Technical Information Brief – Test Development Overview

The NRS-approved CASAS assessments address testing needs in adult education, including identifying an individual's skill level, measuring learning gains over time across the NRS Educational Functioning Levels (EFLs), and providing information about student proficiency in competency areas (National Reporting System for Adult Education, 2017).

Furthermore, the multiple test series are intended for use by local agencies that must comply with NRS requirements to receive funds from the Office of Career, Technical, and Adult Education (OCTAE). Agencies can aggregate Learners' scores and score gains to provide meaningful summative measures of program effectiveness.

The test development process and procedures conform to the *Standards for Educational and Psychological Testing* (AERA et al., 2014).

Introduction

The most recently published NRS-approved assessments from CASAS include Reading STEPS, Listening STEPS, and Math GOALS 2. CASAS developed the assessments in collaboration with psychometricians, state and local agency users, trainers, administrators, and practitioners from all regions of the country. Each test series provides placement into and skill gain measurement across all six NRS EFLs.

Each of these test series are structured with five levels in an overlapping design. Test levels cover two adjacent NRS EFLs (e.g., Level A covers NRS EFLs 1 and 2, and Level B covers NRS EFLs 2 and 3) and each test level has two test forms that are parallel in content and difficulty. This test structure helps to ensure measurement precision and the ability to place and track student progress across the NRS EFLs. Each test series also includes a 14-item Locator test that accurately places new examinees into their initial test level.

Table 1 Test Design Showing Overlapping NRS EFLs across Test Forms

Test Level	NRS EFLs on Test Forms
A	Place into or Exit Level 1 Place into Level 2
B	Place into or Exit Level 2 Place into Level 3
C	Place into or Exit Level 3 Place into Level 4
D	Place into or Exit Level 4 Place into Level 5
E	Place into or Exit Level 5 Place into or Exit Level 6*

*Exit Level 6 for ESL assessments only

Development of Test Specification

Development of test specifications involve the participation of adult education subject matter experts (SMEs) with expertise in the field of adult education These SMEs represent diverse adult education programs from across the country and are knowledgeable about the current educational standards and the instructional needs of students in these programs. SMEs and adult education practitioners meet annually as part of the CASAS National Consortium. This consortium provides input on the purpose, construct, and design of the series. CASAS also convenes smaller groups of 15 to 30 adult education practitioners from around the country. CASAS consults with psychometricians on the design of each test series. These groups provide input regarding test design and content throughout test development.

Test Content Specifications

The CASAS NRS-approved assessments align to the NRS EFLs. In addition, the assessments align with appropriate educational standards. For example, these include the English Language Proficiency Standards for Adult Education (AIR, 2016) and the College and Career Readiness Standards for Adult Education (Pimentel, 2013). For more information on the Content Validity of the CASAS assessments, please refer to the *CASAS Technical Research Brief – Content Validity.*

Item Development

The CASAS Item and Test Development Department continually develops new test items to expand the pool of viable items. The individuals who participate in item writing for CASAS come from a variety of backgrounds and work experience. All individuals have extensive experience in adult education and the specific modality they are writing for (i.e. reading, listening, math). CASAS recruits adult education practitioners from around the country for item writing. Doing so maximizes the degree to which test items reflect instructional practices and ensures that the items are relevant to the adult examinees. Including these practitioners in the test development process also helps practitioners to feel ownership in the assessments. CASAS selects item writers based on:

- experience in adult education (teaching, curriculum development) with adult education populations for which the tests are intended;
- familiarity with the language, cultural issues, and life experience of adult education populations and with the real-life literacy needs of adults in society;
- demonstrated ability to write to specific test blueprint specifications and standards; and
- completion of fairness and sensitivity training.

Item writers receive training from CASAS staff item writers and editors that include the theory and practice of test development. The training includes an overview of the skills described by each NRS EFL, best practices guidelines for writing successful test items, and practical exercises in writing items to specific content standards. Item writers are mentored by master item writers and editors who give specific feedback on their work to build skills.

Incorporated into the design of all test items are the principles of Universal Test Design (UTD) for accessible assessments. UTD helps to ensure that the items are valid, and that the assessment provides reliable scores and valid inferences for all students while minimizing the need for testing accommodations.

During the item development phase, CASAS focused on numerous UTD elements (Thompson, Johnstone, & Thurlow, 2002):

- Precisely defined constructs
- Accessible, non-biased items
- Amenability to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum comprehensibility

All draft items go through a multistage review. Items that are initially accepted from item writers are thoroughly reviewed by CASAS assessment specialists. These specialists review items for how well they match the test specifications, meet fairness and sensitivity requirements, and whether the passages and items are clear and appropriate for the intended level and the adult education target population. CASAS assessment specialists then revise items to ensure they fully meet the required test specifications. The fairness and sensitivity review includes both a panel review with SMEs and an empirical analysis using Differential Item Functioning. For more information on the fairness and sensitivity of the CASAS assessments please refer to the *CASAS Technical Research Brief – Fairness and Sensitivity.*

Item Field Testing

Field testing is a critical aspect of item development. Items are administered to the target population so that statistical information about the items can be evaluated. CASAS conducts item field testing through two methods: embedding items onto operation test forms and creating standalone intact item field-test forms.

Both classical test theory (CTT) and item response theory (IRT) are used to evaluate the psychometric properties of the test items. In addition, differential item functioning (DIF) analyses are performed to evaluate items. Items are not included in the CASAS Item Bank if they failed to meet established psychometric criteria.

Classical Item analyses include the examination of p-values and point biserial correlations. IRT analyses include the examination of high and low group discrimination and infit-outfit statistics.

Items that meet the established statistical criteria were calibrated and linked onto the IRT scale, new test items are calibrated using the IRT one-parameter model or Rasch model (Rasch, 1960). The Rasch model for dichotomous data provides one item parameter, the probability of the outcome X_{ni}= 1 (a correct response to an item), and is given by:

$$Pr\{X_{ni} = 1\} = \frac{e^{\beta_{n-\delta_i}}}{1 + e^{\beta_{n-\delta_i}}}$$

Where:

$$X_{ni} = a\ dichotomous\ random\ variable$$

e = 2.71828

β_n = the ability of person n

δ_i = the difficulty of item i

Development of Operational Test Forms

The operational test forms are constructed of items that have completed field testing and have met the criteria for inclusion using classical and IRT statistics.

Primary among considerations in assembling two parallel forms at each of the five test levels was adherence to the test specifications. In addition to selecting items for assembly onto forms based on test specifications, classical item difficulty and discrimination statistics are considered. Also considered as part of assembly of forms are the IRT item RITs. These item statistics and consideration of RITs help create parallel forms at each level that yield similar examinee score distributions. More information on test reliability can be found in the *CASAS Technical Research Brief – Reliability.*

Results from evaluations of item fairness (both statistical and judgmental) are also considered in assembly of forms. As described earlier, multiple methods are used to determine if items exhibit item bias for examinee groups. No items are included in final forms without thorough review of statistical and judgmental screening.

For more information on the fairness and sensitivity of the CASAS assessments please refer to the *CASAS Technical Research Brief – Fairness and Sensitivity.*

The Score Scale

Each of the NRS-approved CASAS test series uses a stable equal-interval vertical scale using Rasch units (RIT). The use and maintenance of this scale allows for direct comparison of scores across forms and levels and is essential for assessing cohort and individual growth.

A standard setting study is conducted to define the appropriate score ranges that correspond to each NRS Educational Functioning Level (EFL). CASAS employs the Bookmark standard setting method (Lewis, Mitzel, & Green, 1996). For more information on the CASAS standard setting study please refer to the *CASAS Technical Research Brief – Standard Setting.* The scale scores that correspond to each NRS EFL can be found on the CASAS website at https://www.casas.org/training-and-support/wioa-and-nrs-compliance.

Locator Test Construction

In addition to the operational parallel forms at each level, each series also contains a Locator Test to ensure appropriate placement into an initial test level. Due to limited time allotted for testing at local programs, the Locator is designed to be a quick and accurate tool for the purpose of placing the examinee into the appropriate test level.

Items are selected for the Locator based on their statistical characteristics and with content that represents each of the NRS EFLs. Using Rudner's IRT approach (Rudner, 2001, 2005) it was determined that a 14-item locator test made up of items from the viable CASAS Item Bank would place examinees into the correct test level with the precision needed. To ensure accurate placement based on the items selected, Decision Accuracy (DA) estimates are calculated based on the selection of five cut points that define placement into the five test levels.

Operational Form Field Testing

The goal of the field testing of the operational forms is to validate and measure the performance of the operational forms. The test forms are administered to a diverse sample of examinees representative of the adult education population of interest. The sample represents a comprehensive set of examinees, including candidates across a wide variety of key demographic characteristics, such as age, gender, language, ethnicity, and educational background. The demographic characteristics of the student population reported in the NRS Federal Tables guide the recruitment of the sample population during field testing.

CASAS uses multiple methods to ensure that field-test results reflect motivated examinees. Specifically, administration protocols and item response rate analyses. In addition, CASAS analyzes the feedback of field-test participants using student surveys.

In addition to analyzing the statistical properties of the items and test forms, as described above, CASAS conducts a variety of analyses to further examine the validity of the test forms. These include criterion validity analyses to compare the test results with other measures of the students' performance, a learning gains analysis to examine how well the tests measure student performance over time, and a speededness analysis to examine test taking time and student test scores. For more information on the validation studies please refer to the *CASAS Technical Research Brief – Validation Studies.*

Summary

The CASAS NRS-approved assessments are the culmination of work by a dedicated team of test developers in conjunction with psychometricians and subject matter experts from around the country. Each test series is a valid measure of adult learner progress across the six Educational Functioning levels defined by the National Reporting System (NRS).

References

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

American Institutes for Research. (2016). English Language Proficiency Standards for Adult Education. Washington, DC: Author. Retrieved from https://lincs.ed.gov/publications/pdf/elp-standards-adult-ed.pdf.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). Standard setting: A bookmark approach. In D. R. Green (Chair), IRT-based standard setting procedures utilizing behavioral anchoring. Symposium presented at the Council of Chief State School Officers National Conference on Large-Scale Assessment, Phoenix, AZ.

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.

Pimentel, S. (2013) College and Career Readiness Standards for Adult Education. MPR Associates, Inc. Berkeley CA. Washington, DC.

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

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: http://pareonline.net/getvn.asp?v=10&n=13.

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.

