Technical Information Brief – Test Validation Studies

CASAS conducts a variety of studies to examine and provide evidence that the NRS-approved CASAS tests are valid measurement tools for the adult education population. The demographic characteristics of the student population reported in the NRS Federal Tables guide the recruitment of the study samples.

To build a validity argument for a test, there are several types of evidence that can be introduced. The AERA et al. (2014) Standards describe five "sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use" (p. 13). These five sources of validity evidence are based on (a) test content, (b) relations to other variables, (c) internal structure, (d) response processes, and (e) testing consequences.

Introduction

For the CASAS NRS-approved assessments, the three principal score interpretations are: (1) identify the learner's skill level as defined by the NRS EFLs; (2) measure learning gain with respect to the NRS EFLs; and (3) provide input for instruction. The test scores identify the learner's skill level on a score scale aligned to NRS Educational Functioning Levels (EFLs). All tests are scored on this common scale aligned to instructional levels and learner skill levels.

This technical information brief offers a brief description of the studies CASAS conducts to examine the validity of each assessment.

Validity Evidence Based on Content

Test development and content validation procedures provide evidence supporting interpretations of test scores (Cronbach, 1971; Schmeiser & Welch, 2006; Thompson, Johnstone, & Thurlow, 2002). CASAS collaborates with psychometricians to conduct a series of alignment studies to evaluate how well the test items and test forms align with the NRS EFL descriptors. For more information on the Content Validity of the CASAS assessments, please refer to the *CASAS Technical Research Brief – Content Validity.*

Relationship of Test Scores to Other Variables

As the AERA et al. (2014) Standards pointed out, "Evidence based on the relationship with other variables provides evidence about the degree to which these relationships are consistent with the construct underlying the proposed test score interpretation" (p. 16).

Validity evidence based on relationships to other variables is examined through traditional forms of criterion-related validity evidence such as concurrent and predictive validity studies. These external variables are used to evaluate hypothesized relationships between test scores and other measures of student performance. In general, these relationships can be classified as convergent validity and discriminant validity. Convergent validity is a comparison to a variable with which we would expect a high correlation. Discriminant validity is a comparison to a variable with which we would expect low correlation.

Relationship Between CASAS Test Scores and Other Test Scores

To examine the relationship between the CASAS NRS-approved assessments and another assessment developed to assess similar skills of the same population, CASAS compares scores from students who take each of the two assessments. The methodology strictly follows the test administration guidelines used in normal testing practice and the time between the two testing events is controlled to limit significant learning taking place that could affect the comparability of student performance. This ensures that concurrent validity is being examined.

Results focus on an examination of the bivariate correlation between the two test scores and the NRS EFL classification consistency based on the two test scores. Overall results have shown a strong positive relationship between performance on the two assessments which provides evidence of criterion validity.

Relationship Between CASAS Test Scores and Teacher Judgment

To further examine the criterion validity of the NRS-approved CASAS assessments, and specifically concurrent validity, CASAS has submitted evidence that compares the relationship between students' test scores and the teachers' judgment of students' NRS EFL classification based on classroom observation of their skill levels. These NRS classifications provided by teachers are compared to the students' NRS classification based on their CASAS test score. An examination of the results looks at both the bivariate correlations and classification consistency between the two test scores.

As the AERA et al. (2014) Standards point out, "When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and eliciting judgments should be fully described. The description of procedures should include any training and instructions provided, should indicate whether participants indicated their decisions independently, and should report the level of agreement reached."

Each participating teacher completes a training program which includes reference materials on the skills associated with each NRS EFL to prepare for making teacher judgments of student ability levels. Teachers receive detailed instructions about the study and record their judgments on a rating form.

The Teacher Judgment Study is conducted as a "blind experiment." Teachers do not know the students' test scores on the CASAS assessment before making their NRS EFL judgments. Teachers only provide judgments for students who they can observe for sufficient class time and the teachers make their judgments within a specified time period around the testing event. The results of the exact classification consistency, adjacent classification consistency and the bivariate correlation are examined as validity evidence.

Examination of Learning Gains Achieved Over Time on CASAS Tests

A primary purpose of CASAS NRS-approved assessments is to measure student learning gains over a period of instruction in an adult education program. To examine how the CASAS tests measure student progress across time, CASAS conducts a longitudinal study. This study examines the relationship between test results on an initial test (T1) and test results on a subsequent test (T2) which is administered after a recommended amount of instruction takes place between the two testing events. This study assumes the level of effective instruction and is intended to demonstrate that that the CASAS tests measure learning gains as defined by T2-T1.

Whereas the two previously described studies are examinations of concurrent validity, this study examines predictive validity. That is, after a T1 test and an appropriate period of instruction, CASAS examines the hypothesis that test takers will demonstrate learning gains as measured by the CASAS test.

This research also provides valuable examination of the consequential validity of the CASAS test by looking at student outcomes, defined by performance measured on repeated test administrations. This examination of learning gains is especially relevant given the intended use of the NRS-approved CASAS tests to address adult education program accountability.

The study objective is to follow the test administration process that will occur in normal testing practice to measure learning gains. These results provide evidence that students are achieving the expected learning gains when the CASAS test is used to measure student performance.

Validity Evidence based on Internal Structure

Differential Item Functioning (DIF)

DIF is said to occur when equally able examinees differ in their probabilities of answering a test item correctly as a function of group membership (AERA et al., 2014). The examination of DIF is a method to examine internal structure by using empirical data to determine whether test items function differently among subgroups of examinees. The CASAS NRS-approved assessments, built using item response theory (IRT), use the information and item statistics provided by IRT to analyze DIF. For more information on DIF, please refer to the *CASAS Technical Research Brief – Fairness and Sensitivity.*

Reliability and Precision

The degree to which the CASAS test items measure a single dimension and result in reliable scores provides additional validity evidence based on internal structure. For more information on reliability, please refer to the *CASAS Technical Research Brief – Reliability.*

Face Validity based on Examinee Survey

CASAS administers a follow-up survey to students who participated in the field-testing process. The goal of the survey is to gather information about their test-taking experiences. The survey is anonymous and does not collect personal identifiable information from the respondent. These results allow for an examination of face validity, or to what extent examinees consider that the CASAS tests measure the construct it is designed to test.

Summary of Validity Evidence

CASAS compares performance on the NRS-approved tests with another assessment developed and validated to assess the same population. This provides evidence of criterion validity. Additional evidence of criterion validity is examined when comparing NRS EFL classification consistency based on teacher judgments of students' ability with student performance on the CASAS tests. Results of a Learning Gains study show that the students demonstrate an expected level of performance gains when administered two CASAS tests with an intervention (instruction) between the tests. Student survey results provide evidence of face validity, indicating that examinees believe that CASAS tests measure their reading, math, or listening skill level. Empirical analyses using DIF provide compelling evidence that the test items are fair and unbiased towards different population groups.

References

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

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.

Schmeiser, C.B., & Welch, C.J. (2006). Test development. In R.L. Brennan (Ed.). *Educational Measurement* (4th ed., pp. 307-354). Westport, CT: American Council on Education and Praeger.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002, June). Universal design applied to large scale assessments (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

