Technical Information Brief – Reliability Evidence

As part of the development process for the CASAS assessments approved by OCTAE for use in the National Reporting System (NRS) for Adult Education, CASAS examines and provides reliability evidence for all test forms. Each test form in each series is evaluated using methods described in this brief. Results provide evidence that the tests provide reliable score interpretations and a high degree of classification consistency into NRS Educational Functioning Levels (EFLs) (NRS, 2017).

Introduction

Reliability is defined as the consistency of measurements when the testing procedure is repeated on a population. The *Standards for Educational and Psychological Testing* (AERA et al., 2014) also point out that the term reliability has been expressed in two ways in measurement literature. The first way is through the reliability coefficients of classical test theory or the correlation between scores on two alternate and equivalent forms of the test. The second more general way is the consistency of test scores across replications of the testing procedure regardless of how consistency is estimated or reported. The degree of consistency is evaluated with various methods including standard errors, IRT information functions, and indices of classification consistency.

Each CASAS test series is comprised of sets of parallel forms that are constructed so that two forms at each test level can be used independently of each other and be considered equivalent. Examinees with similar ability taking the parallel forms should show comparable performance. Parallel forms are constructed to have comparable content coverage, test statistics, and test difficulty levels. The items comprising the parallel forms contain comparable content across the NRS EFLs, to reflect the same construct and similar Rasch UnIT (RIT) measures. A RIT represents a unit on the logarithmic scale of item difficulty and is used to quantify the difference between a person's ability level and the difficulty of the item on the test (Rasch, 1960).

Sets of parallel test forms at each test level are created to represent comparable empirical difficulty by selection of a combination of items of similar content and RIT values. Calibration makes it possible to calculate RIT measures that relate students' achievement to curriculum on the tests. When calibrated, the scale score range and standard error of measurement across the range of score points, including the cut-points, on each of two fixed-item parallel forms is comparable.

It is important to note that the number of items used in an assessment series has potential implications on reliability. If an assessment is lengthened with comparable tasks or items, the reliability is likely to increase. The lengthening of an assessment is an effective and commonly used method for improving test reliability (AERA et al., 2014). Therefore, test reliability is one of the factors in determining the length of CASAS tests.

Reliability Studies

CASAS uses several methods to examine the reliability of each test form. These include:

- Examination of internal consistency using reliability statistics
- Examination of internal consistency using a split-halves study
- Examination of parallel form test characteristic curves
- Examination of the test information function
- Estimation of classification accuracy (CA) and classification consistency (CC) using IRT
- Examination of classification consistency (CC) using empirical data

Study Descriptions and Methodologies

Summary Statistics

Using examinee data obtained from field testing, CASAS calculates a series of summary statistics to examine the internal consistency of the test forms. Part of this analysis is to ensure that the parallel forms have similar mean scores, standard deviations, and standard errors. The Cronbach alpha reliability coefficient, defined as the average of all possible split-half estimates, is also analyzed. For example, for Reading STEPS form 621, the Cronbach alpha of 0.89 can be interpreted to mean that at least 89 percent of the observed score variance is due to true score variance.

Split Halves Reliability Study

To examine the internal consistency of each test form, a split-halves reliability coefficient is calculated. This analysis measures the extent to which the entire test contributes to the measurement precision.

Each test form is “split” into two halves. As described in Standard 2.5 (AERA et al., 2014), the halves are designed to be of comparable content and statistical characteristics. Independent scores for each half are calculated and the scores are correlated to yield a half-test reliability estimate. This is used to calculate the reliability of the full test using the Spearman-Brown prophecy formula (Fan & Randall, 2018) which is written as follows:

$$r_{full} = \frac{2(r_{half})}{1 + (r_{half})}$$

For example, for Reading STEPS Form 621 the Spearman Brown prophecy formula shows a reliability estimate of .0895.

Parallel Form Test Characteristic Curves

Gulliksen (1950, Chapter 14) states that tests are parallel if they yield the same true scores and error variances. While these conditions cannot be empirically tested, they imply testable hypotheses that the observed scores for parallel tests have equal means, variances, and intercorrelations or alternate form reliabilities (Wilks, 1946). With the advent of automated test assembly techniques using mixed integer linear programming (van der Linden and Adema, 1998), we can create alternate forms that simultaneously have the same mix of content, the same test characteristic curves (TCCs) and test information functions (TIFs). CASAS tests are designed to have identical TCCs and TIFs for each set of parallel forms. The test characteristic curves (TCCs) for each pair of parallel forms are evaluated for their similarity. A dashed line and a solid line represent the parallel test forms. Due to the high degree of match between the forms, the two test characteristic curves are often indistinguishable. To illustrate, Figure 1 shows the test characteristic curves for Forms 621 and 622 from the Reading STEPS series.

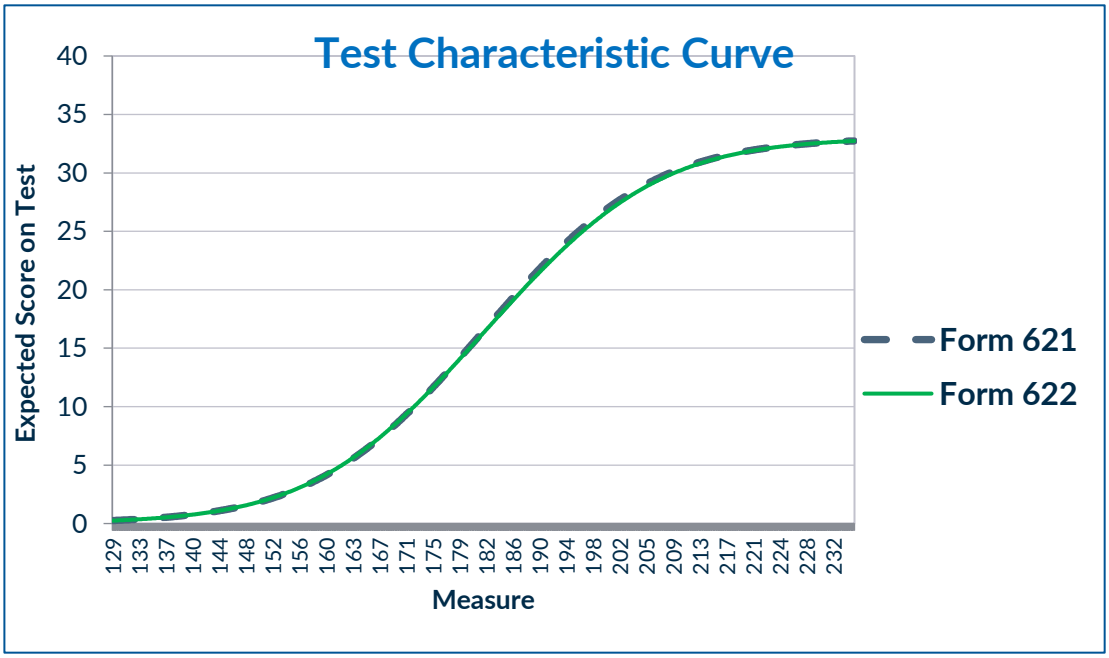

Figure 1 **Test Characteristic Curve**

Test Information Function

In IRT, the classical concept of reliability is extended and replaced by the concept of test information. As the AERA et al. (2014) Standards describe, “IRT addresses the basic issue of reliability/precision using information functions, which indicate the precision with which observed task/item performances can be used to estimate the value of the latent trait for each test taker” (p. 34).

Hambleton and Lam (2009) pointed out that an estimate of reliability can be derived from test information due to the relationship between test information and measurement error. The TIF predicts the accuracy to which we can measure any value of the latent ability (Partchev, 2004). The TIF for each parallel form is compared to the TIF of the alternate parallel form for comparability and error rates. The TIFs of each pair of parallel forms show that the values of TIF across all levels of theta are comparable. This provides evidence that each parallel form provides very similar information and precision at all levels of theta.

The standard errors are higher at the tails of the score distribution for any given test form, but this has less impact on classification accuracy and classification consistency. Overall, the standard errors are relatively small and provide evidence of the reliability of the score points. To illustrate, Figure 2 shows the TIF for Form 621 from the Reading STEPS series.

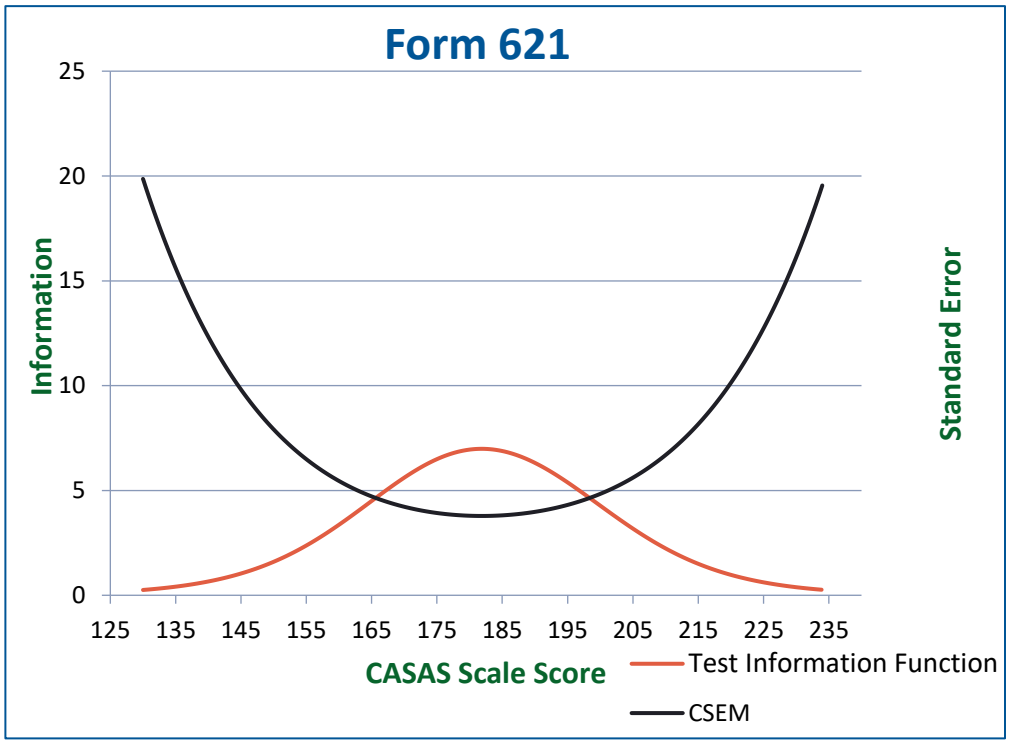

Figure 2 Test Information Function

Classification Accuracy (CA) and Classification Consistency (CC)

Standard 2.16 from the *Standards for Educational and Psychological Testing* (AERA et al., 2014) states that "When a test or combination of measures is used to make classification decisions, estimates should be provided of the percentage of examinees who would be classified in the same way on two replications of the procedure".

The CASAS NRS-approved assessments are criterion-referenced tests developed on an underlying IRT scale with a principal purpose of correctly and consistently classifying examinees into performance categories for adult education students. For purposes of analyzing the suitability for use in the NRS for Adult Education, CASAS considers classification consistency to be the most important estimation of the reliability of scale scores. Classification consistency describes how consistently test forms measure examinees' skills in the context of the NRS EFLs.

Classification Accuracy (CA) is the extent to which the "true" classifications of examinees agree with the observed classifications. Classification Consistency (CC) refers to the degree to which examinees are classified into the same performance levels by independent, parallel forms of a test (Hambleton & Novick, 1973; Lee, 2010). CA can also be defined as "the extent to which observed classifications of examinees based on the results of a single replication would agree with their true classification status. CC refers to the extent to which the observed classifications of examinees would be the same across replications of the testing procedure" (AERA et al., 2014, pg. 40).

Estimating Classification Accuracy and Classification Consistency Using IRT

CASAS used Rudner's method (Rudner, 2001; 2005) and Li's extension of Rudner's approach to measure CA and CC based on Item Response Theory (IRT). A method based on IRT was determined to be appropriate for contemporary educational testing programs (Diao & Sireci, 2018). One advantage is the ability to estimate classification accuracy and classification consistency throughout the test development process. This allows CASAS to estimate measurement precision on different iterations of the test and make adjustments based on

these estimations. Therefore, the CASAS test specifications considered the reliability of scores that would be produced and the validity of the score interpretations (Raymond & Grande, 2019).

Rudner's method involves several steps. First, the test scores, including the cut scores, are transformed from the raw score to the theta score scale. Next, a K-by-K classification consistency table is set up using the categories for the observed scores as the rows and the categories for the true scores as the columns. The distribution of examinee scores at each theta scale score is initially estimated and later entered using actual examinee data from field-test results. The conditional standard error of measurement is entered for each score point. Next, the elements of the contingency table that are conditional probabilities are calculated. Finally, the overall CA is calculated as the sum of the diagonal elements in the contingency table. Li's approach is then used to convert the classification accuracy to classification consistency. To illustrate, Table 1 shows a K-by-K classification consistency table for Form 621 from the Reading STEPS series.

Table 1 Classification Accuracy Estimate – Form 621

	Beginning ESL Literacy	Low Beginning ESL	High Beginning ESL	Low Intermediate ESL	High Intermediate ESL	Advanced ESL
Beginning ESL Literacy	**10.63%**	2.42%	0.00%	0.00%	0.00%	0.00%
Low Beginning ESL	3.27%	**83.69%**	0.00%	0.00%	0.00%	0.00%
High Beginning ESL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Low Intermediate ESL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
High Intermediate ESL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Advanced ESL	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Expected Classification Accuracy for Form 621 = 94.3%

Estimating Classification Consistency Using Empirical Data

Standard 2.16 from the *Standards for Educational and Psychological Testing* (AERA et al., 2014) states that "Although decision consistency is typically estimated from the administration of a single form, it can and should be estimated directly through the use of a test-retest approach, if consistent with the requirements of test security, and if the assumption of no change in the construct is met and adequate samples are available."

CASAS obtains a representative sample of examinees who take both parallel test forms at their appropriate test level within a seven-day period. The same test administration guidelines are followed for both the time 1 (T1) and time 2 (T2) testing events. The results are analyzed for classification consistency based on the test scores from the time T1 and time T2 testing events.

Summary

As stated in the *Standards for Educational and Psychological Testing* (AERA et al., 2014) there is no single, preferred approach to quantification of reliability or precision. CASAS provides evidence of reliability and precision using various methods with a focus on classification accuracy and classification consistency.

The entirety of the evidence for the NRS-approved CASAS assessments shows that the CASAS assessments provide reliable scores used to make valid interpretations regarding student skill levels and learning gains over time. Test scores consistently classify students into the appropriate NRS EFLs.

References

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing. Washington, D.C.

Diao, H. & Sireci, S.G. (2018). Item response theory-based methods for estimating classification accuracy and consistency. *Journal of Applied Testing Technology, 19*(1), 20-25.

Gulliksen, H. (1950). Experimental methods of obtaining test reliability. In H. Gulliksen, *Theory of mental tests* (pp. 193–218). John Wiley & Sons Inc. https://doi.org/10.1037/13240-015.

Hambleton, R.K. and Lam W. (2009). Redesign of MCAS Tests Based on a Consideration of Information Functions. University of Massachusetts, Amherst.

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10(3), 159–170.

Li, S. (2006). *Evaluating the consistency and accuracy of the proficiency classifications using item response theory.* [Unpublished doctoral dissertation]. University of Massachusetts Amherst.

National Reporting System for Adult Education. (2017). Technical Assistance Guide for Performance Accountability under the Workforce Innovation and Opportunity Act. Division of Adult Education and Literacy, Office of Career, Technical, and Adult Education, U.S. Department of Education, Washington, DC: Author.

Partchev, Ivailo (2004). A visual guide to item response theory, Friedrich-Schiller-Universität Jena.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Raymond, M. R. & Grande, J. P. (2019): A practical guide to test blueprinting. *Medical Teacher*, (41 (8), 854-861.

Rudner, L.M. (2001). Computing the expected proportions of misclassified examinees. Practical Assessment, Research & Evaluation, 7(14). http://pareonline.net/getvn.asp?v=7&n=14.

Rudner, L. M. (2005). Expected classification accuracy. Practical Assessment Research & Evaluation, 10(13). Available online: https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1162&context=pare.

Van der Linden Wim J. & Adema, Jos J. Simultaneous Assembly of Multiple Test Forms (1998). Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.

Wilks, S. S. Sample criteria for testing equality of means, equality of variances and equality of covariances in a normal multivariate distribution. *Ann. math. Stat.*, 1946,17, 257–281.

