Center for Advanced Studies in Measurement and Assessment

CASMA Research Report

Number 58

Empirical Investigation of Various Reliability Statistics

Won-Chan Lee
Deborah J. Harris
Huan Liu
Kuo-Feng Chang[†]

December 2025

[†]Won-Chan Lee is Professor and Director, Center for Advanced Studies in Measurement and Assessment (CASMA), 300D Lindquist Center, College of Education, University of Iowa, Iowa City, IA 52242 (email: won-chan-lee@uiowa.edu). Deborah J. Harris is a retired Visiting Professor in Educational Measurement and Statistics at the University of Iowa (email: deborah-harris@uiowa.edu). Huan Liu is Psychometrician at the Riverside Insights (email: huan.liu@riversideinsights.com). Kuo-Feng Chang is Research Scientist at the Human Resources Research Organization (email: kchang@humrro.org).

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

This report presents two illustrative examples to demonstrate the application of reliability statistics derived from classical test theory, generalizability theory, and item response theory, as discussed in the *Reliability in Educational Measurement* chapter of the *5th edition of Educational Measurement* (Lee & Harris, 2025). The first example focuses on a single-domain assessment, while the second involves a composite score derived from a multidomain assessment. For each example, multiple reliability coefficients, conditional standard errors of measurement, and classification consistency and accuracy indices are estimated across raw, scale, and proficiency score metrics. Empirical results from constructed parallel forms serve as benchmarks for evaluating the accuracy of single-administration estimates.

1 Introduction

The primary purpose of this report is to present computational examples that illustrate the theoretical frameworks discussed in the *Reliability in Educational Measurement* chapter (hereafter referred to as *Reliability*) of the *5th edition of Educational Measurement* (Lee & Harris, 2025, pp. 277–381). That chapter introduces the concepts, quantification, and estimation of reliability statistics within three major measurement frameworks: classical test theory (CTT), generalizability theory (GT), and item response theory (IRT). Serving as a companion piece, this report applies the chapter's formulas to empirical data to demonstrate how estimates of reliability statistics differ across these three models.

Two illustrative examples are presented. The first is based on an assessment targeting a single content domain, while the second involves a multi-domain assessment that yields composite scores. Both examples use data from large-scale, multiple-choice testing programs. References to formulas used throughout the report correspond to the equation and page numbers provided in the *Reliability* chapter.

To facilitate comparisons of reliability estimates against a replication benchmark, each original assessment and its associated examinee responses were split to create two statistically and content-balanced half forms. These "half forms" function as pseudo-parallel forms, allowing examinee scores across the two forms to serve as a basis for evaluating score consistency. Importantly, the variability in examinee scores across forms reflects differences in item sampling, as other typical sources of replication error were held constant. All responses were obtained during a single testing session, and the division into half forms was designed to preserve both content coverage and statistical properties, rather than relying on item position. Consequently, factors such as examinee proficiency, testing occasion, motivation, fatigue, and other contextual influences remained consistent across the two forms. This approach ensures that the observed score differences primarily reflect form-related variation, rather than additional sources of measurement error that would typically arise in operational replications (e.g., changes in examinee motivation or proficiency between test administrations).

For each of the two examples, a range of reliability coefficients, conditional standard errors of measurement (CSEMs), and classification consistency and accuracy indices are presented. Results are reported across three scoring metrics: number-correct raw scores, IRT proficiency estimates, and scale scores. These examples are intended to highlight both the similarities and differences among various reliability statistics and to emphasize the importance of clearly specifying how each reliability estimate was derived, rather than simply reporting a single value without context.

2 Illustrative Examples

2.1 Example One: Single Domain

The original dataset consisted of an 80-item multiple-choice assessment with two content categories and administered to several thousand examinees. Two 40-item forms, Form A and Form B, were constructed by dividing the original 80 items into statistically and content-balanced halves. Disattenuated correlations across the two content areas for Form A and Form B were .996 and .992, respectively, supporting the single domain premise. Table 1 presents the average biserial correlations and p-values for both forms, calculated using the full examinee sample. The observed differences fall well within the expected range for alternate forms of a 40-item test. A random sample of 3,000 examinees was drawn from the entire pool of examinees and used to compute the reliability statistics reported in the subsequent tables and figures.

Table 1: P-values and Biserial Correlations for Single Domain Example

Form	*Biserial*	P-value
Form A	.444	.568
Form B	.432	.545

The item responses from the sample of 3,000 examinees were also used to estimate two-parameter logistic (2PL) item parameters for all items in both forms. Item calibrations were conducted separately for each form using *flexMIRT* (Houts & Cai, 2016). Because the same group of examinees was used for both calibrations, the resulting item parameter estimates were on a common theta scale, eliminating the need for a separate scale transformation. Using the calibrated 2PL item parameters, maximum likelihood (ML) and expected a posteriori (EAP) theta estimates were obtained for each examinee.

To equate Form B to Form A, a single-group design was employed. Equipercentile equating with postsmoothing was conducted using *RAGE-RGEQUATE* (Zeng, Kolen, Hanson, Cui, & Chien, 2005), with a smoothing parameter of $S = .05$. This process yielded non-integer raw scores on Form A corresponding to integer raw scores on Form B. The resulting non-integer scores were then rounded and truncated to produce equated, rounded raw scores.

In addition, a score scale ranging from 100 to 130 was developed for Form A. The raw-to-scale score conversion was established by first generating the relative frequency distribution of raw scores, which was then smoothed using the polynomial loglinear method. Percentile ranks were computed from the smoothed distribution, and corresponding z-scores were determined for each integer raw score. These z-scores were subsequently transformed to have a mean of 118 and a standard deviation of 4. The resulting values were then rounded and truncated to produce the final scale scores. Scale scores for Form

B were obtained by equating Form B to Form A. For both Form A and Form B, scale scores based on ML and EAP theta estimates were derived using the test characteristic curve method in conjunction with the Form A raw-to-scale score conversion.

To evaluate the classification consistency and accuracy indices, cut scores were established. For Form A, the cut score was set at a scale score of 120, corresponding to a passing rate of approximately 30%. The associated cut scores on the number-correct and ML/EAP theta metrics were determined by back-mapping from the scale score using the previously established raw-to-scale score and raw-to-theta relationships. This resulted in a cut score of 29 on the raw score metric and 0.375 on the theta metric. For Form B, the cut score on the number-correct metric was obtained through the equipercentile equating process and was determined to be 27. The cut scores on the theta and scale score metrics for Form B remained identical to those for Form A, as both metrics were on a common scale across forms, in contrast to the raw score metric, which differed by form.

2.1.1 Reliability Coefficients

Table 2 presents reliability coefficients for the single-domain example across all three score metrics. For each metric, scores from the 3,000 examinees on Form A and Form B were correlated, with these values reported in the final column labeled "Replication." Although based on only two forms, these correlations align with the conceptualization of reliability as the consistency of examinee scores across replications. While not a formal criterion, they serve as a useful benchmark for interpreting the other reliability coefficients.

Detailed explanations of these reliability coefficients are provided in the sections titled "Reliability in CTT," "Reliability in GT," and "Reliability in IRT" within the *Reliability* chapter. The corresponding symbols and references for the formulas used in this analysis are listed in Table 3, with citations to the relevant equations or page numbers from the chapter.

Across all score metrics and reliability coefficients, the values obtained for Form A and Form B are highly similar and generally align closely with the replication correlation. For the raw score metric, the values of coefficient alpha and the generalizability coefficient based on the $p \times i$ design are identical. This result is expected and highlights the fact that coefficient alpha can be derived under either the assumption of essentially tau-equivalent forms or randomly parallel forms.

Feldt's coefficient is based on the classical congeneric model, while the unidimensional IRT (UIRT) method assumes strictly parallel forms. As discussed in the CTT section of the *Reliability* chapter, Feldt's coefficient is always larger than coefficient alpha; in this example, however, the difference is minimal. The UIRT method yields slightly higher reliability coefficients than the others. This is expected, as the score variability under the assumption of strictly parallel forms is typically smaller than under less restrictive parallel-form assumptions. Overall, the single-administration reliability estimates for both forms,

Table 2: Reliability Coefficients for Single Domain Example

Score Types Reliability Coefficients	Form A	Form B	Replication
Raw Scores			
Coefficient alpha	.874	.870	.874
Feldt's coefficient	.875	.871	
GT $p \times i$	.874	.870	
UIRT	.879	.877	
UIRT Proficiency Estimates			
ML	.799/.881/.870	.776/.884/.870	.857
EAP	.876/.874	.875/.873	.879
Rounded Scale Scores			
UIRT	.864	.867	.859
ML – Polynomial method	.875/.885/.884	.891/.881/.882	.872
EAP – Polynomial method	.876/.885/.884/.886	.842/.884/.879/.884	.879

Table 3: Reliability Coefficients for Single Domain Example: Equations and Page References from the *Reliability* Chapter

Score Types Reliability Coefficients	Symbol	Equation/Page No. (*Reliability* Chapter)
Raw Scores		
Coefficient alpha	${}_{\alpha}R(X)$	Eq. 12
Feldt's coefficient	${}_{F}R(X)$	Eq. 14
GT $p \times i$	$\mathscr{E}\rho^2$	Eq. 25
UIRT	$R(X)_{\text{IRT}}$	Eq. 54
UIRT Proficiency Estimates		
ML	$R(\widehat{\theta})_{\text{VR}}/R(\widehat{\theta})_{\text{EVR}}/R(\widehat{\theta})_{\text{MVR}}$	p. 311
EAP	$R(\tilde{\theta})_{\text{VR}}/R(\tilde{\theta})_{\text{INF}}$	p. 316
Rounded Scale Scores		
UIRT	$R(S)_{\text{IRT}}$	Eq. 58
ML – Polynomial method	$R(S_{\widehat{\theta}})_{\text{VR}}/R(S_{\widehat{\theta}})_{\text{EVR}}/R(S_{\widehat{\theta}})_{\text{MVR}}$	p. 323
EAP – Polynomial method	$R(S_{\tilde{\theta}})_{\text{VR}}/R(S_{\tilde{\theta}})_{\text{INF}}$	Eq. 70 / p. 323

computed under four different parallel-form assumptions, closely approximated the replication-based reliability for raw scores.

While the four raw score methods can be directly compared to one another, comparisons between methods using theta and scale score metrics should not be made within a single metric. The two theta-based methods differ in how item responses are scored (i.e., ML versus EAP), but both rely on the same theta-to-scale score conversion.

Three reliability estimates are reported for the ML scores, and two for the EAP scores. These values differ because they are derived using different reliability formulas. Specifically, ML-based results were calculated using the coefficients $R(\widehat{\theta})_{\mathrm{VR}}$, $R(\widehat{\theta})_{\mathrm{EVR}}$, and $R(\widehat{\theta})_{\mathrm{MVR}}$, while the EAP-based results were computed using $R(\tilde{\theta})_{\mathrm{VR}}$ and $R(\tilde{\theta})_{\mathrm{INF}}$. The relatively large discrepancies among the ML-based coefficients—particularly between $R(\widehat{\theta})_{\mathrm{VR}}$ and the other two—suggest that the contribution of estimation bias is *not* negligible. As discussed in the *Reliability* chapter, $R(\widehat{\theta})_{\mathrm{EVR}}$ is recommended for practical use. It is important to recall that all three ML-based coefficients assume strictly parallel forms. The replication result, a correlation of .857, was derived using two forms with different sets of item parameters—that is, the forms were not strictly parallel—resulting in a lower value than the $R(\widehat{\theta})_{\mathrm{EVR}}$ estimates. This highlights that $R(\widehat{\theta})_{\mathrm{EVR}}$ can overestimate reliability when conceptual replications involve using different sets of items, which is one of the most prominent sources of measurement error. Notably, the replication correlation is not equivalent to $R(\widehat{\theta})_{\mathrm{PF}} = \rho(\widehat{\theta}, \widehat{\theta}')$, as the latter still assumes identical item parameters across forms.

In contrast, for the EAP estimator, both $R(\tilde{\theta})_{\mathrm{VR}}$ and $R(\tilde{\theta})_{\mathrm{INF}}$ are very similar to each other and closely align with the replication result. The information-based estimate $R(\tilde{\theta})_{\mathrm{INF}}$ approximated the conventional Bayesian result remarkably well.

For rounded scale scores, results based on the unidimensional IRT (UIRT) method were obtained using the reliability coefficient $R(S)_{\mathrm{IRT}}$, derived from the raw-to-rounded scale score conversion table for each form. The scale score results for the ML and EAP estimators followed the procedures described in the "Scale Scores Transformed From Proficiency Estimates" section of the *Reliability* chapter.

To produce a continuous, differentiable transformation function for the ML estimator, polynomial models were fitted to the theta-to-rounded scale score conversion tables. Polynomial degrees of 4 and 7 were used for Form A and Form B, respectively. The fitted polynomial functions were then applied to each of the 41 quadrature points from a standard normal distribution to generate "true" *unrounded* scale scores, which were aggregated over the normal density to obtain the true score variance on the scale score metric.The error variance for scale scores was computed by averaging the squared scale score CSEMs, calculated using the first derivative of the polynomial function and the squared CSEMs for $\widehat{\theta}$ at each quadrature point. Observed scale score variance was estimated by transforming each examinee's ML estimate using the fitted polynomial function and rounding the resulting scale score. The reliability coefficients $R(S_{\widehat{\theta}})_{\mathrm{VR}}$,

$R(S_{\widehat{\theta}})_{\text{EVR}}$, and $R(S_{\widehat{\theta}})_{\text{MVR}}$ (scale score analogues of $R(\widehat{\theta})_{\text{VR}}$, $R(\widehat{\theta})_{\text{EVR}}$, and $R(\widehat{\theta})_{\text{MVR}}$) are reported in the table.

A similar approach was used for the EAP estimates. Polynomial functions with degrees 4 and 7 for Forms A and B, respectively, were fitted to the theta-to-rounded scale score conversion. The scale score error variance was computed by averaging each examinee's squared scale score CSEMs, derived using the CSEMs for $\tilde{\theta}$ and the first derivative of the fitted polynomial. The true and observed score variances were computed following the same procedures used for the ML estimator. Results are reported for three variants of the $R(S_{\tilde{\theta}})_{\text{VR}}$ coefficient, as defined in Equation 70 of the *Reliability* chapter, along with the information-based coefficient $R(S_{\tilde{\theta}})_{\text{INF}}$. Unlike the results for $\tilde{\theta}$, the combination of polynomial approximation and rounding introduced meaningful differences across the three variants of $R(S_{\tilde{\theta}})_{\text{VR}}$.

2.1.2 Conditional Standard Error of Measurement

Various estimators of CSEMs are discussed in the section titled "Estimators of CSEMs" in the *Reliability* chapter. References to the corresponding equation and page numbers are provided throughout the main text of this report.

To avoid redundancy, results for CSEMs are presented for Form A only. Figure 1 displays CSEM plots for several raw score methods, including: Lord's method (Equation 78); Keats' correction (p. 327); the GT approach with absolute error variance (Equation 82); the GT approach with relative error variance (Equation 83); the UIRT approach (square root of Equation 52); and empirical estimates (Equation 71).

The GT-based estimates are expressed on the total score scale, obtained by multiplying the mean-score CSEMs from Equations 82 and 83 by the number of items. For the UIRT approach, the version based on a quadrature standard normal distribution—referred to as the D-method (p. 346)—was used.

As shown in the figure, the empirical CSEM estimates derived from Forms A and B tend to exhibit a bumpy pattern, though their overall shape is similar to that of the other methods, generally forming an inverted-U curve. Lord's method, which involves absolute error, tends to yield larger CSEMs compared to the GT approach using relative error, whose results are closely aligned with those obtained using Keats' correction. The UIRT method, which assumes strictly parallel forms, produces the smallest CSEM values across much of the score scale.

Figure 2 presents the estimated CSEMs for rounded scale scores. The estimation methods include the binomial procedure (Equation 93), the polynomial procedure (Equation 87) with degree 3, and the UIRT approach (square root of Equation 57). Across all three methods, the resulting CSEM curves exhibit an elongated "M" shape, with the UIRT method generally producing smaller error estimates than the other two. These patterns reflect a combination of the underlying raw score CSEMs and the rate of change (i.e., slope) in the raw-to-rounded scale score conversion function (not shown here). Among these two factors, the slope pattern typically has the greater influence on the resulting

scale score CSEMs. It is important to note that the horizontal axis in Figure 2 represents different metrics across methods: for the UIRT procedure, it reflects "true" (i.e., expected) scale scores, while for the binomial and polynomial methods, it reflects observed scale scores.

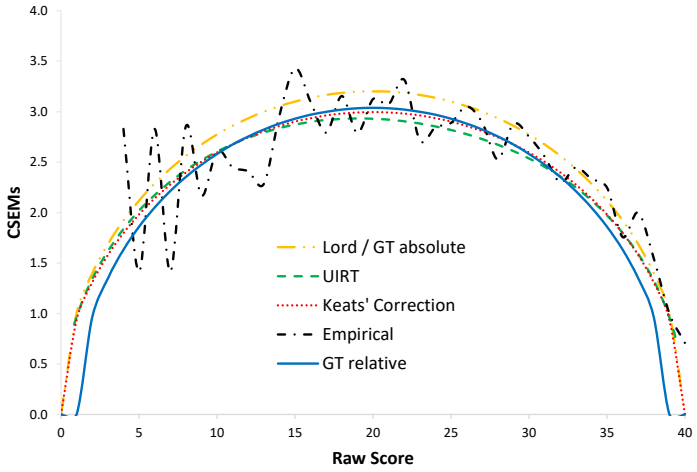

Figure 1: CSEMs for number-correct raw scores.

Figure 3 displays the CSEMs for the ML and EAP estimates. For the ML estimator, the CSEMs were computed using the estimator $\sqrt{1/I(\theta, \widehat{\theta})}$ (p. 329). For the EAP estimator, CSEMs were obtained using the square roots of Equations 44 and 48, representing the conventional and information-based approaches, respectively. Results for both ML and EAP estimates exhibit the expected U-shaped pattern, with larger errors at the lower and upper ends of the proficiency scale.

For the ML and information-based EAP estimates, CSEMs were computed at 41 quadrature points ranging from –5 to 5. In contrast, the conventional EAP CSEMs were estimated individually for each of the 3,000 examinees, whose EAP scores fell within a narrower, data-driven range—roughly within ±3. As expected, the ML estimator yields consistently larger CSEMs across the theta continuum compared to the EAP estimator. Additionally, the two EAP-based CSEM estimates are nearly indistinguishable from one another.

Figure 4 presents the CSEMs for scale scores derived from ML and EAP proficiency estimates. A degree-4 polynomial model was fitted to the theta-to-rounded scale score conversion for both estimators. The resulting CSEMs show close agreement in the center of the scale score distribution but diverge toward the tails. The overall pattern of scale score CSEMs reflects the slope of the transformation function used to convert theta to scale scores.

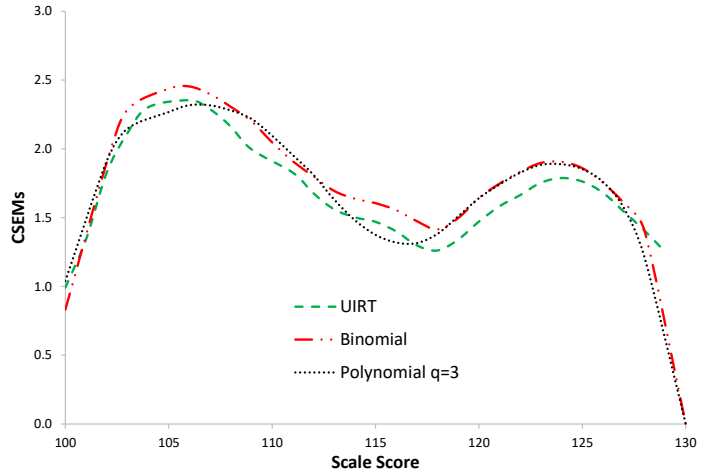

Figure 2: CSEMs for scale scores.

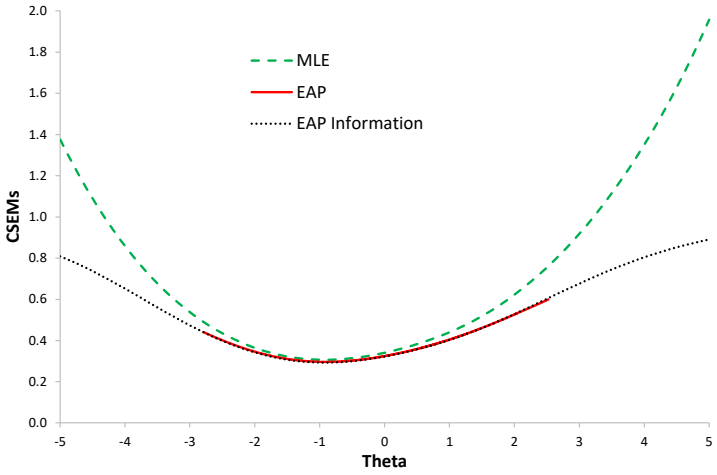

Figure 3: CSEMs for ML and EAP estimates.

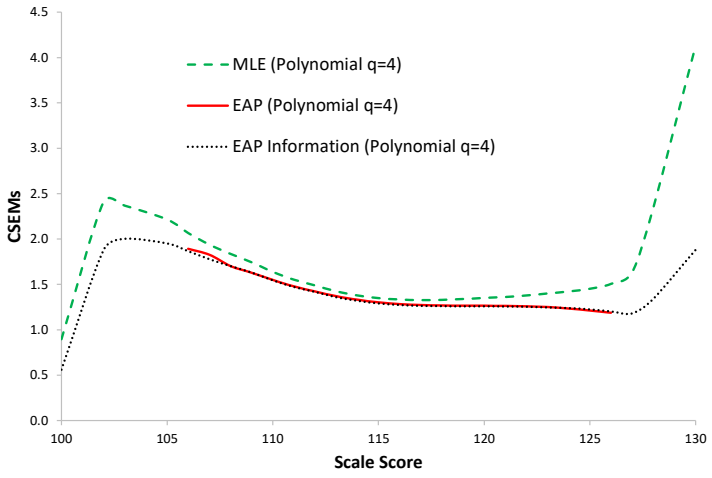

Figure 4: CSEMs for scale scores using polynomial method.

2.1.3 Classification Consistency and Accuracy Indices

The procedures outlined in the "Reliability of Classification Category Scores" section of the *Reliability* chapter were used to obtain the classification consistency and accuracy indices reported in this study. Table 4 presents estimates of the marginal agreement index ϕ for each of the two forms, along with replication-based estimates for comparison. The replication values were derived from 2×2 contingency tables, with the sum of the diagonal cells—representing consistent classifications (i.e., examinees classified as either above or below the cut score on both forms)—reported in the table.

The binary classification cut scores were 29 for Form A and 27 for Form B on the raw score metric, based on the equating results. These raw cut scores corresponded to a scale score of 120 for both forms, which served as the cut score on the scale score metric. The corresponding cut score on the theta metric was 0.375. For methods requiring a reliability estimate, such as the normal approximation and Livingston–Lewis methods, coefficient alpha was used in this example (.874 for Form A and .870 for Form B), although other reliability coefficients based on absolute error variance could also be used.

As shown in the table, all methods produced different but generally comparable results across scoring metrics. The two-parameter beta-binomial model yielded slightly lower agreement values than the other methods, possibly suggesting some degree of model misfit. Notably, all single-administration estimates were smaller than the replication-based values.

Table 5 summarizes the results for the classification accuracy index γ, based on the same set of procedures used to estimate classification consistency. Replication-based values are not reported in this table, as classification accuracy relies on

Table 4: Classification Consistency for Single Domain Example

Score Types and Procedures	Form A	Form B	Replication
Raw Scores (Cut = 29 for Form A and 27 for Form B)			
Two-parameter beta binomial model	.845	.837	.870
Four-parameter beta binomial model	.853	.853	
UIRT	.857	.859	
Normal approximation	.865	.854	
ML Estimates (Cut = .375 for both forms)			
Rudner's Method	.864	.864	.880
Rounded Scale Scores (Cut = 120 for both forms)			
Livingston–Lewis Procedure (with two-parameter beta binomial model)	.854	.847	.869
Livingston–Lewis Procedure (with four-parameter beta binomial model)	.865	.856	
UIRT	.857	.859	
Normal approximation	.859	.856	

the joint distribution of observed and true scores, or equivalently, a contingency table comparing observed and true classifications, which cannot be directly obtained from two operational replications.

Overall, the results in Table 5 are highly consistent across estimation methods and score types. As expected, the accuracy indices reported in Table 5 are generally higher than the consistency indices presented in Table 4.

2.2 Example Two: Composite Score

The original dataset for the composite score example consisted of multiple-choice items spanning three content domains. The items were split to create two parallel forms, Form A and Form B. A small number of items were removed to ensure that both forms contained an equal number of items in each domain and were comparable in terms of both statistical and content specifications. Each form included 13 items in Subcontent 1, 12 items in Subcontent 2, and 6 items in Subcontent 3.

Table 6 presents the average biserial correlations and p-values for the two forms, based on the full sample of examinees. As shown, the differences fall well within the range typically expected for alternate forms of a 31-item assessment.

Table 5: Classification Accuracy for Single Domain Example

Score Types and Procedures	Form A	Form B
Raw Scores (Cut = 29 for Form A and 27 for Form B)		
Two-parameter beta binomial model	.887	.881
Four-parameter beta binomial model	.893	.894
UIRT	.899	.898
Normal approximation	.904	.896
ML Estimates (Cut = .375 for both forms)		
Rudner's Method	.905	.905
Rounded Scale Scores (Cut = 120 for both forms)		
Livingston–Lewis Procedure (with two-parameter beta binomial model)	.895	.889
Livingston–Lewis Procedure (with four-parameter beta binomial model)	.903	.896
UIRT	.885	.898
Normal approximation	.899	.898

A random sample of 3,000 examinees was drawn from the full examinee pool, and this subsample was used to compute the statistics reported in the following tables and figures. The composite total raw score was defined as the simple sum of the number-correct scores across the three subcontent domains.

The item responses from these 3,000 examinees were also used to estimate item parameters using three different IRT models: SS-MIRT (simple-structure multidimensional IRT), BF-MIRT (bifactor multidimensional IRT), and UIRT 2PL. Item calibrations were conducted separately for each form using *flexMIRT* (Houts & Cai, 2016). Because the same group of examinees was used for both calibrations, the resulting item parameter estimates were on the same proficiency scale and did not require additional transformation.

Table 7 presents the correlations among the three content areas for Forms A and B. The values above the diagonal are disattenuated correlations, while the values in parentheses represent the estimated latent trait correlations from the SS-MIRT model as obtained from *flexMIRT*. The fact that these correlations are substantially less than one provides empirical support for the multidimensional structure of the data.

In this example, the BF-MIRT model includes four factors—one general (overall) factor and three content-specific factors—while the SS-MIRT model

Table 6: P-values and Biserial Correlations for Composite Score Example

		Content Domains		
		Subcontent 1	Subcontent 2	Subcontent 3
Biserial				
Form A	.406	.403	.433	.357
Form B	.399	.383	.446	.341
P-value				
Form A	.660	.662	.728	.519
Form B	.662	.647	.737	.543

specifies three correlated content domain factors. In contrast, the UIRT model assumes a unidimensional structure, treating the data as if they arise from a single underlying trait without accounting for domain stratification.

Table 7: Observed and Disattenuated Correlations Composite Score Example

Form		Content Domains		
		Subcontent 1	Subcontent 2	Subcontent 3
Form A	Subcontent 1	1	.955 (.907)	.705 (.699)
	Subcontent 2	.536	1	.513 (.489)
	Subcontent 3	.325	.255	1
		Subcontent 1	Subcontent 2	Subcontent 3
Form B	Subcontent 1	1	1 (.972)	.749 (.560)
	Subcontent 2	.557	1	.516 (.480)
	Subcontent 3	.309	.242	1

Note. Observed correlations are shown below the diagonal; classical disattenuated correlations appear above the diagonal; and IRT-based latent trait correlations from the SS-MIRT model are shown in parentheses above the diagonal.

To equate Form B to Form A, a single-group design was employed. Postsmoothing equipercentile equating was conducted using *RAGE-RGEQUATE* (Zeng et al., 2005), with a smoothing parameter of $S = .05$. This procedure yielded integer raw scores on Form B corresponding to non-integer raw scores on Form A. The resulting non-integer scores were then rounded and truncated to produce equated composite raw scores.

In addition, a scale score metric ranging from 50 to 80 was developed for

Form A. The raw-to-scale score conversion was constructed by generating the relative frequency distribution of raw scores and smoothing it using the polynomial loglinear method. Percentile ranks were then derived from the smoothed distribution, and corresponding z-scores were calculated for each integer raw score. These z-scores were transformed to have a mean of 70 and a standard deviation of 4. The final scale scores were obtained by rounding and truncating the transformed values. Scale scores for Form B were derived through equating to Form A and applying the Form A raw-to-scale score conversion.

To examine classification consistency and accuracy, a cut score was established at a scale score of 70 for Form A, resulting in a passing rate of approximately 50%. The corresponding cut score on the Form A raw score metric was obtained by back-mapping from the scale score conversion and was equal to 21. The equivalent cut score on the Form B raw score metric, obtained via equating, was also 21. The scale score cut score for Form B was the same as that of Form A by definition, since the scale score metric was shared across forms. Although the raw score cut scores happened to align numerically in this case, the equivalence was achieved through equating, not by design. Due to the limited research on the use of IRT-based proficiency scores for composite scoring, only the raw and scale score metrics are considered in this example.

2.2.1 Reliability Coefficients

Table 8 summarizes several reliability coefficients for the composite-domain example. Although the data are stratified and multidimensional, unstratified and univariate reliability estimates are also reported for comparative purposes. Recall that coefficient alpha and the generalizability coefficient from a $p \times i$ design yield identical results. Similarly, stratified alpha is equivalent to the composite-score generalizability coefficient derived from a multivariate GT $p^\bullet \times i^\circ$ design. Stratified alpha assumes essential tau-equivalence within each stratum but not necessarily across strata—an assumption that is clearly more defensible than that of unstratified alpha in this context. As expected, results for stratified alpha are slightly higher and closer to the replication values, on average, than those from unstratified coefficient alpha.

Feldt's coefficient, which relaxes the essential tau-equivalence assumption and instead assumes items are congeneric both within and across strata, provides an alternative approach. In practice, stratification typically has little impact on Feldt's coefficient; indeed, the stratified version of Feldt's coefficient (not shown) produced results nearly identical to the unstratified version.

The D-method was used for all IRT-based procedures. For the SS-MIRT and BF-MIRT models, 11 quadrature points were used for each dimension, while 41 points were used for the single dimension of the UIRT model. The SS-MIRT model employed a trivariate standard normal distribution with inter-factor correlations estimated through *flexMIRT* calibration. In contrast, the BF-MIRT model assumed a four-dimensional uncorrelated standard normal distribution.

An additional model, labeled "BF-MIRT General," was included for comparison. In this specification, only the general factor is retained in the model, while

all specific content factors are treated as nuisance and excluded. Although this approach may not fully capture the dimensional structure of the data, it is not entirely misspecified and is therefore considered a useful point of comparison.

The two MIRT models (SS-MIRT and BF-MIRT) generally yielded higher reliability estimates than both the UIRT and BF-MIRT General models. This finding is consistent with prior results reported by Lee et al. (2020). On average, the single-administration estimates produced by the MIRT models were closer to the replication-based values than those from UIRT. Similar patterns were observed for composite scale scores derived from the IRT procedures.

There is no coherent framework within CTT or GT for computing reliability on the scale score metric. One practical approach is to estimate scale score CSEMs using the compound binomial model and then aggregate these values to obtain the overall error variance. Because no model-based estimate of observed score variance is available under CTT or GT for scale scores, the actual observed variance of the scale scores was used in the reliability calculation. As shown in Table 8, this approach resulted in slightly underestimated reliability values compared to the replication-based estimate.

Table 8: Reliability Coefficients for Composite Score Example

Score Types Reliability Coefficients	Form A	Form B	Replication
Composite Raw Scores			
Coefficient alpha	.730	.719	
Feldt's coefficient	.731	.720	
Stratified alpha	.735	.722	
GT $p \times i$	.730	.719	
MGT $p^{\bullet} \times i^{\circ}$	.735	.722	.732
UIRT	.733	.718	
BF-MIRT general	.718	.699	
SS-MIRT	.737	.721	
BF-MIRT	.744	.726	
Composite Rounded Scale Scores			
Compound binomial model	.678	.669	
UIRT	.713	.698	
BF-MIRT General	.698	.678	.719
SS-MIRT	.717	.703	
BF-MIRT	.727	.708	

2.2.2 Conditional Standard Error of Measurement

Estimates of CSEMs for composite raw scores are presented in Figure 5, based on Form A data only. The estimation methods include: Lord's binomial error model that does not account for stratification (Equation 78); the compound binomial model, which incorporates stratification appropriately (Equation 94); the UIRT model assuming unidimensionality (square root of Equation 52); and the SS-MIRT and BF-MIRT models (both using the square root of Equation 62).

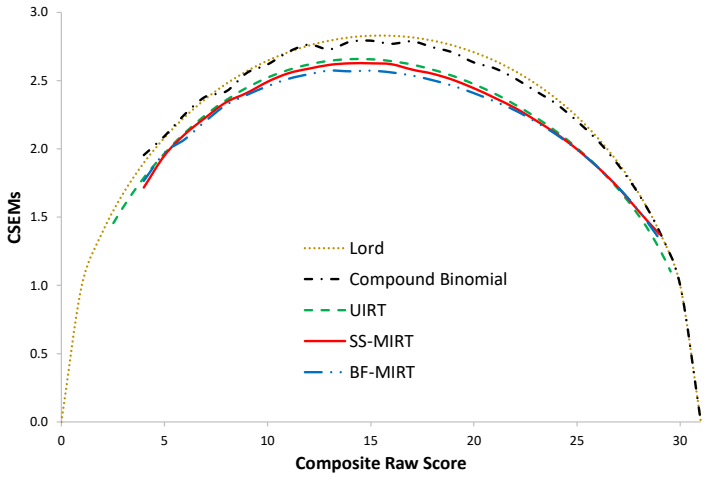

Figure 5: CSEMs for composite raw scores.

In the application of the compound binomial model, examinees with the same composite score could have different combinations of subdomain scores, resulting in different CSEMs. To produce a single CSEM line for this model, the square root of the average CSEMs across examinees with the same composite score was computed.

For the SS-MIRT and BF-MIRT models, the M-method (p. 346) was used, which involved drawing 10,000 proficiency combinations from a specified multivariate distribution—either a trivariate standard normal with estimated correlations (for SS-MIRT) or a four-dimensional uncorrelated standard normal (for BF-MIRT).

For plotting purposes, expected (i.e., true) composite scores computed under each of the three IRT models were rounded. When multiple draws resulted in the same rounded score, their corresponding CSEMs were averaged. This procedure led to a truncated range of composite scores in the final plot.

As expected, the three sets of IRT-based CSEMs tend to cluster together and are consistently lower than those obtained from the binomial and compound binomial models. Among the IRT methods, the two MIRT models generally produced smaller CSEMs than the UIRT model, which aligns with the reliability

results presented in Table 8. CSEMs from Lord's binomial model were slightly, but consistently, higher than the average CSEMs produced by the compound binomial model.

Figure 6 displays the CSEMs for scale scores. Results from Lord's method are not included in this figure. Across methods, a similar pattern emerges, with the IRT-based approaches yielding smaller CSEMs than the compound binomial model. The jagged appearance of the CSEM curves is primarily attributable to fluctuations in the slope of the raw-to-scale score conversion function.

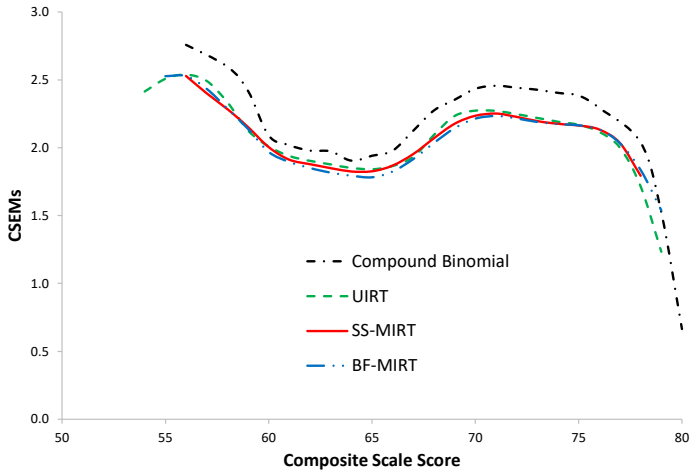

Figure 6: CSEMs for composite scale scores.

2.2.3 Classification Consistency and Accuracy Indices

Table 9 presents estimates of the classification consistency index ϕ for the composite-domain example. Stratified alpha was used as the reliability input for both the normal approximation and strong true score models, with values of .735 for Form A and .722 for Form B. The beta-binomial models, which do not account for stratification, treat the data as a single unified test rather than as three distinct domains. As a result, they tend to underestimate classification consistency.

The three IRT-based methods yielded very similar results, with the UIRT model producing slightly lower estimates than the two MIRT models. Results for rounded scale scores closely mirrored those for raw scores. In fact, estimates from the normal approximation and the three IRT methods were identical across the two score metrics. This equivalence is due to the one-to-one mapping between raw and scale scores in the conversion process.

Interestingly, the Livingston-Lewis method performed well in combination with the beta-binomial model for scale scores, despite the beta-binomial model

Table 9: Classification Consistency for Composite Score Example

Classification Consistency	Form A	Form B	Replication
Composite Raw Scores (Cut = 21 for both forms)			
Two-parameter beta binomial model	.749	.744	.767
Four-parameter beta binomial model	.754	.749	
Normal approximation	.763	.757	
UIRT	.769	.766	
SS-MIRT	.770	.767	
BF-MIRT	.774	.769	
Rounded Scale Scores (Cut = 70 for both forms)			
Livingston–Lewis Procedure (with two-parameter beta binomial model)	.768	.762	.767
Livingston–Lewis Procedure (with four-parameter beta binomial model)	.775	.768	
Normal approximation	.763	.757	
UIRT	.769	.766	
SS-MIRT	.770	.767	
BF-MIRT	.774	.769	

tending to underestimate classification consistency when applied to raw scores.

Table 10 presents a summary of the classification accuracy index γ for both composite raw and scale scores. Overall, the accuracy estimates are higher than the consistency estimates reported in Table 9, as expected. Similar to the consistency results, the beta-binomial models tend to yield somewhat lower accuracy estimates compared to the other methods.

Notably, the performance of the UIRT model is nearly comparable to that of the MIRT models, despite the multidimensional nature of the data—an unexpected but informative finding. In general, all estimation methods produced closely aligned results, with the exception of the beta-binomial models applied to the raw score metric, which continued to show lower estimates relative to the other approaches.

2.3 Computer Programs

Several computer programs were used for computing results reported in this report. An R program, *emreliability* (Liu, Lee, & Liang, 2025), was developed to compute reliability coefficients and CSEMs. *BB-CLASS* (Brennan, 2004) was used to obtain classification consistency and accuracy results for strong

Table 10: Classification Accuracy for Composite Score Example

Classification Accuracy	Form A	Form B
Composite Raw Scores (Cut = 21 for both forms)		
Two-parameter beta binomial model	.813	.808
Four-parameter beta binomial model	.818	.813
Normal approximation	.828	.823
UIRT	.831	.826
SS-MIRT	.830	.827
BF-MIRT	.832	.828
Rounded Scale Scores (Cut = 70 for both forms)		
Livingston–Lewis Procedure (with two-parameter beta binomial model)	.832	.827
Livingston–Lewis Procedure (with four-parameter beta binomial model)	.836	.831
Normal approximation	.828	.823
UIRT	.831	.826
SS-MIRT	.831	.829
BF-MIRT	.833	.830

true score models and Livingston-Lewis procedure. The results for the normal approximation procedure were computed using *NM-CLASS* (Kim, 2019). IRT-based classification consistency and accuracy indices were computed using *IRT-CLASS* (Lee & Kolen, 2008) and *MIRT-PP* (Lee, 2015), and Rudner's results were obtained by *cacIRT* (Lathrop, 2015).

3 Summary and Conclusions

This report provides a set of computational examples to illustrate the application of reliability statistics grounded in CTT, GT, and IRT, as presented in the *Reliability in Educational Measurement* chapter of the *5th edition of Educational Measurement* (Lee & Harris, 2025). Two examples are explored: one based on a single-domain assessment, and the other involving a composite score from a multidomain assessment.

For each example, the report presents multiple reliability coefficients, CSEMs, and classification consistency and accuracy indices across three score metrics: number-correct raw scores, IRT proficiency scores, and scale scores. Both univariate and multivariate methods are considered within each of the three measurement theories. Empirical replication results, obtained from constructed

pseudo-parallel forms, serve as benchmarks to evaluate the accuracy of single-administration estimates.

Findings from the two examples show that, while results are generally comparable, notable differences emerge across the three measurement models and across the different score metrics used. These differences reflect the distinct assumptions, estimation procedures, and error structures inherent to each framework.

The report highlights the critical importance of aligning estimation methods with the conceptual and structural characteristics of the assessment. Model assumptions, such as dimensionality, form parallelism, and the treatment of domain scores, directly influence the interpretation and comparability of reliability estimates. Consequently, careful attention to these assumptions is essential for drawing valid inferences about score precision.

4 References

Brennan, R. L. (2004). *BB-CLASS: A computer program that uses the beta-binomial model for classification consistency and accuracy* (Version 1.0, CASMA Research Report No. 9). Iowa City: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Available at `http://www.education.uiowa.edu/casma`

Houts, C. R., & Cai, L. (2016). *flexMIRT user's manual version 3.5: Flexible multilevel multidimensional item analysis and test scoring.* Chapel Hill, NC: Vector Psychometric Group.

Kim, S. Y. (2019). *NM-CLASS: For classification consistency and accuracy using the normal approximation procedure* [Computer software]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. Available at `http://www.education.uiowa.edu/casma`

Lathrop, Q. N. (2015). *cacIRT: Classification accuracy and consistency under item response theory* (R package version 1.4). Retrieved from `http://CRAN.R-project.org/package=cacIRT`

Lee, W. (2015). *MIRT-PP: Multidimensional item response theory psychometric properties* [Computer software]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.

Lee, W., & Harris, D. J. (2025). Reliability in educational measurement. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 277–381). Oxford University Press. DOI: `10.1093/oso/9780197654965.003.0005`

Lee, W., & Kolen, M. J. (2008). *IRT-CLASS: A computer program for item response theory classification consistency and accuracy* (Version 2.0) [Computer software]. Iowa City, IA: Center for Advanced Studies in Mea-

surement and Assessment, University of Iowa. Available at `http://www.education.uiowa.edu/casma`

Lee, W., Kim, S. Y., Choi, J., & Kang, Y. (2020). IRT approaches to modeling scores on mixed-format tests. *Journal of Educational Measurement*, *57*, 230–254.

Liu, H., Lee, W., & Liang, M. (2025). *emreliability: Test reliability and CSEM in educational measurement* (R Package Version 1.0.0). Comprehensive R Archive Network (CRAN).