data provided in this manual should be sufficient to establish the UTAGS as a useful addition to current methods of identifying and screening students for special needs, at both extremes of school success (e.g., those with intellectual and learning disabilities and those who are gifted). The UTAGS format requires that an examinee's behavior be compared to his or her peers as rated by a qualified individual—usually the student's teacher—and benefits from the capacity to make local comparisons. Results presented in this manual help establish validity of the instrument. Data strongly support use of UTAGS for screening students who are talented and/or gifted.

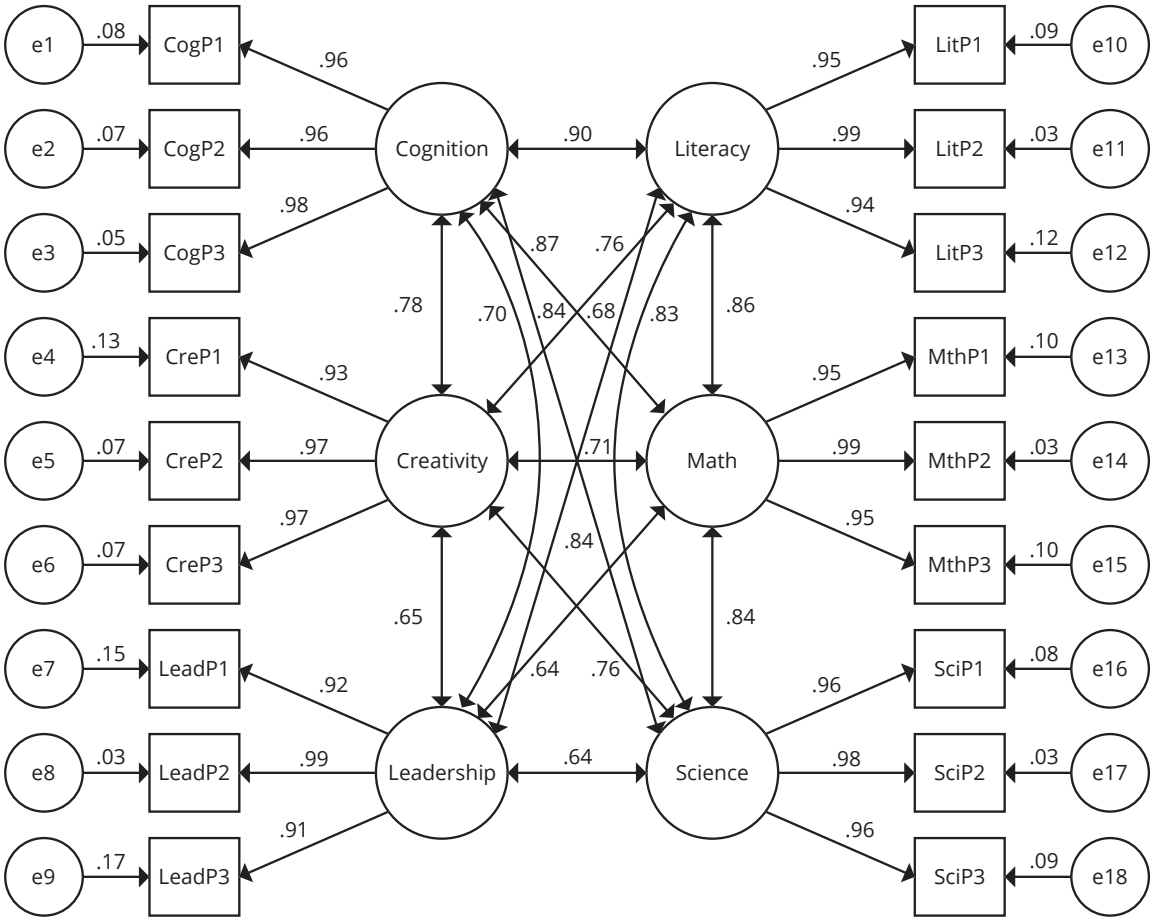

Figure 3. UTAGS confirmatory factor analysis results.

Note. CogP1 – CogP2 = Cognition Parcel 1–3; CreP1 – CreP3 = Creative Parcel 1–3; Lead P1 – Lead P3 = Leadership Parcel 1–3; LitP1 – Lit P3 = Literary Parcel 1–3; MthP1 – MthP3 = Math Parcel 1–3; SciP1 – Sci P3 = Science Parcel 1–3; df = degrees of freedom; CFI = Comparative Fit Index (Bestler, 1990); TLI = Tucker-Lewis Index (Tucker & Lewis, 1973); NFI = Normed Fit Index (Bentler & Bonett, 1980); RMSEA = Root Mean Square Error of Approximation (Brown & Ludeck, 1993).

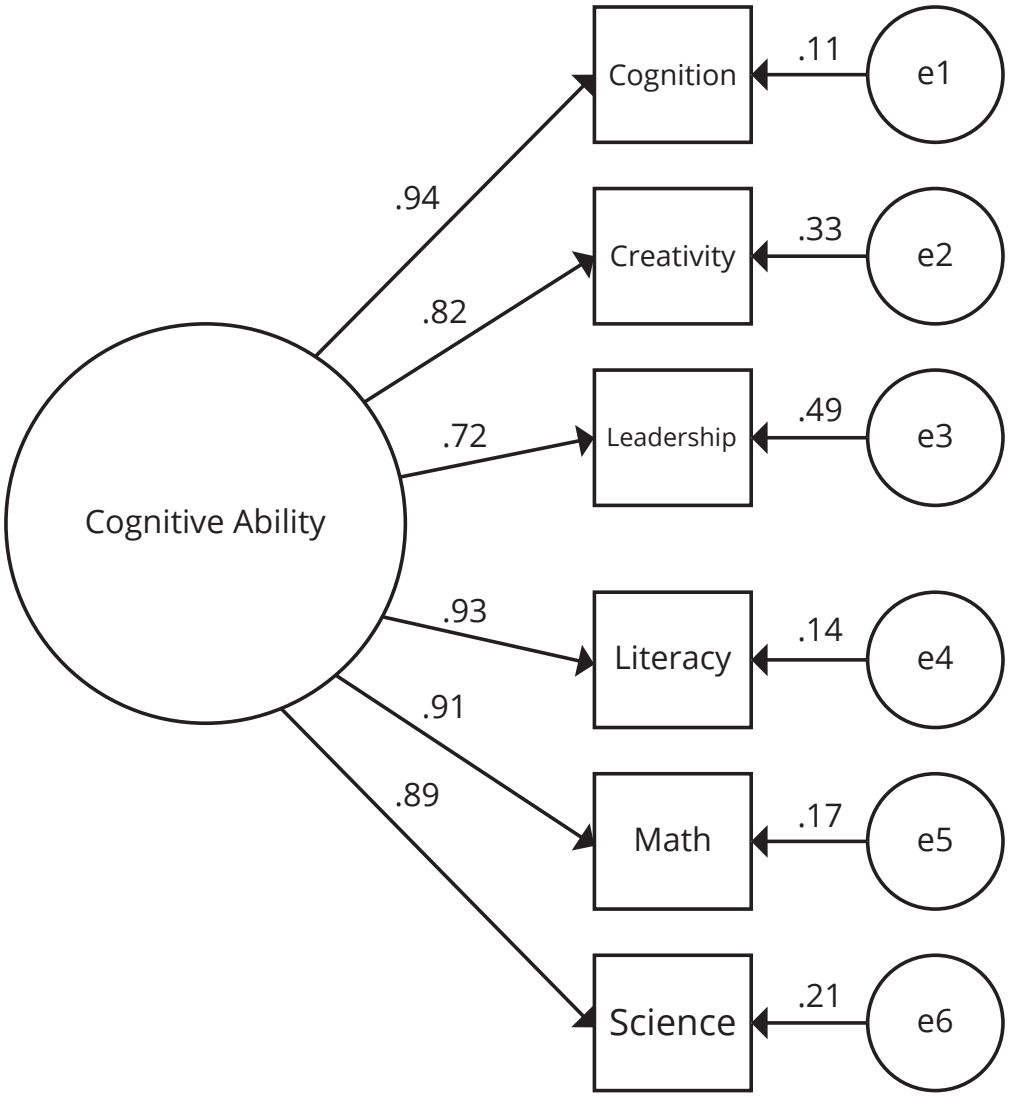

Figure 4. UTAGS confirmatory factor analysis results.

Note. df = degrees of freedom; CFI = Comparative Fit Index (Bestler, 1990); TLI = Tucker-Lewis Index (Tucker & Lewis, 1973); NFI = Normed Fit Index (Bentler & Bonett, 1980); RMSEA = Root Mean Square Error of Approximation (Brown & Ludeck, 1993).