CLT

2024 TECHNICAL REPORT

CLT3-8

Table of Contents

Contributors

Report Lead Noah Tyler

Psychometrician & Writer Eren Asena

Research/Statistical Analyst, CLT; MSc, Methodology & Statistics, University of Amsterdam

Review Natalie Walker

Cover Design Meg Pilcher

1. Introduction

This technical report describes the psychometric characteristics of the CLT3-8 suite of assessments. As such, it serves as the technical counterpart of the CLT3-8 Assessment Content Briefs, which provide a detailed description of the purpose and the content of each test. Specifically, this report describes the psychometric procedures used in the development, scaling, and scoring of the CLT3-8, and analyzes its reliability. Evidence for the validity of the CLT3-8 can be found in the reports of the linking studies CLT conducted with MetaMetrics, Inc. (MetaMetrics, 2023). To facilitate the interpretation of the technical findings presented in this report, we begin with a brief overview of the content, structure, and the scale of the tests.

A BRIEF OVERVIEW OF THE CLT3-8 SUITE OF ASSESSMENTS

All the items on the CLT3-8 are dichotomous multiple choice questions. No points are lost for an incorrect response. The tests are scored on two main sections: Verbal Reasoning (VR) and Quantitative Reasoning (QR). In CLT3-6, the VR section has 60 items and the QR section has 50 items. CLT7 preserves the structure of the CLT8, which has the three sections of VR, Grammar/Writing (GW), and QR, each containing 40 items. Notwithstanding, the VR and the GW sections of the CLT7 and CLT8 are also scored as a single VR section. Both the VR and the QR sections are scored on a 150-300 scale, and this is true for both the CLT3-6 and the CLT7-8. That is, a student's score is not the total number of questions they answered correctly - statistical methods are used to transform the number-correct scores to *scale scores* between 150-300. These statistical methods are described in Chapter 2.

THE LEXILE® FRAMEWORK FOR READING AND THE QUANTILE® FRAMEWORK FOR MATHEMATICS

In addition to CLT scale scores, students who take the CLT3-8 receive Lexile® reading measures and Quantile® measures. These measures are associated with the national norms described in Chapter 5. To further facilitate score interpretations, the VR and QR scales are centered around the national 50th percentile of Lexile® and Quantile® measures. Specifically, the scale scores are calculated such that a score of 225 always corresponds to the national 50th percentile of Lexile®/Quantile® measures in each grade. This readily allows students, families, and teachers to place their CLT3-8 scores below or above the national 50th percentile. Given the link between CLT3-8 and Lexile® and Quantile® measures, CLT3-8 scores can be mapped to scores from any standardized test that reports Lexile® or Quantile® measures. This mapping can be used to evaluate the degree to which a student's achievement in reading and mathematics matches the standards

that the linked test is used to meet.

As stated in the Quantile® Framework linking study report, Quantile® measures can be used to "determine if a student is ready for a new mathematics skill or concept" and "link big mathematical concepts with state curriculum objectives" (MetaMetrics, 2023, p.1). The Quantile® Framework for Mathematics has defined over 500 mathematics skills and/or concepts. Each of these concepts has a measure, and each measure shows how difficult one skill is in relation to the others. These skills and concepts are aligned to state content standards for all 50 states. The alignments can be accessed using the Quantile® Math Skills Database tool that allows users to see Quantile® measures for each state standard. Educators can use Quantile® measures to match student mathematics ability with identified mathematics skills aligned with their state standards.

Similarly, the Lexile® Framework for Reading can help parents and teachers to understand a student's reading level in the context of state ELA standards. State ELA content standards enable students to read and understand texts of steadily increasing complexity, focusing on how well students read and comprehend the text. The Lexile® Framework for Reading describes the text complexity necessary for students to meet the demands of colleges and careers at each grade level. These grade and Lexile® bands are the basis for determining at what text complexity level students should be reading—and at which grades—to make sure they are ultimately prepared for the reading demands of college and careers (See more information at Prepare for College & Careers). Note that Lexile® reading measures are an indicator of inferential comprehension which means that they measure reading ability across the reading continuum rather than specific skills.

2. Field Testing and IRT Calibrations

CLT3-8 will be administered in an operational setting for the first time in Spring 2024. The CLT8 was launched in 2018 and the CLT7 is created from the CLT8 item pool. The difference between the CLT7 and the CLT8 is that the CLT7 is easier and more suitable for 7th grade students. The CLT3-6 items were field tested in Spring 2023, using a spiral design in which six test forms were randomly assigned to approximately 3000 students in each grade. 25% of the items in each form consisted of common items shared with the other forms. In addition, the forms contained Lexile® and Quantile® items written by MetaMetrics which allowed MetaMetrics to link the VR scale to the Lexile® Framework for Reading and the QR scale to the Quantile® Framework for Mathematics. In parallel with the field test, we administered a CLT8 Norm Referencing test to 7th and 8th grade students which contained MetaMetrics items to include the CLT7 and CLT8 in the linking study. The details of the linking studies can be found in the Lexile® Framework for Reading and Quantile® Framework for Mathematics linking study reports prepared by MetaMetrics (MetaMetrics, 2023).

CLASSICAL ITEM ANALYSIS AND DATA REVIEWS

After the CLT3-6 field test, the items were analyzed using Classical Test Theory (CTT) and the items that performed poorly were reviewed by content experts. Each content expert voted to either keep, reject or modify an item. Rejected and modified items are not used in operational test forms. The items that were not rejected during the data reviews were calibrated using the Rasch model as explained in the next section. Item performance was evaluated in terms of difficulty and discrimination. Difficulty was measured by the proportion of students who answered the item correctly, referred to as the *p* value (Crocker & Algina, 2008) (Equation 1). The higher the *p* value of an item, the easier the item is.

$$p_i = \frac{\sum_{n=1}^{N} X_{ni}}{N} \tag{1}$$

where X_{ni} is the response of student n to item i, coded as 1 for a correct answer and 0 for an incorrect answer. N is the total number of students. Items with a *p* value below 0.3 or above 0.9 were flagged to be reviewed by content experts. Item discrimination was assessed using the point-biserial correlation (r_{pb}), which is the correlation between the responses to an item and the total scores on a section, ranging between 0 and 1 (the calculation of the total scores excludes the item being analyzed). If an item has a high discrimination, it is more likely to be answered correctly by students with high ability than low ability. Thus, the correlation between the responses to the item and the total scores obtained on the test will have a large, positive correlation. A negative point-biserial means that high ability students are less likely to answer the item correctly than low ability students, which may indicate an issue with the answer key and needs to be reviewed. Items with a point-biserial correlation lower than 0.15 were flagged to be reviewed by content experts. In addition to the point-biserial correlation between the correct answer choice and the total scores, we calculated the point-biserial correlation between the responses to the distractors and the total scores, which allowed the content experts to evaluate the quality of each answer choice.

ITEM CALIBRATION AND SCALING

Each test type/grade level in the CLT3-8 suite has multiple operational test forms that contain different sets of items. Given that students take different versions of the same test, it is important that every test-taker is scored fairly and consistently. However, if the items on two different forms vary in content and difficulty, and the forms are scored simply based on the number of correct responses, then the scores on the two forms cannot be compared. CLT uses Item Response Theory (IRT) to adjust for potential difficulty differences between the forms. IRT consists of a family of

latent variable models that model the probability of a correct response to an item based on the interaction between item parameters and latent ability parameters. Item and ability parameters obtained from different test forms are placed on the same scale through a process of calibration that is described below (Kolen & Brennan, 2014). Using IRT enables two key outcomes: 1) measurement of an individual student's ability that is independent of the items on a particular test form, and 2) evaluation of test items that is independent of any particular group of test-takers. However, the scale of latent ability estimates obtained from IRT models is hard to interpret, so the ability estimates are transformed to scale scores that can be interpreted and understood more easily by stakeholders.

CLT uses a particular IRT model called the Rasch model. The Rasch model quantifies the probability that a given test-taker will answer a given item correctly as a function of two variables: the test-taker's ability and the item's difficulty (Equation 2). The more capable the student and the easier the item, the higher the probability that the student will get the item right.

$$P_{ni} = \frac{\exp^{\theta_n - b_i}}{1 + \exp^{\theta_n - b_i}} \tag{2}$$

where P_{ni} is the probability that test-taker n will answer item i correctly, θ_n is the ability of test-taker n, and b_i is the difficulty of item i. Both the ability estimates and the difficulty estimates are on the log-odds scale, also called the logit scale. Most observed logit values fall in the -3 to 3 range. The Rasch model assumes unidimensionality, which means that the items on the test measure only a single construct/ability. The CLT3-8 were designed to measure the two constructs of Verbal Reasoning and Quantitative Reasoning, so we fit the Rasch model to the two sections separately.

The fit of the items to the Rasch model can be examined using the outfit and infit mean squares (MSQ) (Wright & Masters, 1982). The outfit MSQ of an item is the average of its squared standardized residuals, which are the squared differences between the observed responses in the data and the response probabilities predicted by the model, divided by the modeled variance of the response (Equation 3).

$$OutfitMSQ_i = \frac{\sum_n z_{ni}^2}{n} \tag{3}$$

where $z_{ni} = \frac{X_{ni} - E(X_{ni})}{\sqrt{Var(X_{ni})}}$, X_{ni} is the observed response of student n to item i, $E(X_{ni})$ is the expected response of student n to item i, and $Var(X_{ni}) = E(X_{ni})(1 - E(X_{ni}))$ is the variance of a student's response to an item. Outfit statistics are sensitive to outliers such as lucky guesses on hard questions by low-ability students or careless mistakes on easy questions by high-ability students. In contrast, the infit MSQ accounts for outliers by weighing the squared residuals by the proximity between

an item's difficulty and a student's ability (Equation 4). For instance, for hard items, prediction errors for high-ability students are weighted more heavily than the prediction errors for low-ability students.

$$InfitMSQ_i = \frac{\sum_n z_{ni}^2 \times Var(X_{ni})}{\sum_n Var(X_{ni})} \tag{4}$$

where z_{ni} and $Var(X_{ni})$ are defined as above. Items with an infit or outfit value above 1.5 are excluded from operational forms as it indicates substantial model misfit (Linacre, 2002).

IRT models have scale indeterminacy, which means that for any given value of item difficulty, we can find an ability level that retains the same probability of a correct response to that item. To solve the issue of scale indeterminacy, CLT follows the common practice of constraining the mean of student abilities to be 0, which sets a reference point for the estimation of both item difficulties and student abilities. Although constraining the mean of the ability distribution determines the scale, a calibration process is necessary to ensure that the logit values derived from different forms are on a consistent scale and thus comparable. This is because if the students who took the different forms differ in ability, the "0" that serves as the reference point will represent different ability levels in each analysis. Items administered to different groups of students can be placed on the same scale using the common items shared between the forms (Kolen & Brennan, 2014). This study design is known as the common-item nonequivalent group design, and the common items are referred to as anchor items when used to link the scales of two forms. The psychometric literature suggests that to achieve a stable calibration, at least 20% of the items in a form should be anchor items (Kolen & Brennan, 2004).

The CLT3-6 items were calibrated using the concurrent calibration method in which all the items were analyzed in a single run of the Rasch model. Given that the unique items are analyzed simultaneously with the anchor items, all the item parameters are automatically placed on the same scale. CLT uses the WINSTEPS® software (Linacre, 2023) for Rasch calibrations. The CLT8 items were calibrated using the fixed-parameter calibration method. First, the Spring 2023 Norm Referencing form was taken as a base form and calibrated freely. Then, the CLT8 form that shared the largest number of common items with the base form was calibrated by fixing the parameters of the anchor items to the values obtained from the free calibration. The resulting item pool, which included the items of both the base form and the new form, was used to calibrate the items of a third form. This chain-equating process was repeated until all the forms were calibrated. We checked the stability of the anchor items to ensure that they functioned the same in the forms that were linked. Items may function differently across forms due to item drift which refers to the fact that the

difficulty of an item may change over time. To evaluate drift, we used the displacement statistic from WINSTEPS® (Linacre, 2023). Displacement is the difference between the anchored difficulty of an item and the difficulty estimate that would be obtained if the item was calibrated freely in the new form (i.e., if it was "unanchored"). Items that showed a displacement of 0.5 or more logits in absolute value were not used as anchor items (O'Neill et al., 2013). Instead, their difficulty parameters were estimated freely and updated to reflect the most recent estimate. Further, items with a point-biserial correlation of less than 0.10 were excluded from the anchor item pool.

DATA CLEANING AND MISSING VALUES

Before the analyses, we applied certain exclusion rules to ensure that the calibration samples were representative of the population, the assumptions of the Rasch model would be met, and the parameter estimates would be unbiased. For the calibrations of CLT8, repeat attempts were excluded from the item calibrations: that is, only the first attempt of each student was used to calibrate the items. For all the tests (CLT3-8), we excluded students who did not attempt a given section from the calibrations of that section. Third, missing/blank responses were treated differently in item calibration and scoring; during calibrations, we made a distinction between omitted and not-reached items. Omitted items are items to which a student did not respond but after which the student continued the test. Given that the student had responses to the subsequent items, we assume that the student saw the omitted items and decided not to respond because they thought that the item was too difficult. In contrast, not-reached items are the missing responses at the end of the test – the missing responses that are not followed by any response. We assumed that the student did not actually encounter these items either because they ran out of time or decided to stop the test. Since the student never reached these questions, we do not know if they could have answered them correctly or not. Therefore, these questions were left as missing values during the calibrations. The distinction between omitted and not-reached items are only made during item calibration. When scoring students, all missing responses are treated as 0. For a detailed treatment of this approach with examples and an explanation of its advantages, we refer the reader to Ludlow and O'Leary (1999).

3. Test Assembly and Scoring

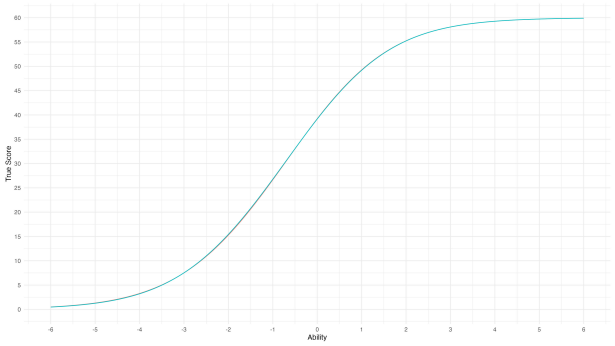
AUTOMATED TEST ASSEMBLY (ATA)

CLT uses automated test assembly (ATA) to construct test forms that are parallel in content and statistical specifications. ATA involves computer algorithms that translate a set of constraints defined by psychometricians and content experts into mathematical optimization problems. Constraints

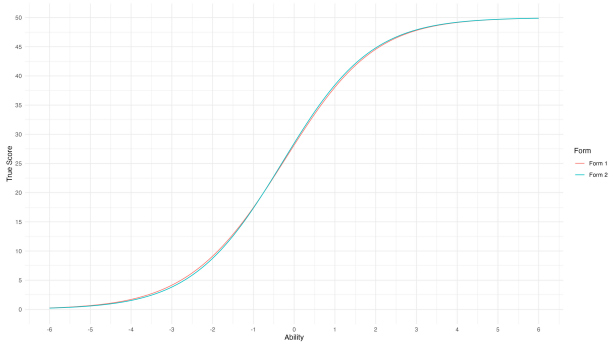
related to the content of the tests come from the blueprints created by our test development team and include the number of items each form should include from each subject, domain, subdomain, passage type, and question type. Furthermore, ATA allows us to construct forms that have a consistent level of difficulty that is appropriate for the intended grade level. For example, although CLT7 is built entirely out of CLT8 items, ATA allows us to automatically select items that have the right level of difficulty for a 7th grade student. This is accomplished by defining an objective function, which is the statistical outcome that the ATA algorithm strives to achieve. When the objective function is defined as a target difficulty level, the software finds the combination of items that minimize the differences between the target difficulty and the difficulty of the forms while satisfying the content constraints. The items are pulled from an item bank that is maintained and updated by our Test Development team and psychometricians. Item difficulties are estimated using IRT, which was discussed above. ATA is conducted using the eatATA package (Becker et al., 2021) in the R programming language (R Core Team, 2023). Figure 1 shows the test characteristic curves (TCCs) of the CLT3-8 test forms that will be operational in Spring 2023. A TCC shows the expected number-correct score on a form given an ability level and the item difficulties. The abilities are on the logit scale as explained in above. Each curve in the plots is the TCC of a single module. A high overlap between the curves means that the difficulty differences between the modules are small. Given that there are a finite number of items from which the modules can be created, it is difficult to assemble forms that are identical in difficulty. The section on IRT calibrations explains how our scoring process adjusts for the differences between the forms to ensure that scores obtained from different forms are on the same scale and can be compared. Once the parallel test forms have been constructed, the passages and the items are uploaded into the test delivery platform.

Figure 1

The Test Characteristic Curves (TCCs) of the CLT3-8 Operational Forms

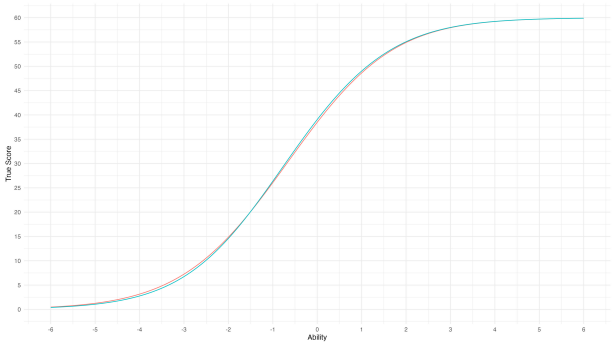


(a1) Verbal Reasoning

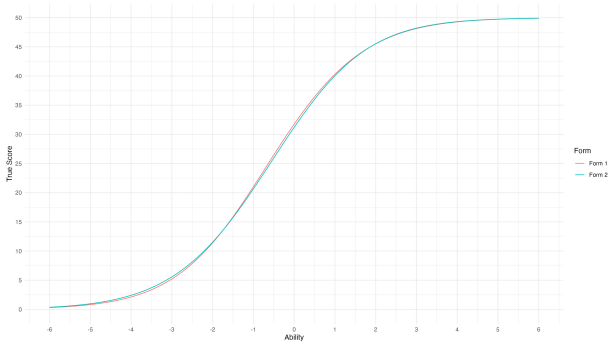


(a2) Quantitative Reasoning

(a) TCCs of CLT3 Forms

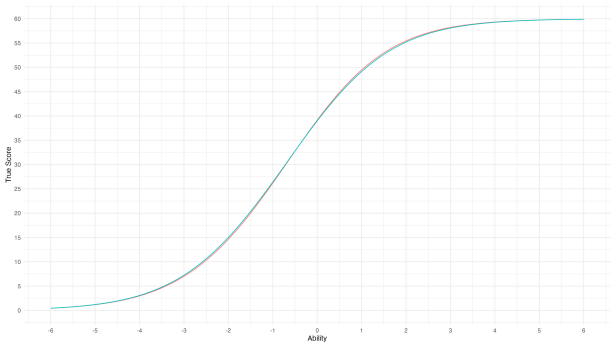


(b1) Verbal Reasoning

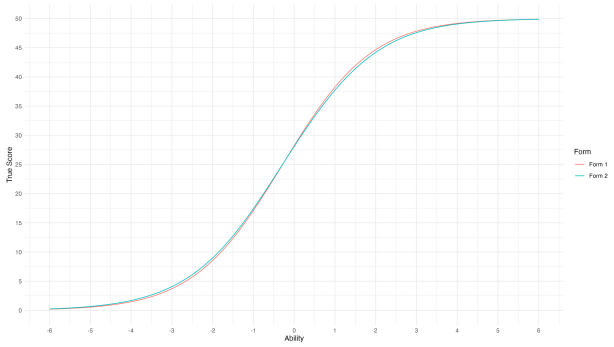


(b2) Quantitative Reasoning

(b) TCCs of CLT4 Forms

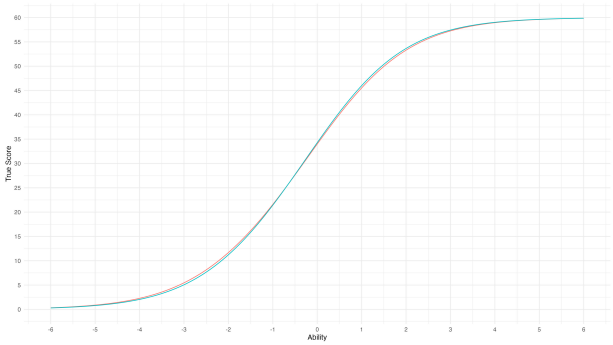


(c1) Verbal Reasoning

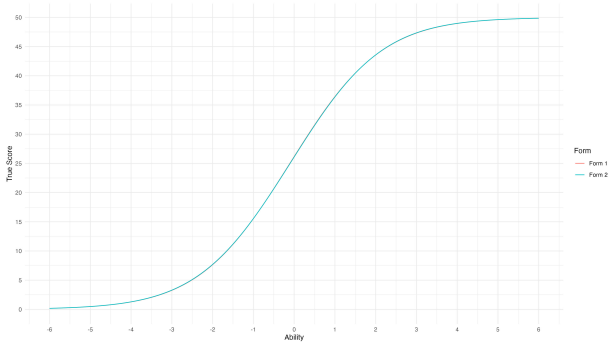


(c2) Quantitative Reasoning

(c) TCCs of CLT5 Forms

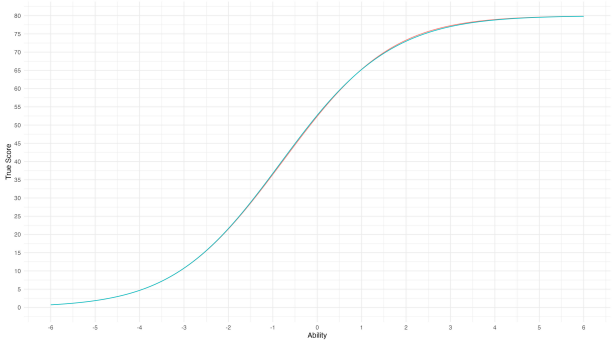


(d1) Verbal Reasoning

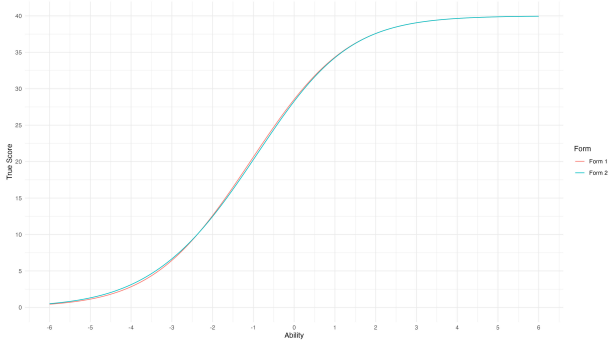


(d2) Quantitative Reasoning

(d) TCCs of CLT6 Forms

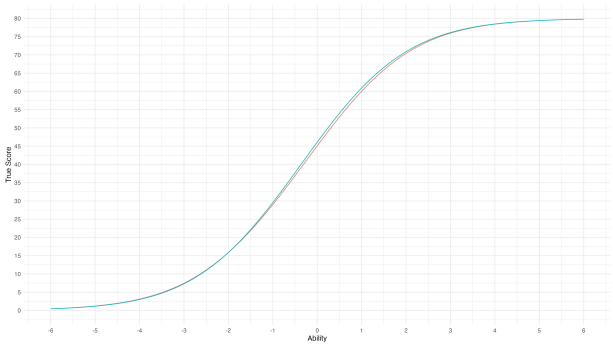


(e1) Verbal Reasoning

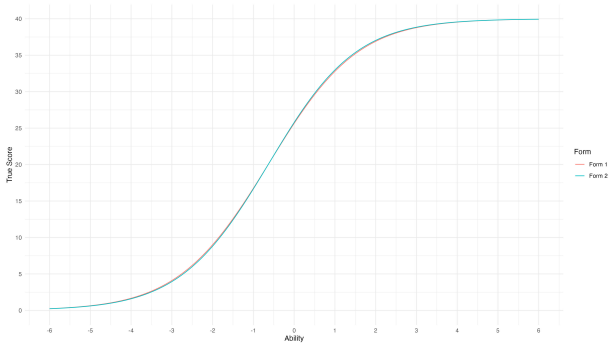


(e2) Quantitative Reasoning

(e) TCCs of CLT7



(f1) Verbal Reasoning



(f2) Quantitative Reasoning

(f) TCCs of CLT8 Forms

SCALING AND REPORTED SCORES

Once the items have been calibrated and their difficulty parameters have been estimated, we use WINSTEPS® (Linacre, 2023) to obtain raw-to-theta tables. These tables assign an ability value θ to each raw score on the test. These ability estimates are linearly transformed to scale scores between 150-300 that are reported to students. The scaling equation is of the form given in Equation 5:

$$SS = A + B \times \theta \tag{5}$$

where A is the intercept and B is the slope. As indicated in the Introduction, the a slope and an intercept is chosen such that the score point 225 corresponds to the national 50th percentile of Lexile® or Quantile® measures. The slope is found by transforming the theta range to the 150-300 scale score range (Equation 6):

$$B = \frac{300 - 150}{\theta_{\text{Max}} - \theta_{\text{Min}}} \tag{6}$$

Since the theta range is theoretically infinite, the best values of θ_{Max} and θ_{Min} were found through experimentation and by observing the resulting scale score distribution. Once the slope has been identified, the intercept can be found by solving Equation 7 for A:

$$\begin{aligned} A + B \times \theta_{\text{Median}} &= 225 \\ A &= 225 - B \times \theta_{\text{Median}} \end{aligned} \tag{7}$$

where θ_{Median} is the ability level that corresponds to the national 50th percentile of Lexile® or Quantile® measures for a given grade level.

4. Reliability

The reliability of test scores pertains to the precision and consistency of the scores a test produces. Test scores can be influenced by errors stemming from various random factors. For instance, a student might perform sub-optimally due to poor sleep the previous night or score higher than their true ability would suggest due to sheer luck (e.g., guessing correctly on items). CTT formalizes this concept by separating test scores into two components: a true score and an error component (Equation 8):

$$X = T + E \tag{8}$$

where X represents the observed score (number of correct answers), T signifies the true score, and E denotes the error. A larger error implies larger variability of observed scores around the true score. The standard error of measurement (SEM) corresponds to the standard deviation of the observed scores around the true score. In other words, SEM quantifies the spread of the error term. The Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) recommends that in addition to SEM, conditional standard errors of measurement (CSEM) are reported for each score point. This is because measurement precision is not constant across the score scale, and modern measurement theories such as IRT highlight that a test will measure certain ranges of abilities more precisely than others. Therefore, we use CTT to report reliability at the test level and IRT to report measurement precision at different ability levels. Given that we do not have operational data, we analyze the reliability of CLT3-6 using the data obtained from the Spring 2023 field test. However, the field test forms are not analyzed as a whole – we only use the items that were not excluded from the item pool to ensure that they represent the items that will be administered in operational forms. The reliability of CLT7 and CLT8 is evaluated based on the CLT8 form used in the 2021 CLT8 Technical Report (CLT, 2021), considering it representative of the CLT8 item pool from which both the CLT7 and the CLT8 operational forms were created.

QUANTIFYING RELIABILITY

Reliability can be quantified as the proportion of observed score variance that is due to true score variance (Harvill, 1991):

$$r_{XX'} = \frac{s_T^2}{s_X^2} \tag{9}$$

where $r_{XX'}$ denotes the reliability of the test scores, s_T^2 is the variance of true scores, and s_X^2 is the variance of observed scores. This expression can be re-written as

$$r_{XX'} = 1 - \frac{s_E^2}{s_X^2} \tag{10}$$

where s_E^2 is the error variance. Thus, the error variance becomes $s_E^2 = s_X^2(1 - r_{XX'})$ and the SEM is:

$$SEM = s_E = \sqrt{s_X^2(1 - r_{XX'})} = s_X\sqrt{(1 - r_{XX'})} \tag{11}$$

The most commonly used reliability coefficient is Cronbach's alpha (Cronbach, 1951), which measures the internal consistency of a test by examining the covariance between the items (Tavakol & Dennick, 2011). Internal consistency is the degree to which the items in a test measure the same latent construct and are related to each other. The formula for Cronbach's alpha is given in Equation 12 (Bland & Altman, 1997):

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum s_i^2}{s_X^2}\right) \tag{12}$$

where k is the number of items in the test, s_i^2 is the variance of item i, and s_X^2 is the variance of the total number-correct scores. Cronbach's alpha is affected not just by the variances and covariances of items and total scores but by test length as well; adding similar items to a test form will increase alpha. Cronbach's alpha takes values between 0 and 1, and the psychometric literature has suggested acceptable values that range from 0.70 to 0.95 with no consensus on what value of alpha is "good" or "high" (Taber, 2018; Tavakol & Dennick, 2011).

Table 1 shows the ranges of Cronbach's alpha, sample sizes, and the number of items analyzed for each test. The tables show that each section of the CLT base form has high reliability for each analyzed group according to criteria often cited in the psychometric literature (Taber, 2018) as well as criteria used by state scholarships and education departments. For example, Florida's Tax Credit Scholarships require that students take a standardized test with internal consistency/reliability of at least 0.80 (Florida Department of Education, 2023), and Texas Education Agency considers a reliability coefficient between 0.80-0.89 as good (Texas Education Agency, 2022).

Table 1

The Reliability of CLT3-8

Section	Test	N Students	N Items	Alpha	SEM
VR	CLT3	498-550	73-77	0.92-0.92	3.44-3.55
QR	CLT3	511-577	58-60	0.91-0.92	3.15-3.28
VR	CLT4	487-554	75-76	0.91-0.91	3.53-3.73
QR	CLT4	497-574	56-59	0.91-0.92	2.97-3.11
VR	CLT5	458-526	72-76	0.89-0.90	3.51-3.76
QR	CLT5	465-534	54-58	0.91-0.91	3.06-3.14
VR	CLT6	431-496	70-75	0.88-0.90	3.59-3.74
QR	CLT6	430-499	55-59	0.90-0.91	3.10-3.22
VR	CLT7	1222	80	0.91	3.87
QR	CLT7	1222	40	0.81	2.26
VR	CLT8	1808	80	0.92	3.72
QR	CLT8	1808	40	0.85	2.74

TEST INFORMATION FUNCTION AND CONDITIONAL STANDARD ERRORS OF MEASUREMENT (CSEM)

The test information function (TIF) computes the amount of information a set of item responses provide about the latent ability parameter. The information provided by an item about the ability parameter depends on the item's difficulty, and is maximized when the item's difficulty matches the

student's ability. In the Rasch model, test information is simply the sum of the information provided by individual items (Equation 13):

$$I(\theta) = \sum_{i=1}^{k} P_i(1 - P_i) \tag{13}$$

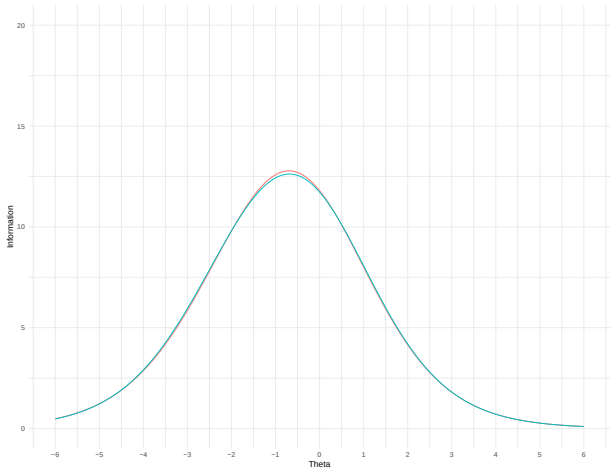
where P_i is the probability of a correct response to item i for a person with ability θ, and k is the total number of items. Test information determines the precision with which a student's ability is estimated by a given set of items. Specifically, test information is inversely related to SEM. Given that test information is a function of the proximity between the test's difficulty and a student's ability, a test will measure different abilities with different levels of precision. Equation 14 gives the SEM for a given ability level:

$$SEM(\theta) = \frac{1}{\sqrt{I(\theta)}} \tag{14}$$

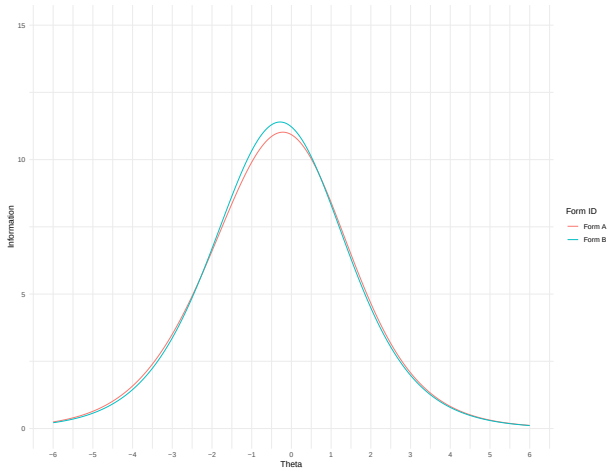
where $I(\theta)$ is the test information function as defined above. Figure 2 displays the TIF of the operational forms for the logit range [-6, 6]. Figure 3 shows the CSEM for the same ability range. IRT ability estimates have larger errors at the tails of the distribution than in the middle, which is reflected in Figure 3. Table 2-7 scale the CSEM of one of the operational forms from each test type by the slope of the scaling equation to express the CSEM on the 150-300 scale. The CSEM are displayed around 7 score points: 150, 175, 200, 225, 250, 275, 300. Not all these scale scores are obtainable in all the forms, so Tables 2-7 display the CSEM for the nearest score point.

Figure 2

The Test Information Functions (TIF) of the CLT3-8

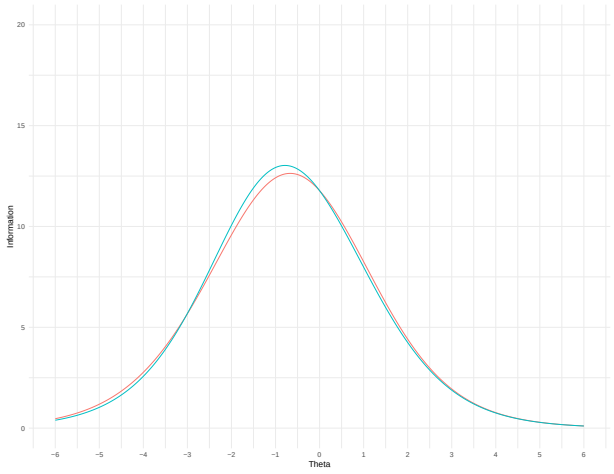


(a1) Verbal Reasoning

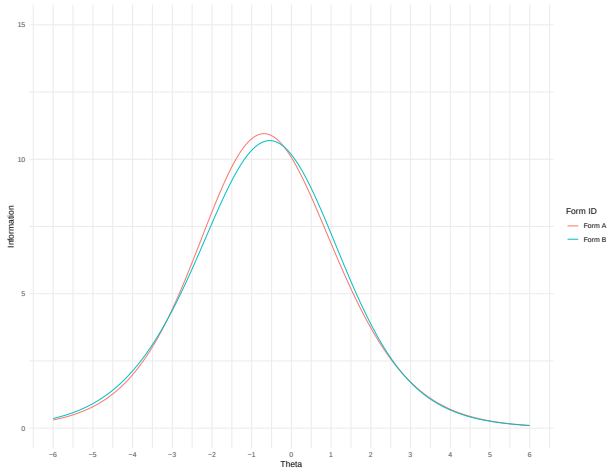


(a2) Quantitative Reasoning

(a) TIF of CLT3

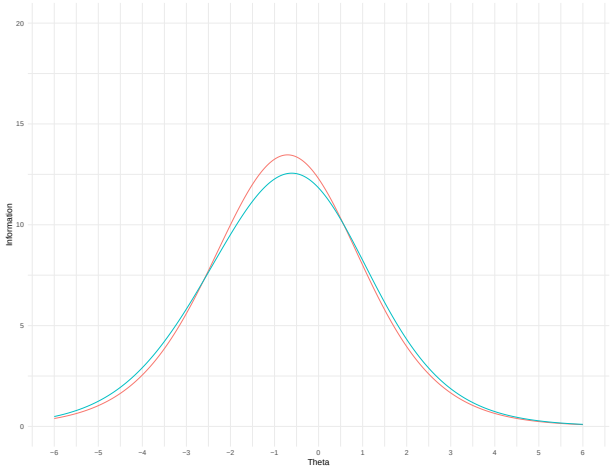


(b1) Verbal Reasoning

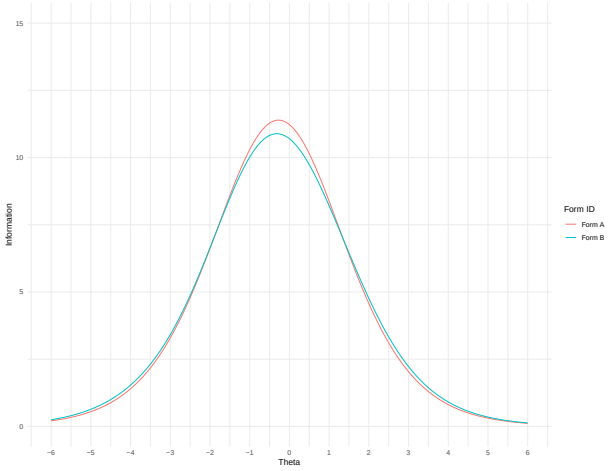


(b2) Quantitative Reasoning

(b) TIF of CLT4

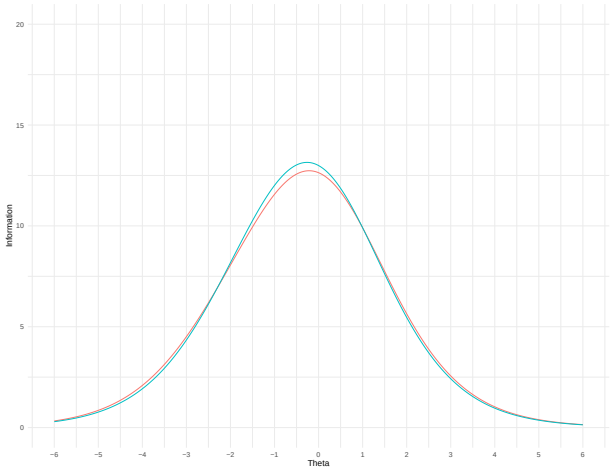


(c1) Verbal Reasoning

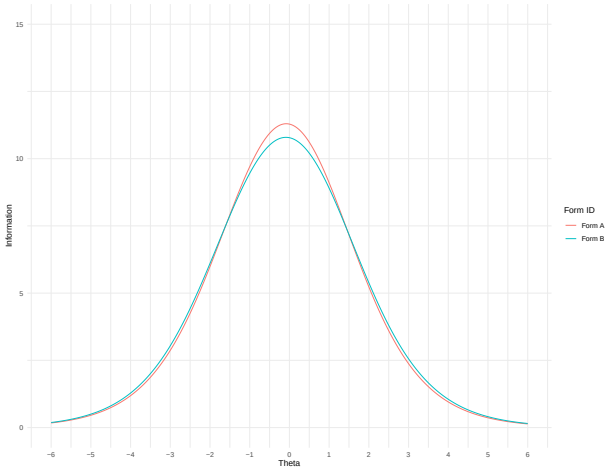


(c2) Quantitative Reasoning

(c) TIF of CLT5

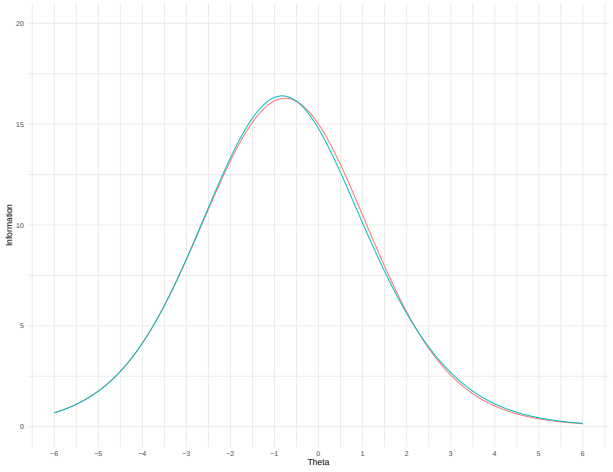


(d1) Verbal Reasoning

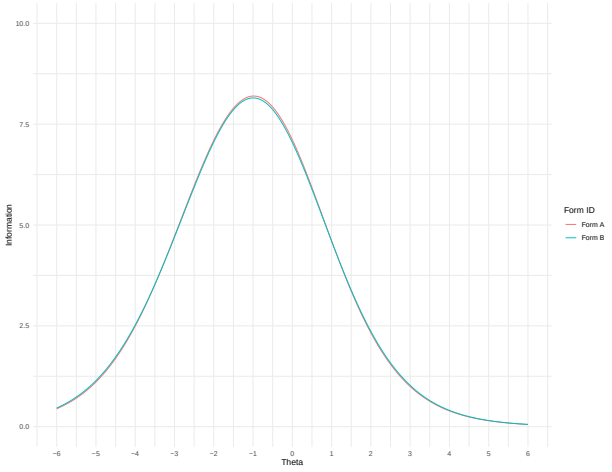


(d2) Quantitative Reasoning

(d) TIF of CLT6

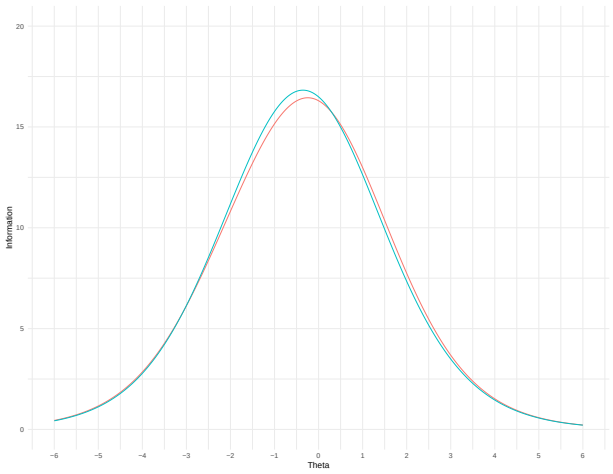


(e1) Verbal Reasoning

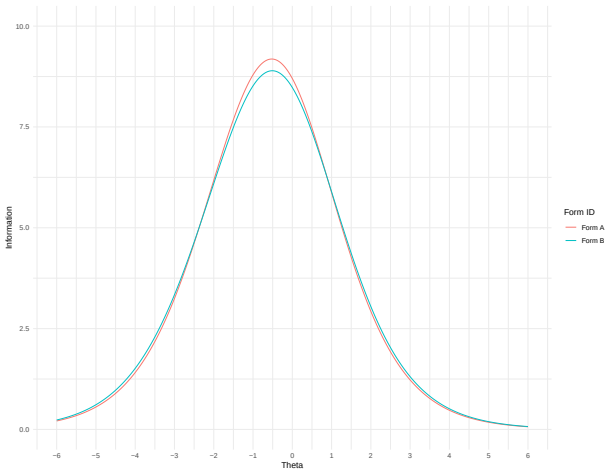


(e2) Quantitative Reasoning

(e) TIF of CLT7



(f1) Verbal Reasoning

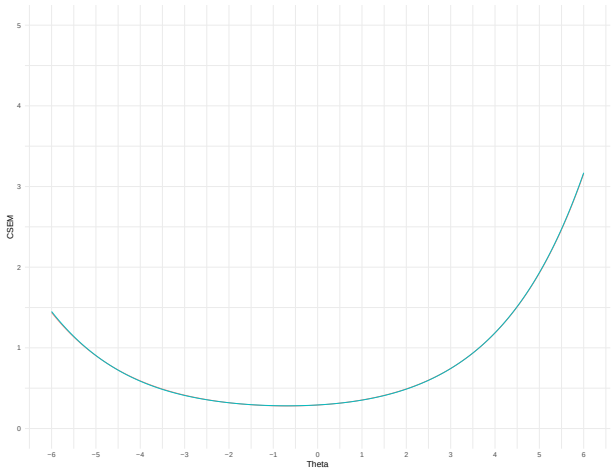


(f2) Quantitative Reasoning

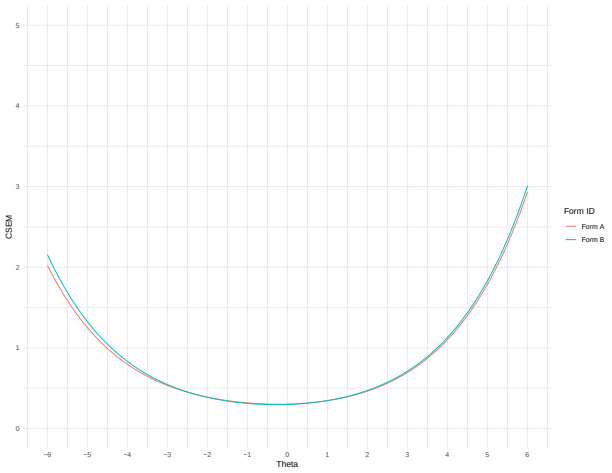
(f) TIF of CLT8

Figure 3

The Conditional Standard Error(s) of Measurement (CSEM) of the CLT3-8 (On the Logit Scale)

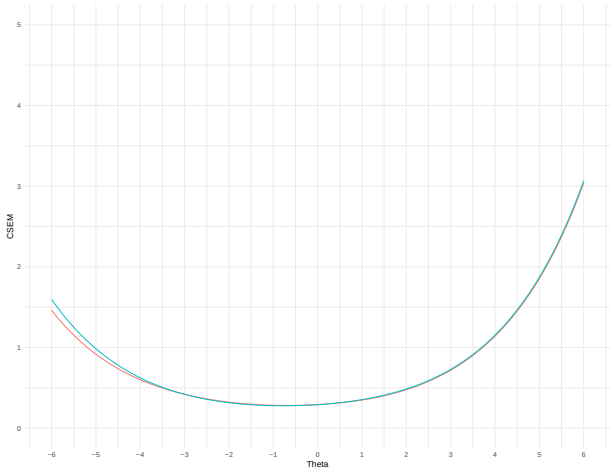


(a1) Verbal Reasoning

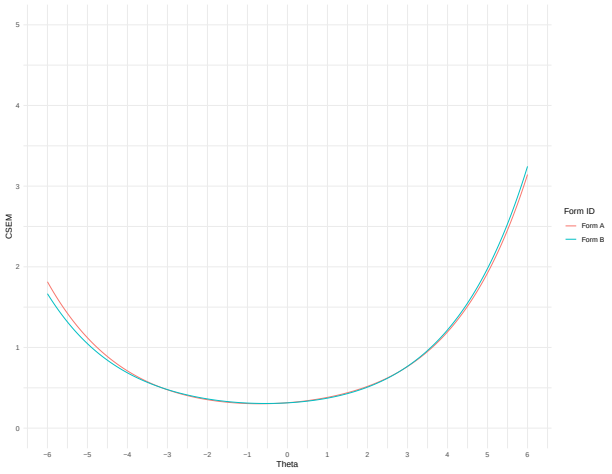


(a2) Quantitative Reasoning

(a) TIF of CLT3

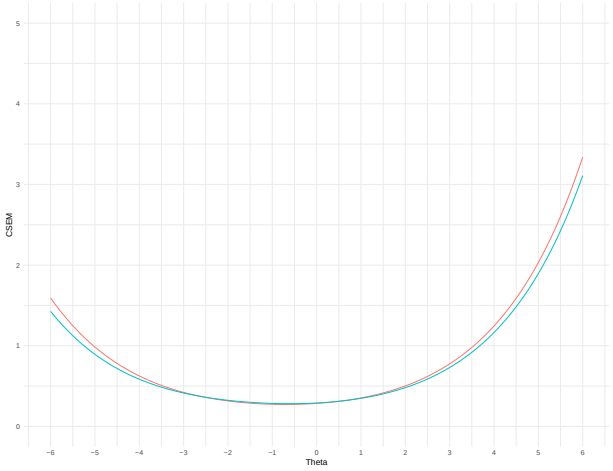


(b1) Verbal Reasoning

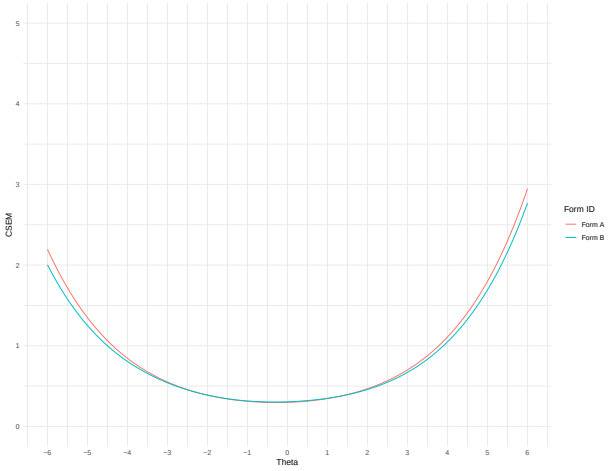


(b2) Quantitative Reasoning

(b) TIF of CLT4

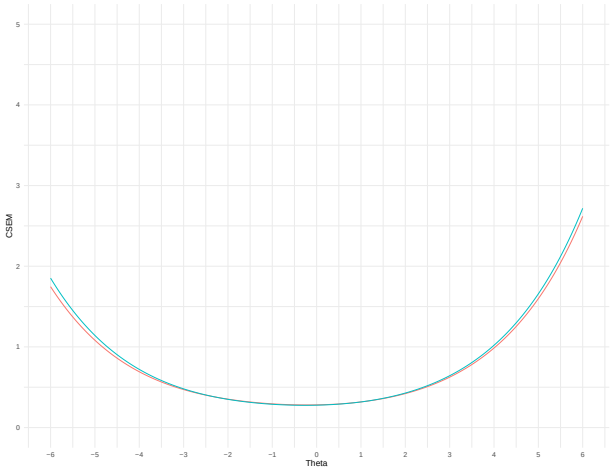


(c1) Verbal Reasoning

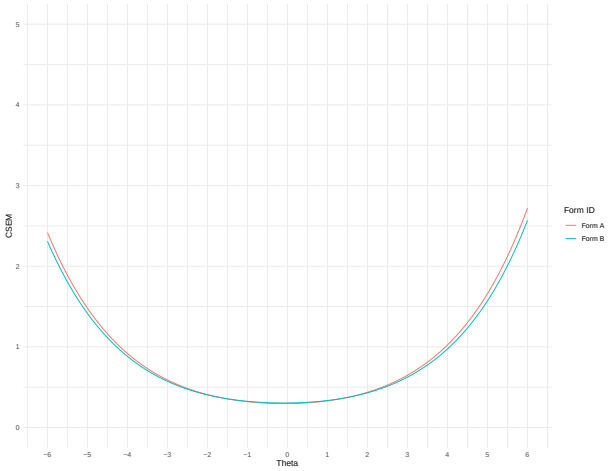


(c2) Quantitative Reasoning

(c) TIF of CLT5

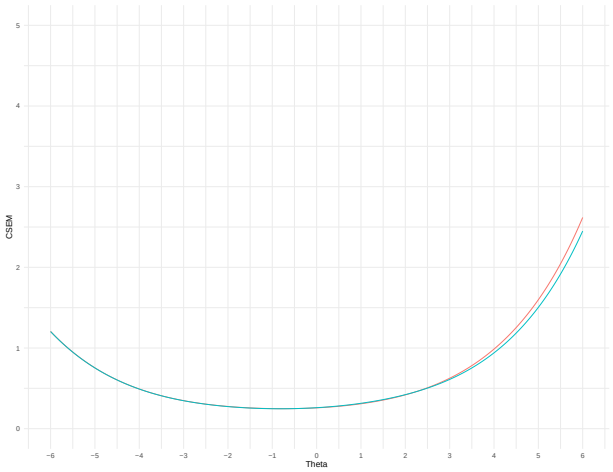


(d1) Verbal Reasoning

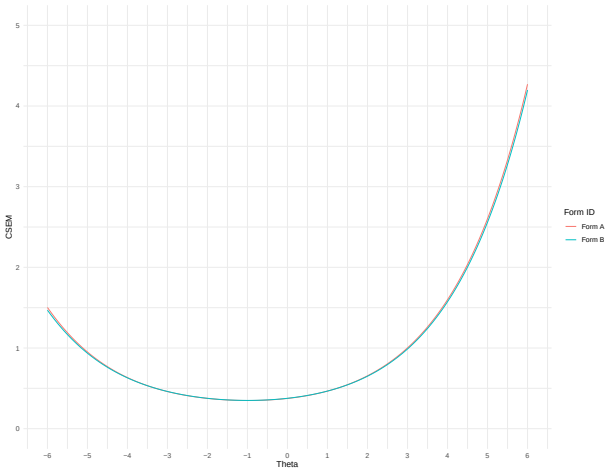


(d2) Quantitative Reasoning

(d) TIF of CLT6

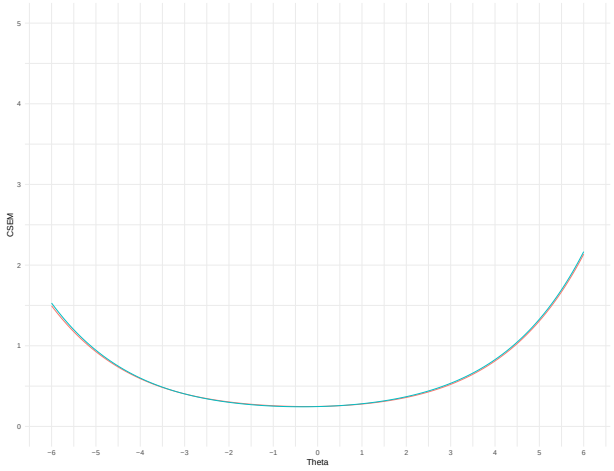


(e1) Verbal Reasoning

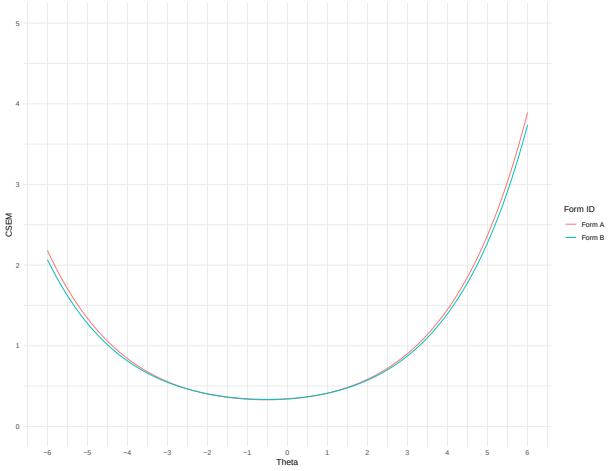


(e2) Quantitative Reasoning

(e) TIF of CLT7



(f1) Verbal Reasoning



(f2) Quantitative Reasoning

(f) TIF of CLT8

Table 2

The CSEM of CLT3 Scale Scores

Scale Score	Verbal Reasoning SEM	Quantitative Reasoning SEM
150 (VR) – 150 (QR)	16	16
174 (VR) – 175 (QR)	9	10
200 (VR) – 199 (QR)	6	7
226 (VR) – 225 (QR)	5	6
251 (VR) – 251 (QR)	6	6
276 (VR) – 276 (QR)	9	10
300 (VR) – 300 (QR)	16	16

Table 3

The CSEM of CLT4 Scale Scores

Scale Score	Verbal Reasoning SEM	Quantitative Reasoning SEM
150 (VR) – 150 (QR)	13	16
174 (VR) – 173 (QR)	8	9
200 (VR) – 200 (QR)	6	6
224 (VR) – 225 (QR)	5	6
251 (VR) – 251 (QR)	7	7
274 (VR) – 276 (QR)	10	11
300 (VR) – 300 (QR)	16	22

Table 4

The CSEM of CLT5 Scale Scores

Scale Score	Verbal Reasoning SEM	Quantitative Reasoning SEM
150 (VR) – 150 (QR)	12	16
176 (VR) – 177 (QR)	7	9
200 (VR) – 200 (QR)	5	6
224 (VR) – 225 (QR)	5	6
250 (VR) – 250 (QR)	7	7
273 (VR) – 277 (QR)	11	11
300 (VR) – 300 (QR)	21	22

Table 5

The CSEM of CLT6 Scale Scores

Scale Score	Verbal Reasoning SEM	Quantitative Reasoning SEM
150 (VR) – 150 (QR)	16	16
177 (VR) – 176 (QR)	9	9
200 (VR) – 201 (QR)	6	6
224 (VR) – 225 (QR)	5	6
251 (VR) – 251 (QR)	6	7
275 (VR) – 275 (QR)	8	11
300 (VR) – 300 (QR)	16	21

Table 6

The CSEM of CLT7 Scale Scores

Scale Score	Verbal Reasoning SEM	Quantitative Reasoning SEM
150 (VR) – 150 (QR)	12	11
176 (VR) – 176 (QR)	7	8
200 (VR) – 200 (QR)	5	7
225 (VR) – 226 (QR)	5	7
250 (VR) – 252 (QR)	5	10
276 (VR) – 281 (QR)	8	16
300 (VR) – 300 (QR)	12	22

Table 7

The CSEM of CLT8 Scale Scores

Scale Score	Verbal Reasoning SEM	Quantitative Reasoning SEM
150 (VR) - 150 (QR)	12	13
176 (VR) - 174 (QR)	8	8
199 (VR) - 199 (QR)	5	6
225 (VR) - 225 (QR)	5	7
249 (VR) - 249 (QR)	5	10
275 (VR) - 277 (QR)	7	16
300 (VR) - 300 (QR)	12	22

5. Norm Referencing

In the summer of 2023, CLT conducted a linking study with MetaMetrics, the developer of the Lexile® Framework for Reading and the Quantile® Framework for Mathematics. The reports of this study (MetaMetrics, 2023) can be found on our website. As a result of this study, the Verbal Reasoning scales of the CLT3-8 suite of assessments were linked to Lexile® reading measures and the Quantitative Reasoning scales were linked to Quantile® measures. This link allows CLT to report the national user norms established by MetaMetrics, which were derived in a research study of 3 million students across the United States who took assessments that report Lexile® and Quantile® measures between 2010-2019 (MetaMetrics, 2023). Although the participants in the study were self-selected, they are expected to represent a broad spectrum of the general student population in the United States due to the fact that the assessments that report Lexile® and Quantile® measures include common benchmarks of academic performance such as the ERB®, Stanford Achievement Test 10®, The Iowa Tests of Basic Skills®, and the State of Texas Assessments of Academic Readiness® (STARR).

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association.

Becker, B., Debeer, D., Sachse, K. A., & Weirich, S. (2021). Automated Test Assembly in R: The eatata package. Psych, 3(2), 96–112. https://doi.org/10.3390/psych3020010

Bland, J. M., & Altman, D. G. (1997). Statistics notes: Cronbach's alpha. BMJ, 314(7080), 572–572. https://doi.org/10.1136/bmj.314.7080.572

Classic Learning Test. (2021). 2021 Technical Report: The Classic Learning Test for Seventh and Eighth Grade. https://www.cltexam.com/wp-content/uploads/2022/04/210915_TechnicalReport_v4.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16(3), 297–334. https://doi.org/10.1007/bf02310555

Crocker, L. M., & Algina, J. (2008). Introduction to classical and modern test theory. Cengage learning.

Florida Department of Education. (2023). Annual assessment requirement. Retrieved from https://www.fldoe.org/schools/school-choice/k-12-scholarship-programs/ftc/annual-assessment-requirement.stml

Harvill, L. M. (1991). Standard error of measurement. Educational Measurement: Issues and Practice, 10(2), 33–41. https://doi.org/10.1111/j.1745-3992.1991.tb00195.x

Kolen, M. J., & Brennan, R. L. (2004). Test equating, scaling, and linking: Methods and practices. Springer.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking: Methods and practices. Springer.

Linacre, J. M. (2002). What do Infit and Outfit, Mean-square and Standardized mean? Rasch Measurement Transactions, 16(2), p.878 https://doi.org/10.1177/0013164499594004

Linacre, J. M. (2023). Winsteps® Rasch measurement computer program (Version 5.6.0). Portland, Oregon: Winsteps.com

Ludlow, L. H., & O'Leary, M. (1999). Scoring omitted and not-reached items: Practical Data Analysis Implications. Educational and Psychological Measurement, 59(4), 615–630.

https://doi.org/10.1177/00131649921970053

MetaMetrics. (2023, November). Linking the Classic Learning Test of Verbal Reasoning to the Lexile® Framework for Reading: Linking Study Report. Classic Learning Initiatives. https://www.cltexam.com/wp-content/uploads/2024/01/Classic-Learning-Test-of-Verbal-Reasoning-Linking-Study-Report-November-2023_Redacted.pdf

MetaMetrics. (2023, November). Linking the Classic Learning Test of Quantitative Reasoning to the Quantile® Framework for Mathematics: Linking Study Report. Classic Learning Initiatives. https://www.cltexam.com/wp-content/uploads/2024/01/Classic-Learning-Test-of-Quantitative-Reasoning-Linking-Study-Report-November-2023_Redacted.pdf

O'Neill, T., Peabody, M., Tan, R. J. B., & Du, Y. (2013). How much item drift is too much? Rasch Measurement Transactions, 27(3), 1423-1424.

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.https://www.R-project.org/.

Taber, K. S. The Use of Cronbach's Alpha When Developing and Reporting Research Instruments in Science Education. Res Sci Educ 48, 1273–1296 (2018). https://doi.org/10.1007/s11165-016-9602-2

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Texas Education Agency. (2022). Standard Technical Processes. In 2021-2022 Technical Digest (Chapter 3). Retrieved from https://tea.texas.gov/student-assessment/testing/2021-2022-technical-digest-chapter-3.pdf

Willse. J. T. (2018). CTT: Classical Test Theory Functions. R package version 2.3.3, https://CRAN.R-project.org/package=CTT.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, IL: MESA Press