Technical Information Brief – Standard Setting

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 must define the appropriate score ranges that correspond to each NRS Educational Functioning Level (EFL). Each NRS EFL is defined by a descriptor which details the skills associated with that NRS EFL (OCTAE, 2016).

CASAS collaborates with a team of psychometricians and adult education subject-matter experts to conduct standard setting studies to determine the test score, or "cut score," that represents an examinee's transition from one NRS EFL to the next. The standard setting study provides the information to set psychometrically defensible cut scores for each NRS EFL. The standard setting study results and evaluation feedback support the validity of the final cut scores for each CASAS test series.

Introduction

All CASAS test items are field tested, calibrated, and placed onto an existing CASAS item bank scale. For each CASAS test series, there is a unique item bank and associated item bank scale: ESL Reading, ABE Reading, ESL Listening and ABE Math. Each test item is assigned a Rasch Unit (RIT) value.

Performance standards are established for each CASAS test series to cover the entire spectrum of the CASAS item bank scale. Once performance standards are set, the test taker can be accurately classified into an NRS EFL based on their test score.

Standard Setting Approach

The Standards for Educational and Psychological Testing (AERA et al., 2014) states that, "where the results of the standard-setting process have highly significant consequences, those involved in the standard-setting process should be concerned that the process by which cut scores are determined be clearly documented and that it be defensible. When standard-setting involves judges or subject matter experts, their qualifications, and the process by which they were selected are part of that documentation."

Numerous standard setting methods are used to establish defensible and appropriate cut scores on educational assessments (Hambleton & Pitoniak, 2006). CASAS employs the Bookmark standard setting method (Lewis, Mitzel, & Green, 1996). The bookmark approach is classified as an item mapping approach (Hambleton & Pitoniak, 2006) that focuses on the relationship between item difficulty and examinee performance on the test. It is appropriate for tests that employ item response theory models and is one of the most used standard setting approaches. The selection of the Bookmark method (Lewis, Mitzel, & Green, 1996) reflects consideration of the characteristics of the assessment tasks as well as key requirements of the standard setting method itself.

The Bookmark standard setting method is an examinee-centered approach. Panelists are presented with an ordered item book (OIB) where each page of the book presents one item. The first page in the OIB is the easiest item (empirically) and the last page is the most difficult item. Panelists are asked to place a "bookmark" between two items along the item difficulty continuum that, based on their best judgment and understanding of the performance level descriptors (i.e., NRS EFLs), separate the items that borderline examinees (i.e., exiting one level and entering the next) will likely answer correctly from those that borderline examinees will not likely answer correctly. The panelists are instructed to estimate whether a given transitioning examinee would likely answer the item correctly. The bookmark locations indicate the expected level of performance of a student who is just entering a particular performance level described in the NRS EFLs. Specifically, the bookmark page number indicates the first item the panelist judges the student transitioning into a particular level will answer incorrectly. Therefore, the page before the bookmark indicates the last item they felt this student will answer correctly.

To place a bookmark for a given performance level, panelists are instructed to review each item and ask themselves: “Does a typical minimally-competent candidate at a given performance level have at least a 50% chance of answering this item correctly?” If yes, the panelist moves on to the next item to ask the question again. If no, the panelist places the bookmark for the given performance level in front of the item. In the case of the Reading STEPS standard setting, 50% response probability (RP) is used. Wang (2003) argued that an RP of 0.50 was most appropriate for test scales that use the Rasch (1960) model, given that the likelihood of an examinee getting a correct response is 50% when their ability is equal to the item difficulty. Kolstad et al (1998) conducted a review of different studies, including studies that focused on adult literacy, such as the National Adult Literacy Scale (NALS), and found that from a traditional IRT perspective, the RP of 0.50 was optimal because it maximized the information that could be obtained from each item. On the other hand, Kolstad noted a number of other studies that used response probabilities with higher values, such as 0.67, that were justified by the researchers because they were testing for mastery of a given topic. Again, the intended interpretation and use of scores, in combination with the context, informed the policy decision of the appropriate RP criterion for this study.

Figure 1 illustrates the conceptual task that panelists are presented with to consider the examinee with the minimum knowledge and skills required for NRS EFLs 1-6 for the Reading STEPS test series (the arrows in the figure point to each of the five key transition points). This graphic is presented in the training to help the panelists understand the function of a cut score as the minimum entry point into a particular level.

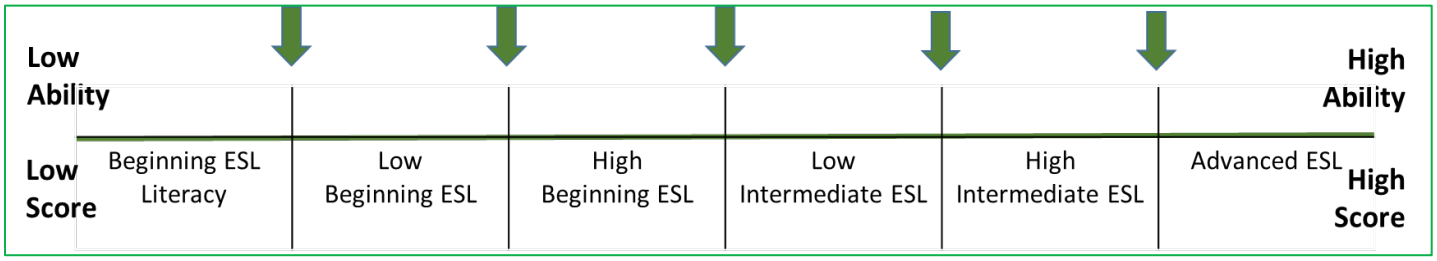

Figure 1 Graphic Representation of the Reading STEPS Standard Setting Tasks

Methodology

Bookmark standard setting methodology is dependent on qualified subject matter experts (SMEs) to contribute their expertise and knowledge of the content domain as well as the intended population of examinees (Hambleton & Pitoniak, 2006). Therefore, the primary criteria for panelist inclusion are that the panelists a) have familiarity with the content and b) are familiar with the abilities of examinees in the target population.

The selected panelists first attend an orientation and training session. This includes a practice round to help ensure that panelists understand the process. The Bookmark standard setting process is then conducted in two rounds. Panelists first perform their Round 1 bookmark ratings independently (offline). After analyzing the Round 1 bookmark results, the panelists reconvene for feedback on the first round of ratings. Round one feedback includes the group’s recommended cut scores (median), the variability of the recommendations, and the impact of the median recommended cut score. The panelists first meet with the psychometrician leading the standard setting to discuss their rating process, ask questions about the initial findings, and review any concerns about distinguishing one level from the next. The panelists then meet in their small groups to further discuss their ratings, the anticipated impact, and how they placed their bookmark for each performance level.

After the small group discussions of the Round 1 results, panelists consider the information and discussion to independently perform their Round 2 bookmark ratings. Their Round 2 ratings may be consistent with their Round 1 ratings or may be adjusted based on the feedback and discussion after their initial findings. Panelists

then submit their final judgments. Figure 2 illustrates the comparability of Round 1 and Round 2 judgments for the Reading STEPS test series.

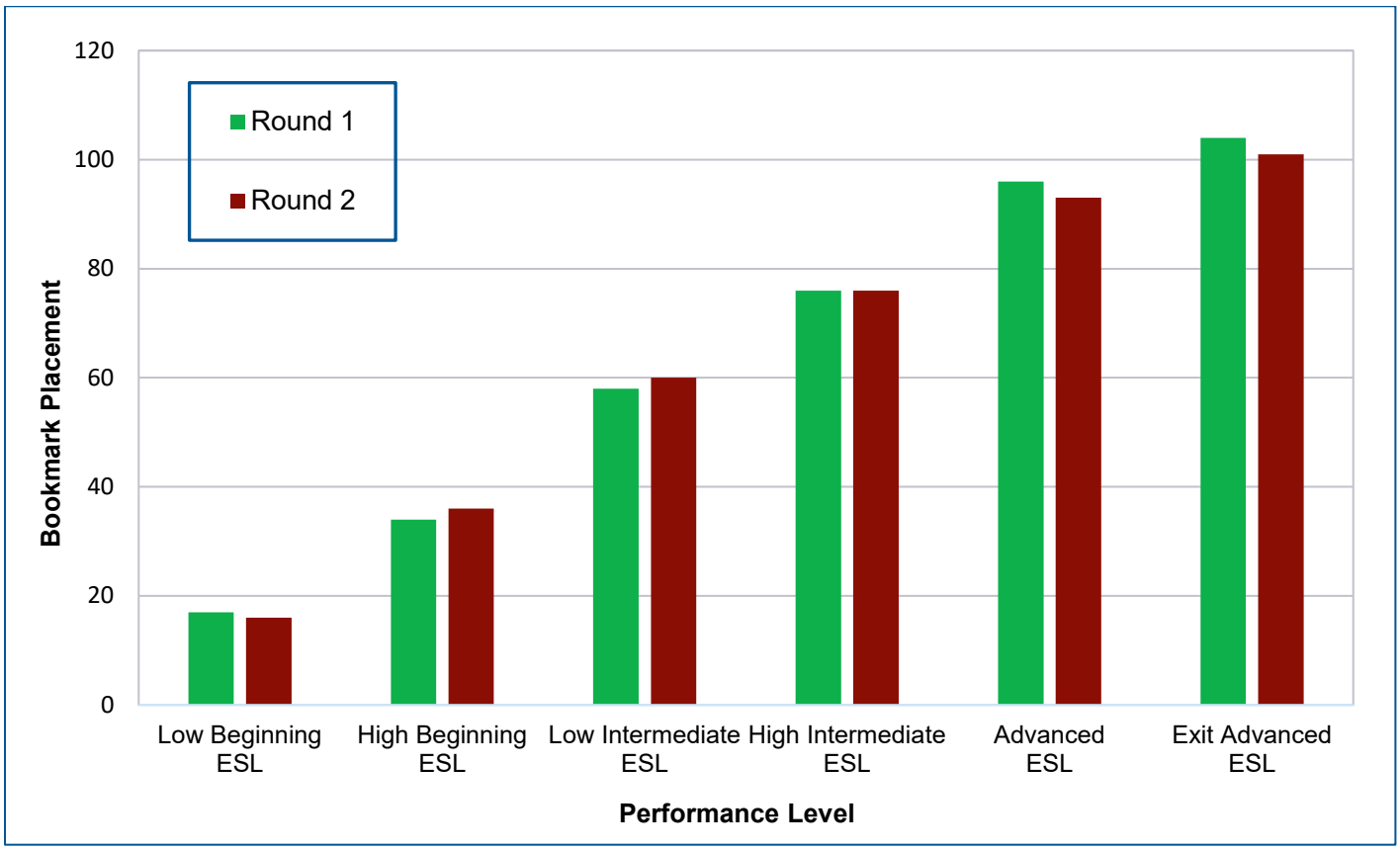

Figure 2 Median Bookmark Placements by Round

Summary

Panelist bookmark placements are the basis for the cut score recommendations for each CASAS test series. Specifically, the results by round include the range of observed bookmark placements (min-max), the average bookmark location, the median bookmark location, and the standard error (SE) of bookmark locations across all panelists. Standard error is calculated as the standard deviation of the panel's cut scores divided by the square root of the panelist count (12). The standard error values are used to create a range around the median bookmark locations to better model the variability among the panel. When the standard errors decrease between rounds this indicates that the panelists Round 2 ratings are in more agreement after the presentation of the Round 1 results and discussion in small groups.

To determine the recommended cut scores and associated ranges, all Round 2 bookmark locations are converted to CASAS scale scores (RIT scale score associated with page before the bookmark page) and the results re-estimated. In addition, the conditional standard error of measurement is calculated (CSEM) as is the combined error (combined Error = sqrt(SE2 + CSEM2)). These calculations are the basis for creating a range of recommended cut scores around the median that reflects the variability in the cut score recommendations as well as the error in exam scores at that point on the scale.

In addition, CASAS analyzes "impact data" from test takers to examine the impact that using the median recommended cut score will have. A split-half analyses is also conducted to address the replicability of the standard setting study.

The validity evidence from the standard setting studies indicates strong support for the cut score recommendations for each CASAS test series. The procedural evidence shows the appropriate panelist selection, choice of methodology and application of the methodology. In addition, panelist feedback showed that the panelists' perspectives about the implementation of the methodology is very positive. The internal evidence shows a high degree of consistency of panelist ratings, specifically through the examination of the standard errors, and the convergence of the recommendations between rounds.

Because of the strength of the panel and validity evidence of the suggested cut scores, CASAS is very confident in the final cut scores based on the content-based judgments of the expert panelists. The standard setting study results and evaluation feedback support the validity of the final cut scores for the OCTAE-approved test series.

References

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

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed). Educational measurement (4th ed., pp. 433–470). Westport, CT: Praeger Publishers.

Kolstad, A, Cohen, J., Baldi, S., Chan, T, DeFur, E., & Angeles, J. (1998, May). The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard? Washington, DC: National Center for Education Statistics.

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.

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

U.S. Department of Education, Office of Career, Technical, and Adult Education (2016). Implementation Guidelines – Measures and Methods for the National Reporting System for Adult Education.

Wang, N. (2003). Use of the Rasch IRT model in standard setting: an item-mapping method. Journal of Educational Measurement, 40, 231-253.

