12. VALIDITY

Overview

The ultimate use of test scores is to draw inferences about students' ability, competence, or behaviors in situations beyond that observed in the testing scenario. The high reliability of test scores implies consistency in test scores but cannot assure that the inferences drawn from test scores are valid and defensible.

Validity is another critical aspect that needs to be addressed in test development and evaluation according to the *Standards for Educational and Psychological Testing*,[1] and is related to fairness. While reliability addresses the consistency in test scores obtained from different forms, administrations, and time, validity addresses whether a test measures what it intends to measure. Validity refers to the degree to which evidence collected in the test scores and in the process of test development and test administration supports the inferences based on test scores as intended.[2]

According to the *Standards*, validity evidence is collected from the following aspects: content, response process, internal structure, relations with other variables, and consequences of testing. Validity evidence related to test content can be collected based on test specifications, alignment of test content with curriculum, and instruction if relevant for the purpose of the test. Further, test administration and scoring reflect more dimensions for content-related evidence of validity. Chapters 2, 3, 4, and 9 in this technical report provide such content-related evidence of validity.

Response process related evidence of validity can be collected in multiple ways. For example, students taking the test can be interviewed about how they respond to the items. Some think-aloud procedure can help item developers better understand test-takers' thinking and evaluate whether test-takers' thinking is consistent with what the item was intended to be. Further, students' problem-solving strategies could be investigated by observing students' responding behaviors, analyzing process data such as item response time and log files, and the relationship between responses and response process data.

---

1 American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

2 Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13 - 103). New York: American Council on Education and Macmillan.

Evidence Based on Internal Structure

In addition, the investigation of the internal structure of a test can provide important validity evidence. The internal structure can be evaluated in terms of dimensionality, construct equivalence, measurement precision (in terms of reliability, standard error of measurement, and test information) and differential item functioning. In general, construct equivalence across the subgroups of the test-taker populations and differential item functioning are related to test fairness. Test fairness, as part of validity evidence, means that comparable opportunities have been provided to test-takers to demonstrate their abilities on the constructs a test intends to measure.[3] In evaluating test fairness, such questions as whether the test measures the same construct in all relevant populations should be addressed. An investigation of the factor structure of a test and the invariance of the factor structure across subgroups of the student population can provide evidence of construct-related evidence of validity.[4]

To collect another source of evidence of validity, an investigation may be conducted of the relationship among the test scores and other variables such as SAT/ACT scores, high school and college GPA, and graduation. A multitrait-multimethod study will serve this purpose. Further, the influence of the CLT on instruction and school dropout rates can be examined to evaluate the intended and unintended consequences of testing.

The collection of validity evidence is an ongoing process. This technical report provides evidence from different sources in the test development and administration process. This chapter focuses on collecting evidence related to the internal structure of the CLT. It evaluates the internal structure of the three subjects: Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning in two test administrations, Form 1517 and Form 1618. It further investigates the influences of gender and ethnicity on the internal structure of these three subjects in each of the two forms. Both exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) are conducted using the domain scores.

EVALUATION OF THE INTERNAL STRUCTURE OF CLT USING EFA AND CFA

The evaluation of the internal structure of the CLT is conducted using the domain scores. Table 12.1 summarizes the number of items measuring each domain within each CLT subject test. The number of items assessing each domain across subjects ranged from 10 to 27. The number of items assessing each domain remained the same across the two forms. The descriptive statistics for the domain scores are summarized in Tables 12.2 and 12.3 respectively for Form 1517 and Form 1618. The distributions of the scores for each domain were similar across the two forms.

Table 12.1 The Number of Items Measuring Each Domain

Subject	Domain	Number of Items
Verbal Reasoning	Analysis	13
	Comprehension	27
Grammar/Writing	Grammar	20
	Writing	20
Quantitative Reasoning	Algebra	10
	Geometry	14
	Mathematical Reasoning	16

3 American Educational Research Association, et al., 2014.

4 Messick, S. (1995). Standards-based score interpretation: Establishing valid grounds for valid inferences. In *Proceedings of the joint conference on standard setting for large-scale assessments of the National Assessment Governing Board (NAGB) and the National Center for Education Statistics (NCES), Vol. II* (pp. 291-305). Washington, DC: National Assessment Governing Board and National Center for Education Statistics.

Table 12.2 Descriptive Statistics for the Domain Scores for Form 1517

Domain	N	Minimum	Maximum	Mean	Std. Deviation
verbal_Analysis	780	.00	13.00	6.74	2.37
verbal_Compreh	780	4.00	26.00	15.84	4.54
writing_Grammar	780	3.00	20.00	14.82	3.17
writing_Writing	780	2.00	20.00	10.17	3.20
quan_Algebra	780	.00	10.00	5.45	2.08
quan_Geometry	780	.00	14.00	7.26	3.18
quan_Reasoning	780	.00	16.00	8.23	3.13

Table 12.3 Descriptive Statistics for the Domain Scores for Form 1618

Domain	N	Minimum	Maximum	Mean	Std. Deviation
verbal_Analysis	276	0.00	13.00	7.89	2.76
verbal_Compreh	276	2.00	27.00	17.21	5.53
writing_Grammar	276	0.00	20.00	13.63	3.82
writing_Writing	276	0.00	20.00	9.42	4.04
quan_Algebra	276	0.00	10.00	4.63	2.45
quan_Geometry	276	0.00	13.00	6.38	2.73
quan_Reasoning	276	0.00	15.00	6.73	2.97

CORRELATIONS BETWEEN THE DOMAIN SCORES AND THE SUBJECT TEST SCORES

The correlations between the domain scores and the subject scores for each test form were computed and summarized in Tables 12.4 and 12.5 for Forms 1517 and 1618, respectively. Across forms, similar patterns were observed. In general, the correlation between the Verbal Reasoning scores and the Grammar/Writing scores was higher than those between the Quantitative Reasoning score and the Verbal Reasoning or Grammar/Writing scores. The correlations between the Verbal Reasoning subject scores and the Grammar/Writing domain scores were higher than those between the Verbal Reasoning subject scores and the Quantitative Reasoning domain scores. The correlations between the Grammar/Writing subject scores and the Verbal Reasoning domain scores were higher than those between the Grammar/Writing subject scores and the Quantitative Reasoning domain scores. On the other hand, the correlations between the Quantitative Reasoning subject scores and the Grammar/Writing domain scores and those between the Quantitative Reasoning subject scores and the Verbal Reasoning domain scores were all comparatively lower.

The correlations reported in Tables 12.4 and 12.5 in the box at the right corner are all related to the domain scores. In general, the domain scores from the same subject test tended to be more highly correlated. The same is true for the correlations between the domain score between the Verbal Reasoning and Grammar/Writing subjects. On the other hand, the correlations between the domain score from the Quantitative Reasoning test and those from either the Verbal Reasoning test or the Grammar/Writing test were all lower. The patterns observed across multi-traits fall within expectations and provided both convergent and divergent validity evidence.

Table 12.4 Correlations among the Domain Scores and the Subject Scores for Form 1517

	V_ Orig	W_ Orig	Q_ Orig	V_ Analysis	V_ Compreh	W_ Grammar	W_ Writing	Q_ Algebra	Q_ Geometry	Q_ Reasoning
V_Orig	1									
W_Orig	.746**	1								
Q_Orig	.538**	.583**	1							
V_Analysis	.837**	.610**	.442**	1						
V_Compreh	.958**	.722**	.520**	.646**	1					
W_ Grammar	.661**	.904**	.520**	.541**	.640**	1				
W_Writing	.688**	.906**	.535**	.562**	.667**	.638**	1			
Q_Algebra	.443**	.479**	.812**	.353**	.434**	.432**	.436**	1		
Q_ Geometry	.451**	.510**	.892**	.364**	.439**	.462**	.461**	.611**	1	
Q_ Reasoning	.504**	.524**	.890**	.427**	.480**	.458**	.491**	.611**	.661**	1

*** Correlation is significant at the 0.01 level (2-tailed).*

Table 12.5 Correlations among the Domain Scores and the Subject Scores for Form 1618

	V_ Orig	W_ Orig	Q_ Orig	V_ Analysis	V_ Compreh	W_ Grammar	W_ Writing	Q_ Algebra	Q_ Geometry	Q_ Reasoning
V_Orig	1									
W_Orig	.805**	1								
Q_Orig	.635**	.663**	1							
V_Analysis	.911**	.726**	.571**	1						
V_Compreh	.979**	.792**	.625**	.806**	1					
W_Grammar	.729**	.917**	.596**	.655**	.718**	1				
W_Writing	.755**	.926**	.626**	.682**	.742**	.700**	1			
Q_Algebra	.533**	.562**	.861**	.470**	.529**	.516**	.520**	1		
Q_Geometry	.559**	.560**	.870**	.510**	.546**	.491**	.541**	.657**	1	
Q_Reasoning	.559**	.601**	.871**	.503**	.550**	.543**	.565**	.622**	.611**	1

EXPLORATORY FACTOR ANALYSIS

Exploratory factor analyses were conducted based on the seven domain scores from the three subject tests for both test forms using the software Statistical Package for the Social Sciences (SPSS). Eigenvalues, eigenvalue differences and the percentage of variance explained by each factor were examined.

Form 1517, Eigenvalues and Eigenvalue Differences

The eigenvalues and the eigenvalue differences between factors for Form 1517 are summarized in Table 12.6. The eigenvalues for the first two factors were larger than 1. Kaiser (1960) recommends extracting the component based on the eigenvalue that is larger than 1.[5] According to Kaiser's rule, two components were extracted.

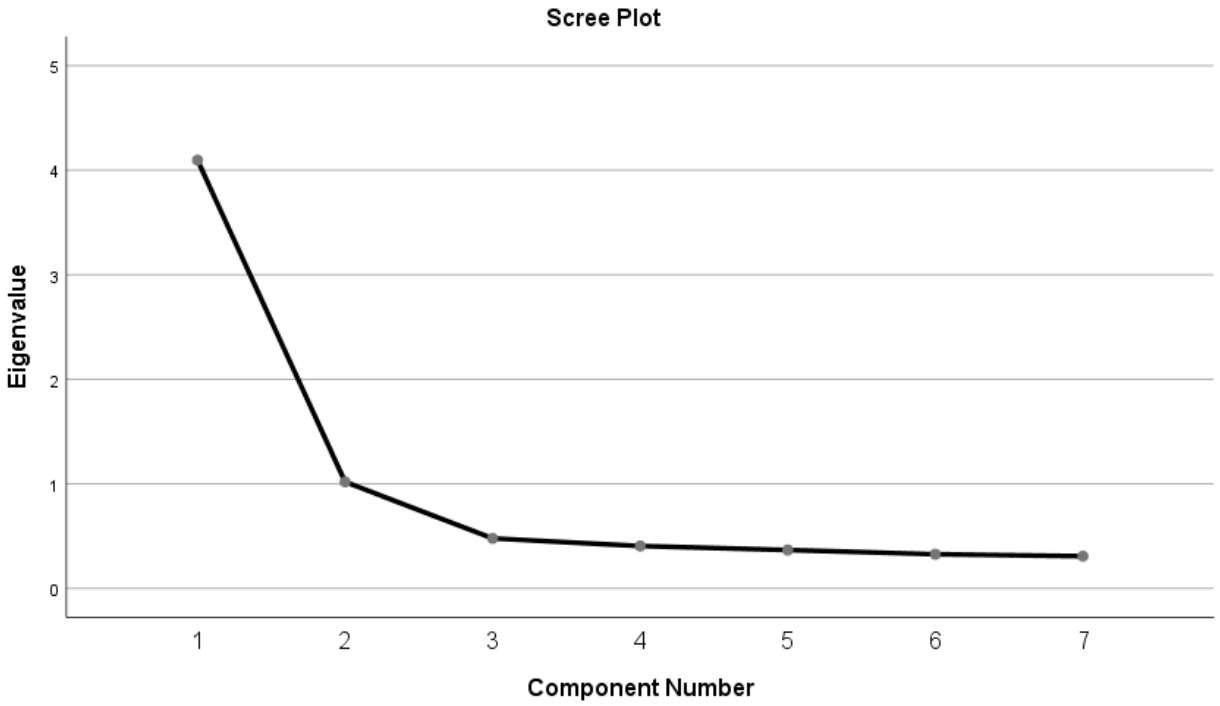
Hattie (1985) suggests using the ratio of the difference between the first factor and the second factor to the difference between the second and the third factor to examine the relative strength of the first factor.[6] This ratio was dubbed as the Factor Difference Ratio Index (FDRI) in Johnson, Yamashiro, and Yu (2003).[7] If this ratio is larger than 3, the first factor is relatively strong. The eigenvalues for the first factor were larger than 4, and the difference between the first two factors was around 3. Based on this criterion, the first factor was relatively strong. The scree plot presented in Figure 12.1 for Form 1517 in general supports one dominant factor.

Table 12.6 Variance Explained for Form 1517

Component	Total	% of Variance	Cumulative %	Differences between Eigenvalues	Factor Difference Ratio Index
1	4.096	58.518	58.518	3.077	5.688
2	1.019	14.558	73.076	0.541	
3	.478	6.826	79.903		
4	.406	5.794	85.696		
5	.367	5.243	90.939		
6	.326	4.663	95.602		
7	.308	4.398	100.000		

5 Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141-151.
6 Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.
7 Johnson, J. S., Yamashiro, A., & Yu, J. (2003). *ECPE annual report: 2002*. Ann Arbor, MI: English Language Institute, University of Michigan.

Figure 12.1. The scree plot for Form 1517 based on domain scores.



Percentage of Variance Explained

The percentage of variance explained by each factor is also presented in Table 12.6. Reckase (1979) suggested that a test is unidimensional if the first factor accounts for 20% or more of the total variance.[8] For Form 1517, the first component explained over 58.5% variance while the second component explained about 14.6% of the total variance. In general, one dominant component was supported.

Table 12.7 presents the loading of each component. The domain scores in the Verbal Reasoning and Grammar/Writing subjects load positively on component 1 but negatively on component 2. This indicates that component 1 is related to Verbal Reasoning and Grammar/Writing skills. On the other hand, quantitative domain scores load on both component 1 and 2, but with a larger weight on component 1 as well. This indicates that the Quantitative Reasoning domain scores are highly related to both the Verbal Reasoning and the Grammar/Writing components as well.

Table 12.7 Component Matrix

	Component	
	1	2
verbal_Analysis	.727	-.401
verbal_Compreh	.811	-.341
writing_Grammar	.784	-.265
writing_Writing	.801	-.269
quan_Algebra	.716	.470
quan_Geometry	.742	.468
quan_Reasoning	.769	.400

Extraction Method: Principal Component Analysis.

a. 2 components extracted.

8 Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics*, *4*(3), 207-230.

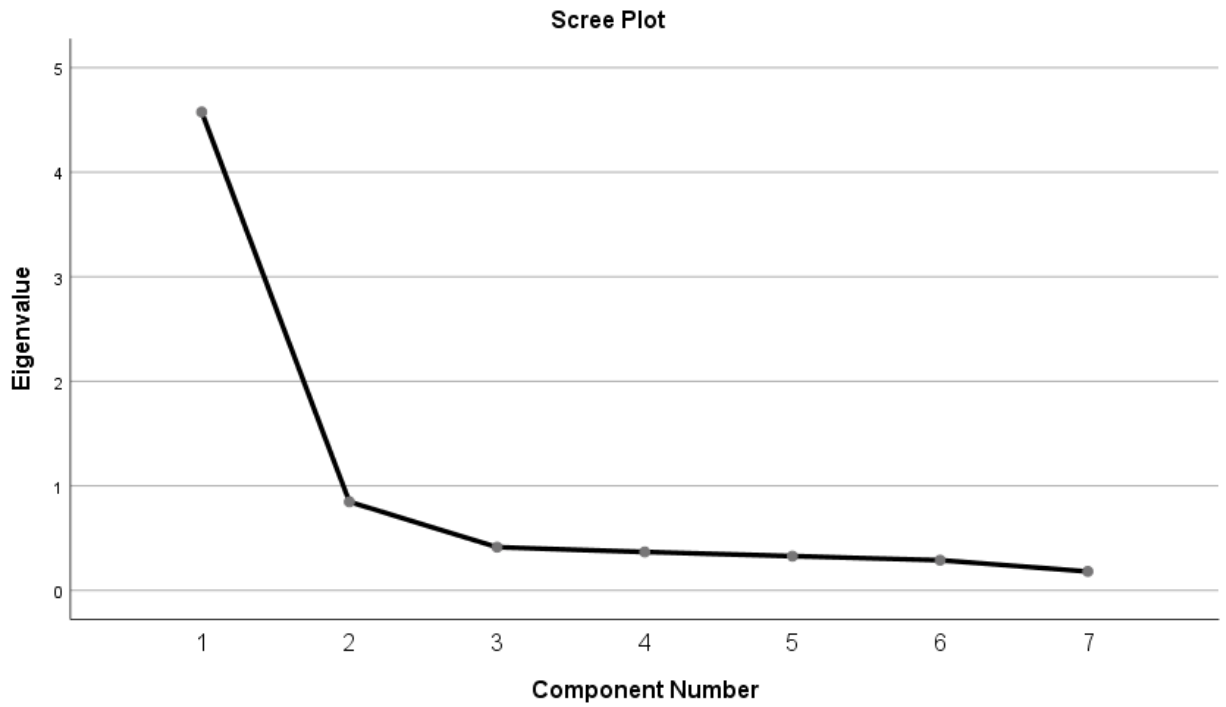
Form 1618, Eigenvalues and Eigenvalue Differences

The eigenvalues and the eigenvalue differences between factors for Form 1618 are summarized in Table 12.8. Only the eigenvalue for the first factor was larger than 1. The eigenvalue for the first factor were larger than 4, and the difference between the first two factors was above 3. According to Hattie (1985), the Factor Difference Ratio Index (FDRI) is larger than 3. The first factor is relatively dominant. Kaiser's rule (Kaiser, 1960) also suggested one component. The scree plot presented in Figure 12.2 for Form 1618 supports one factor.

Table 12.8 Variance Explained for Form 1618

Component	Total	% of Variance	Cumulative %	Differences between Eigenvalues	Factor Difference Ratio Index
1	4.576	65.367	65.367	3.729	8.592
2	.847	12.098	77.465	0.434	
3	.413	5.900	83.365		
4	.368	5.252	88.616		
5	.328	4.681	93.297		
6	.288	4.118	97.415		
7	.181	2.585	100.000		

Figure 12.2. The scree plot for Form 1618 based on domain scores.



Percentage of Variance Explained

The percentage of variance explained by each component for Form 1618 is presented in Table 12.7. The first component accounts for 65.4% of the total variance while the second component explained about 12% of the total variance. The results from EFA supported unidimensionality for Form 1618.

CONFIRMATORY FACTOR ANALYSIS

In general, the results from EFA conducted for Forms 1517 and 1618 were not consistent in identifying the number of factors extracted. In general, the empirical data from Form 1517 supported a two-factor model while that for Form 1618 supported a one-factor model. In theory, CLT was developed to assess students' ability in three content areas: Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning. Thus, three models: a one-factor model, a two-factor model, and a three-factor model were fitted to the domain scores for each form respectively. The three models are presented in Figures 12.3, 12.4, and 12.5, respectively.

Figure 12.3. The One-Factor Model.

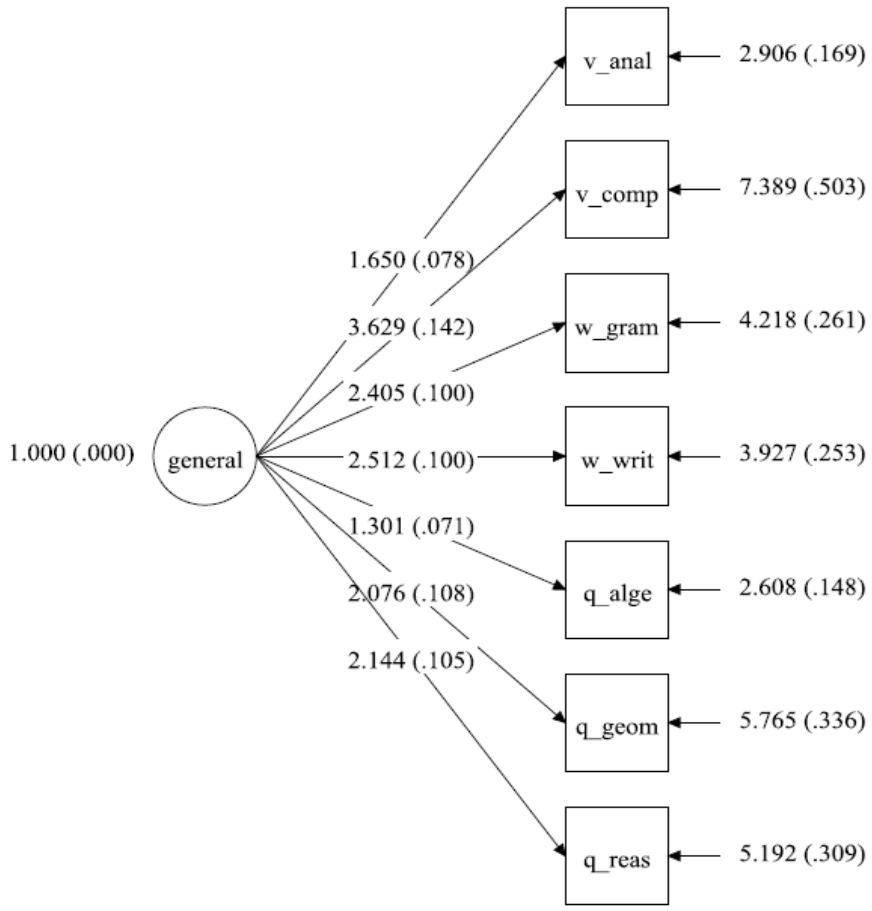


Figure 12.4. The Two-Factor Model.

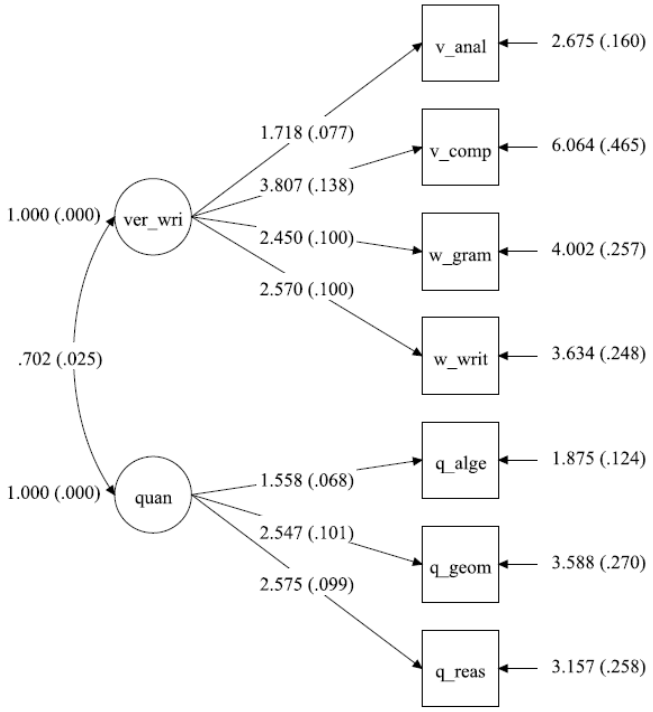
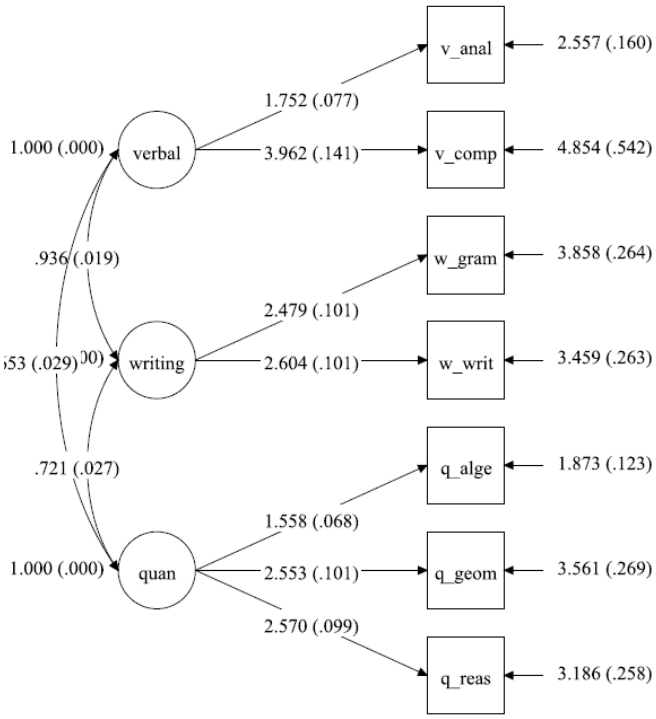


Figure 12.5. The Three-Factor Model.



Several goodness-of-fit indices (GFI) were used to evaluate the model-data fit. These include chi-square ($\chi 2$), the comparative fit index (CFI), the Tucker-Lewis Index (TLI), the root mean square error of approximation (RMSEA) and the standardized root mean square error residual (SRMR). Based on Hu and Bentler (1999), the cut values for a model with good fit are CFI > 0.95, TLI > 0.95, RMSEA < 0.06, and SRMR < 0.08.[9] In practice, a rough guideline is that for absolute fit indices and incremental fit indices (such as CFI and TLI), cutoff values could be just above 0.90 and for fit indices based on residual matrices (such as RMSEA and SRMR), values below 0.10 or 0.05 are usually considered adequate. In general, these fit indexes supported the conclusion that both the two-factor and the three-factor models were the best fitting models.

Further, information-based relative fit indexes were also examined. They are Akaike's information criterion (AIC; Akaike, 1974)[10] and Bayesian information criterion (BIC; Schwarz, 1978).[11] The best fitting model is the one that minimizes AIC or BIC. The fit indexes are computed as follows.

$$AIC = \overline{D(\xi)} + 2p$$

$$BIC = \overline{D(\xi)} + p \ln$$

where $\overline{D(\xi)}$ is the posterior mean of the deviance, a measure of fit; p is the number of model parameters to be estimated; N is the sample size; and $p_D = \overline{D(\xi)} - D(\hat{\xi})$ which is the difference between the posterior mean of the deviance ($\overline{D(\xi)}$) and the deviance of the posterior model ($D(\hat{\xi})$) based on the posterior estimates of the parameters. Both AIC and BIC as reported in Table 12.9 identified the three-factor model as the best fitting model. Also, the chi-square test as reported in Table 12.9 supported the three-factor model as the best fitting model. This is consistent with the theoretical design of CLT with content coverage in Verbal Reasoning, Grammar/Writing, and Quantitative Reasoning.

Table 12.9 Model Comparison

Form	Model	AIC	BIC	χ^2 (*df*)	*p*-value	RMSEA	CFI	TLI	SRMR
1517	1-factor	25229.349	25327.194	396.356 (14)	<.001	0.187	0.859	0.789	0.072
	2-factor	24862.271	24964.776	27.279 (13)	0.011	0.038	0.995	0.992	0.016
	3-factor	24848.073	24959.896	9.081 (11)	0.614	<.001	1.000	1.001	0.009
1618	1-factor	9044.060	9120.088	131.844 (14)	<.001	0.175	0.906	0.858	0.062
	2-factor	8939.704	9019.353	25.488 (13)	0.020	0.059	0.990	0.984	0.023
	3-factor	8925.044	9011.933	6.828 (11)	0.813	<.001	1.000	1.006	0.013

In summary, this section assessed the internal structure of two test forms of CLT, Form 1517 and 1618. It intends to provide validity evidence related to the internal structure of the CLT forms. EFA supported different numbers of components to be extracted. CFA empirically identified the best fitting model, the three-factor model. This is consistent with the theoretical framework and the theoretical content model in the CLT design and development.

9 Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6, 1-55.*

10 Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-723.

11 Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Evidence Based on Content

In addition to being technically valid, the content of the CLT also passes a reasonableness test. According to the Standards for Educational and Psychological Testing (2014), "Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."[12] Parameters for the CLT are designed to ensure that test results yield appropriate indicators of individuals' capacity for higher-level thinking as well as preparation for college. The range of question types in each of the three test sections provide a reasonable assessment of the kind of knowledge and skills that colleges value.

On the Verbal Reasoning section, questions are broken down into two types: Comprehension questions, which include the subdomains "Passage as a Whole," "Passage Details," and "Passage Relationships"; and Analysis questions, which include the subdomains "Textual Analysis" and "Interpretation of Evidence."

As a result, students are asked to engage with a text on two essential levels: first, their understanding of the text's meaning, the author's intent, and the information conveyed by the passage, and second, their ability to analyze and synthesize information in the text to draw valid conclusions. This reflects the multi-level analysis that students are required to engage in during high school, college, and beyond: students are asked not only to assess and comprehend a text, but to draw new ideas and conclusions from it.

On the Grammar/Writing section, questions are also broken down into two types: Grammar questions, which include the subdomains "Agreement" and "Punctuation and Sentence Structure"; and Writing questions, which include the subdomains "Structure," "Style," and "Word Choice."

Here, Grammar questions serve to evaluate a student's ability to use English standards and conventions properly, so as to clearly convey ideas and information. Writing questions serve to evaluate a student's ability to use language and style to accurately and appropriately convey the tone, argument, and intent of the text. Both skills are essential for high-level writing: to succeed at the college level, students must not only have a grasp of the conventions required to convey their arguments properly, but also the ability to clearly and concisely communicate their ideas.

On the Quantitative Reasoning section, questions are broken down into three types: Algebra, Geometry, and Mathematical Reasoning. Algebra questions include the subdomains "Arithmetic and Operations" and "Algebraic Expressions and Equations." Geometry questions include the subdomains "Coordinate Geometry," "Properties of Shapes," and "Trigonometry." Mathematical Reasoning questions include the subdomains "Logic" and "Word Problems."

The breakdown of Quantitative Reasoning questions into three types mirrors the types of logical reasoning and analysis skills that will serve students well in college and beyond. Algebra questions test students' ability to understand and work with symbols. Geometry questions test students' spatial abilities and understanding of shapes such as lines, triangles, squares, and other 2-D and 3-D shapes. Mathematical Reasoning questions test students' logical abilities. These skills are not only necessary for students interested in pursuing higher-level mathematics or science coursework in college, but are important indications of a student's ability to think clearly and logically, a crucial skill regardless of academic discipline.

Summary

In terms of both content and internal structure, the CLT exam demonstrates a high level of validity. Analysis of the test's structure found that a three-factor model is a good fit for evaluating its domain scores, as the exam is composed of three equally weighted subject tests. The types of questions in each subject test correspond to key skills in reading, writing, and mathematics. CLT test scores are thus a legitimate measure of students' aptitude and preparation for academic work at the college level and beyond.

12 AERA, 2014.