Using Expected Classification Accuracy and Classification Consistency to Guide the Test Development Process for an Adult Education Assessment with Multiple Cut Scores

Jared Jacobsen
Senior Research Analyst, CASAS

Abstract – Measurement precision is an important component in the process of evaluating the validity of an assessment. One way to examine measurement precision is through the reliability coefficients of classical test theory. Another is through the examination of consistency of scores across replications of a testing procedure. The *Standards for Educational and Psychological Testing* states that the latter approach employs different methods to examine this consistency in terms of standard errors, reliability coefficients per se, generalizability coefficients, error/tolerance ratios, item response theory (IRT) information functions, or various estimates of classification consistency (AERA, APA, NCME, 2014). For criterion-referenced tests that classify examinees into performance categories, a primary focus of measurement precision should be the degree of classification accuracy (CA) or classification consistency (CC). The CASAS GOALS test series are examples of 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. By estimating and examining the CA and CC during the development process, CASAS ensures that the final test forms meet the desired CA and CC, while adhering to the test blueprint. This paper describes and summarizes a now popular approach for using an item response theory-based method to estimate CA and CC and examines how this information informs test construction considerations such as item selection and test length.

Index Terms **–** test development, classification accuracy, classification consistency, measurement precision, adult education, reliability

I. Introduction

Criterion-referenced tests compare an examinee's knowledge against a predetermined standard, learning goal, performance category, or other criterion. CASAS GOALS tests are criterion-referenced tests that align with the National Reporting System (NRS) for Adult Education Educational Functioning Levels (EFLs). The College and Career Readiness Standards for Adult Education govern the content of each EFL. The intended population for the GOALS tests is adult students who are enrolled in adult education programs and are functioning across the entire spectrum of the NRS EFLs for reading, listening, and math, from beginning literacy levels through transition into postsecondary education and training.

To evaluate an examinee's knowledge with respect to a performance category, a standard setting study relates performance on the test content to the defined performance categories. This is achieved by determining the score ranges or "cut scores" that define each performance category.

Two prevalent methods for setting cut scores are the Angoff method (Angoff, 1971) and the Bookmark method (Lewis, Mitzel, & Green, 1996). The Angoff method entails convening a panel of judges with content expertise and familiarity with the target population to make item-level judgments on the likely performance of defined target examinees.

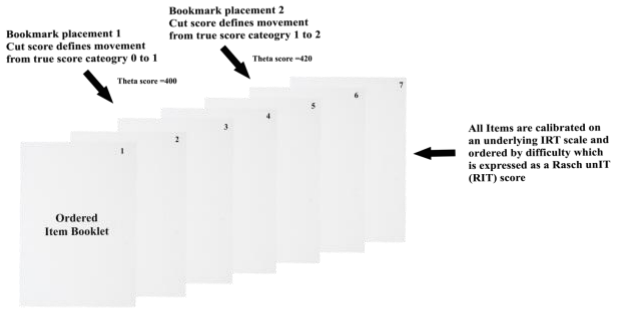
The Bookmark method has become popular since its introduction in 1996. An essential feature of the Bookmark method is the use of item response theory (IRT) to map items, by order of difficulty, onto a proficiency distribution where cut scores are then set by qualified subject matter experts. Items are ordered by difficulty with the goal of simplifying the cognitive tasks required by the participating experts. Once the appropriate cut scores have been established, an examinee's performance can be interpreted and reported with respect to the performance categories being measured by the test.

A detailed comparison of the two methods is presented in *A Comparison of Angoff and Bookmark Standard Setting Methods* (Buckendahl, Smith, Impara & Plake, 2002).

CASAS chose the Bookmark method to set cut scores to define each NRS EFL. The Bookmark method is an examinee-centered approach that involves judgments on the examinees and their typical performance (Cizek, 2006). Because judges are not asked to make item level predictions relative to the absolute item difficulty, the judgmental burden is reduced. A detailed compilation of information on the Bookmark method appears in *The Bookmark Standard-Setting Method: A Literature Review* (Karantonis & Sireci, 2006).

Figure 1 provides a simple visual description of the Bookmark method. In this example the judges have placed two bookmarks to set multiple cut scores. The corresponding theta values for these two cut scores are 400 and 420.

Figure 1 Visual of Bookmark Standard Setting Method



Part of the test development and validation process includes demonstrating that the test has appropriate cut scores, content, length, and difficulty, and how these components relate to the goal of consistently classifying examinees into the defined performance categories. These development considerations must include the desired classification accuracy and classification consistency for each form. Classification accuracy (CA), also referred to as decision accuracy, measures the extent to which observed classifications of examinees based on the result of a single replication test administration would agree with their true classification status. Classification consistency (CC), also referred to as decision consistency, measures the degree to which the observed classifications of examinees would be the same across replications of the testing procedure (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014).

The most straightforward approach for measuring CC is to administer two tests to the same examinees and then compare the degree of agreement in classifications across the two administrations. Hambleton and Novick (1973) introduced the topic of CC as the consistency of examinee classifications resulting from either two administrations of the same examination or from parallel forms of an examination. However, this approach is often neither practical nor feasible for most programs because of the requirement of two test administrations. The challenges include recruiting examinees to take two complete tests without instruction between testing events and maintaining consistent examinee motivation.

CASAS has previously conducted CC studies that require multiple test administrations and found that these challenges apply to adult education programs. Administering multiple field-tests to students, in addition to their required testing, results in an undue burden on educators and students. It is costly for adult education agencies both in terms of time and resources. CASAS found that when multiple tests were obtained, there were considerations regarding how to follow a valid study methodology while attempting to limit the burden on educators and students. For example, the multiple testing events were sometimes administered during a very short period but, in other instances, could be administered only over a longer time period. In the first scenario, which intended to limit instruction between testing events, the challenge was to avoid examinee fatigue and maintain examinee motivation throughout the process. In the second scenario, there was the risk of instruction taking place between testing events, which increased the likelihood of time sampling error. Including examinee performance data with varying degrees of time sampling error would make it impossible to compare the CA or CC of a test.

The administration of multiple field-tests to a single examinee also should consider the characteristics of the population served. The adult education population often includes inexperienced test takers. In addition, these test takers have situations that limit their availability to take multiple tests over a larger block of

time. At the time of this writing, there were also unique limitations because of the COVID-19 pandemic that made it even more difficult to recruit students to take multiple field-tests that are not part of their regular testing schedule.

Because of these limitations, more practical approaches have been developed to estimate classification accuracy and consistency based on a single test administration. These include both classical test theory (CTT) and item response theory (IRT) approaches.

In classical test theory, a linear model is postulated and links the observable test score (X) to the sum of two latent variables, true score (T) and error score (E):

$$X=T+E \quad (1)$$

The assumptions of classical test theory include (a) true scores and error scores are uncorrelated; (b) the average error score in the population of examinees is zero, and (c) error scores on parallel tests are uncorrelated (Hambleton & Jones, 1993). The primary criticism of CTT is its "circle dependency" or that the characterization of an examinee is test dependent and the characterization of the items or test is examinee dependent. For example, with CTT, the difficultly of an item is not an inherent property of the item but is relative to the group on which the item is administered.

IRT is a general statistical theory about examinee item and test performance and how performance relates to the abilities measured by the items in the test. Item response theory rests on two basic suppositions: (a) the performance of an examinee on a test item can be predicted by a set of factors called traits, latent traits, or abilities, and b) the relationship between examinees' item performance and the set of traits underlying item performance can be described by a monotonically increasing function called an item characteristic curve or ICC (Hambleton, Swaminathan, & Rogers, 1991). IRT is commonly used in education to calibrate and evaluate test items and to score examinees on their latent ability.

Theoretically, IRT overcomes the major weakness of CTT, that is, the circular dependency of CTT's item/person statistics. However, one common misperception of IRT parameter invariance is that the parameters are entirely independent of the tested population. Items calibrated from different groups of test takers must be placed on the same scale before they can be used interchangeably (Rupp & Zumbo, 2006; Rudner, 1983). As a result, once scaled, IRT models produce item statistics independent of examinee samples and person statistics independent of the specific test items administered (Fan, 2008). Thus, prior to placing calibrated items on the same scale, IRT has the same circle dependency as CTT. It is also important to note that circle dependency of CTT can also be eliminated with equating.

The use of the IRT model requires three main assumptions. The first is the assumption of unidimensionality. Unidimensionality means that only one trait or ability is measured by the items. However, tests are not perfectly unidimensional. What is required for a test to meet the unidimensionality assumption for use in IRT is a sufficiently dominant component or factor that influences the ability measured by the test (Hambleton, Swaminathan, and Rogers, 1991). The second assumption is local independence. Local independence states that a response to any one item is independent of the response to any other item, controlling for ability and item parameters. The third assumption is item parameter invariance, which states that item characteristics do not vary across subgroups of the population.

There are several methods to calculate CA and CC based on classical test theory. One of the more common is the Livingston-Lewis procedure. This procedure estimates the accuracy and consistency of classifications based on "effective test length." This can be described as the number of equally difficult items that must comprise the test to produce a total score with the same precision (Livingston & Lewis, 1995).

There are also different methods to calculate CA and CC using item response theory. These include Rudner's, Guo's, Lee's, Hambleton and Han's, and Lathrop and Cheng's methods. These approaches are summarized by Diao and Sireci (2018). The following sections will examine CA and CC using Rudner's method (Rudner, 2001; 2005). This is the method used by CASAS.

II. DESCRIPTION OF RUDNER'S CLASSIFICATION ACCURACY AND CLASSIFICATION CONSISTENCY METHOD

The hypothetical we will use to illustrate the process is a fixed-length, multiple-choice test that consists of dichotomous items, that is, items with only two possible outcomes – correct and incorrect. The design uses a Rasch IRT model and is calibrated and placed on an IRT scale. To estimate the CA and CC indices, we applied Rudner's method. This approach is described well in the Massachusetts Adult Proficiency Tests College and Career Technical Manual or MAPT-CCR (Zenisky et al., 2018) and is summarized below.

Rudner's method (Rudner, 2001; 2005) computes CA and CC when the cut scores are placed on the underlying IRT scale, referred to as the θ metric. This approach assumes that for any true score, the corresponding observed score, or latent ability, is expected to be normally distributed with a mean of θ and a standard deviation of $se(\theta)$. Previous research has examined and supported this assumption (Guo, 2006; Lathrop, 2015). The probability of an examinee with a given true score of θ having an observed score in the interval [a,b] (the score range covered between two cut scores) on the theta scale is expressed by:

$$p(a < \hat{\theta} < b \mid \theta) = \emptyset\left[\frac{b-\theta}{se(\theta)}\right] - \emptyset\left[\frac{a-\theta}{se(\theta)}\right] \quad (2)$$

where $\emptyset(z)$ is cumulative normal distribution function.

Rudner stated that by multiplying equation (2) by the expected proportion of examinees whose true score is θ yields the expected proportion of examinees whose true score is expected to be in interval [a,b] or the classification category being measured. Summing or integrating overall examinees in a different interval [c,d], the interval between the cut scores gives the expected proportion of examinees who have a true score in [c,d] and an observed score in [a,b]. Setting [a,b] and [c,d] to correspond to the true score intervals defined by the cut scores yields a classification table showing the expected proportion of all examinees with observed and true scores in each cell.

Rudner's method involves a series of steps. First, the cut scores must have been converted from the raw score to the theta score scale. Next, a K-by-K classification accuracy table is produced 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 estimated. This estimation will be discussed in the final section of this paper.

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. The conditional probabilities that are below and above the diagonal elements at each expected score category can be classified as "false positives" and "false negatives." In our context, with a test containing multiple classification categories that describe and measure NRS Educational Functioning Levels, false positives are examinees with an expected classification above the true classification and false negatives are examinees with expected classification below the true classification.

Table 1 Sample Classification Accuracy Table

		Expected Score Category					
		350-399	400-419	420-435	436-451	452-467	468-485
		0	1	2	3	4	5
True Score Category	0	26.1	2.9	0.0	0.0	0.0	0.0
	1	1.9	21.8	2.4	0.0	0.0	0.0
	2	0.0	1.9	15.0	2.1	0.0	0.0
	3	0.0	0.0	1.7	10.9	1.3	0.0
	4	0.0	0.0	0.0	1.1	6.0	0.7
	5	0.0	0.0	0.0	0.0	0.6	4.6

In Table 1 shows a hypothetical classification accuracy table which simulates a system like the NRS with multiple categories that need to be measured by the test. The expected score category for each true score category is presented in the rows. The sum of the total expected score category classification percentages for each true score category, the diagonal across all true score categories, provides the overall classification accuracy for the test. In the example, the sum of the diagonal elements results in a classification accuracy of 84.4%. So, 84.4% of examinees, based on the result of a single replication, are expected to be classified in agreement with their true classification status. In an analysis of the results by the individual expected score category, results show that at the expected score category of 1, the CA is 82.0% (21.8/(21.8+1.9+2.9)).

Upon further examination, 10.9% of examinees (1.9/(21.8+1.9+2.9)) classify as "false negatives" (true 2's that are expected to be classified as 1's) and 10.9% ((2.9/(21.8+1.9+2.9)) classify as "false positives" (true 0's expected to be classified as 1's).

For a test designed to measure learning gains across multiple test administrations, the expected CC is of heightened importance. Li (2006) expanded Rudner's approach to provide an estimate of CC in addition to CA. To estimate CC, she introduced a parallel to equation (2) to estimate the probability of an examinee with a given true score of having an observed score in an interval [c,d] on an independent, parallel administration of the test without having acquired any practice effects, as was described in equation (2).

The responses to the tests are independent and the probability of an examinee with a given true score θ having an observed score that classifies into the interval [a,b] on the first administration of the test who then classifies into the interval [c,d] on the second administration of the test can be shown by:

$$P(a < \hat{\theta} < b|\theta) * P(c < \hat{\theta} < d|\theta) = \left\{\phi\left[\frac{b-\theta}{se(\theta)}\right] - \phi\left[\frac{a-\theta}{se(\theta)}\right]\right\} * \left\{\phi\left[\frac{d-\theta}{se(\theta)}\right] - \phi\left[\frac{c-\theta}{se(\theta)}\right]\right\} \quad (3)$$

By applying this logic to all candidates in the test, or to the entire theta scale range, as presented in equation (3), we arrive at the expected proportions of all examinees who have observed scores in the interval [a,b] on one form and observed scores in the interval [c,d] on the other form:

$$\int_{\theta=-\infty}^{\infty} P(a < \hat{\theta} < b|\theta) * P(c < \hat{\theta} < d|\theta) f(\theta) d\theta = = \int_{\theta=-\infty}^{\infty} \left\{\left[\phi\left[\frac{b-\theta}{se(\theta)}\right] - \phi\left[\frac{a-\theta}{se(\theta)}\right]\right] * \left[\phi\left[\frac{d-\theta}{se(\theta)}\right] - \phi\left[\frac{c-\theta}{se(\theta)}\right]\right]\right\} \Phi\left(\frac{\theta-\mu_\theta}{\sigma_\theta}\right) d\theta \quad (4)$$

Therefore, in the above example with a DA of 84.4%, the DC is calculated using Li's extension of Rudner's method as presented in equation (4). The DC index, estimating the likelihood a person would get the same classification on two parallel forms, is 72.5%.

III. USE OF CLASSIFICATION ACCURACY AND CLASSIFICATION CONSISTENCY TO INFORM TEST DEVELOPMENT

Estimating classification accuracy and classification consistency during the test development process allows the developer to estimate measurement precision continually on different iterations of the test and make adjustments based on these estimations. This is a useful tool to help the test developer adhere to the key properties of the test blueprint. The test blueprint describes the key elements of a test, including the content to be covered, the amount of emphasis allocated to each content area, and other important features. The test blueprint also should consider the reliability of scores that will be produced and the validity of the score interpretations (Raymond & Grande, 2019).

For example, if the test developer is building a fixed-form test to measure gains across multiple administrations, the principal concern may be with examine classification consistency, the degree to which examinees are classified into the same performance levels between independent parallel forms of a test using Li's extension of Rudner's method. It is important to note that although this example refers to a fixed-form test, this method also could be used to measure the classification consistency of a computer adaptive test or Multi-stage adaptive test. An example of a Multi-stage adaptive test that uses Rudner's method to estimate CC is the Massachusetts Adult Proficiency Test or MAPT (Zenisky et al., 2018).

To calculate classification accuracy using Rudner's method, the cacIRT software package may be used and then expanded on to calculate the classification consistency (Lathrop, 2015). CacIRT has the ability to compute several classification accuracy and consistency indices under item response theory. In addition to Rudner's method, these include the total score IRT-based methods in Lee, Hanson & Brennen (2002), and Lee (2010) and the total score nonparametric methods in Lathrop & Cheng (2014).

The test developer needs the following information to calculate the CC using Rudner's method:

A. *Theta Score and Conditional Standard Error of Measurement*

The test needs to have been calibrated and placed on an IRT scale. The scale score and conditional standard error of measurement (CSEM) at each score point are entered into cacIRT. The CSEM is a measure of the variation of observed scores for an individual examinee with a particular true score. Each true score point has a CSEM. When examining the CSEM at score points compared to the CC indices for the entire test, the developer may consider changes to the item selection that could lower the CSEM, especially near the cut scores.

B. *Cut Scores*

By conducting a standard setting study, the test developer has determined psychometrically defensible cut scores or cut score ranges based on the judgments of subject matter experts. The cut scores are determined for each category that the test is designed to measure and placed on the same underlying IRT scale. CASAS, for example, used the Bookmark method to determine the cut scores for GOALS series.

The cut scores are entered into cacIRT and provide the standards to calculate the classification of each examinee.

It is important to note that the initial cut scores should not be treated as sacrosanct (Glass, 1977). When final cut score decisions are made, the test developer should take into account other important considerations, one of which may include error of measurement. Use of the reliability of a test to modify the standard may be reasonable when test reliability is low (Geisinger & McCormick, 2010).

If a cut score range is very narrow, especially relative to the conditional standard errors for the score points that fall within each range, classification consistency may be lower than desired. Any changes to the cut scores must be made with caution. This is especially true when a change yields more passing scores and the cost of incorrectly passing an examinee is high. An alternative, if feasible, may be to combine categories. The effects of a change to the cut score should be carefully considered, and the justifications for the change documented. The effect on classification consistency, false negatives, and false positives should be examined.

C. *Score Distribution of Examinees*

The distribution of examinees across each score point of the test is another variable that affects the classification consistency indices and must be entered into cacIRT when calculating the final estimation of the test's classification consistency. This distribution may be determined from field-test data or, if the test has already been used in practice, real test data that fully represents the test taker population.

If the test is in the development phase and the test developer needs to get a reasonable estimate of the expected reliability before actual examinee data is obtained, there are several options to estimate the examinee distribution. The developer could use examinee distribution from a previous version of the test if there is an expectation of a similar distribution; use a normal distribution to report examinee distribution across the entire score spectrum of the test; or, in the absence of better information, use a constant distribution of examinees to estimate the examinee distribution. The goal is to allow the test developer to get a reasonable estimate of classification consistency across different test iterations prior to administering the test to examinees. However, it is important to note that these initial CC estimates will change if the examinee distribution is adjusted or found to be inaccurate.

IV. CONCLUSION

For criterion-referenced tests that classify examinees into performance categories, the calculation of classification accuracy (CA) and classification consistency (CC) is an especially important component of estimating the measurement precision of the test. Without estimating the CA or CC, a test user cannot conclude if classification decisions can be reliably interpreted.

The objectives and format of the test will guide the test developer in determining the appropriate CA and CC indices. This determination should be considered when the test blueprint is created. Although it is always

important to strive for a high degree of CA and CC, an appropriate degree of CA and CC will not be the same for all tests. For example, if the test is deemed "high stakes," in which the examinee classification has a high impact, such as in certification tests, higher CA and CC indices are critical.

Because the calculation of CA and CC using multiple test administrations to the same group of examinees is often impossible to obtain, more practical approaches have been developed to estimate CA and CC based on a single test administration. The use of these methods, such as Rudner's IRT-based approach described in this paper, have been supported in recent literature (Deng, 2011; Diao & Sireci 2018).

Another benefit of the use of these indices is the ability to estimate CA and CC throughout the development process. This allows the test developer to examine different test iterations with respect to CA and CC while adhering to properties of the test blueprint. This process can be much more efficient and economical compared to an approach of determining estimated measurement precision upon completion of the intact test.

The CASAS GOALS series are developed using IRT methods and placed on an underlying IRT scale. In the past, CASAS has undertaken the analyses of estimating test precision by administering multiple tests to a single examinee. These results indicated findings similar to the CA and CC indices estimated from the IRT-based approach. But these studies resulted in significant burdens to educators and students. Therefore, CASAS determined that estimation of the measurement precision of the GOALS series is best achieved by estimating CA and CC using an IRT-based approach.

As described, a principal purpose of the GOALS test series is consistent classification of examinees into the National Reporting System (NRS) for Adult Education Educational Functioning Levels (EFLs). Using the IRT-based approach described in this paper, CASAS overcomes the drawbacks of administering multiple tests to a single examinee. This approach supports the development of forms with indices of CA and CC that are appropriate for the uses of the tests.

It is also important to emphasize that the tests must be maintained through continual and consistent examination of the psychometric properties to ensure they are functioning as expected in practice and over time.

We hope this paper provides useful information regarding a single-administration IRT-based approach to estimate classification accuracy and classification consistency that CASAS employs throughout the test development process.

Acknowledgments

The author would like to thank Dr. Larry Rudner for his review of an earlier draft of this paper. The author would also like to thank Patricia Rickard, Jane Eguez, Linda Taylor, Debalina Ganguli, and Daniel Esko of CASAS for their reviews and contributions to this paper.

References

Angoff, W. H. (1971). *Scales, norms, and equivalent scores.* In R. L. Thorndike (Ed.),Educational Measurement (2nd ed.). Washington, DC: American Council on Education.

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

Buckendahl, C. W., Smith, R. W., Impara, J.C., & Plake, B. S. (2002). A comparison of Angoff and bookmark standard setting methods. *Journal of Educational Measurement*, 39(3), 253–263.

Comprehensive Adult Student Assessment Systems (2019). *Math GOALS Technical Manual*.

Comprehensive Adult Student Assessment Systems (2019). *Reading GOALS Technical Manual*.

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225–258). Mahwah, NJ: Lawrence Erlbaum.

Deng, N., (2011). *Evaluating IRT- and CTT-based methods of estimating classification consistency and accuracy indices from single administrations.* [Unpublished doctoral dissertation]. Amherst, MA: University of Massachusetts.

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.

Fan, X., 1998. Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58(3), 357-381.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, 29(1), 38–44.

Glass, G.V. (1977). Standards and criteria. University of Colorado, December 1977.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment Research & Evaluation*, 11(6), 1-9.

Haertel, E. H. (2006). Reliability. In R.L. Brennan (Ed.). *Educational Measurement (4th ed.)* (pp. 65-110). Westport, CT: Praeger Publishers.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory. Newbury Park, CA: Sage.

Han, C. and Rudner, L.M. (2016). Decision consistency. In Wells, C.S. and Faulkner-Bond, M. *Educational Measurement. From Foundations to Future*. New York, NY: The Guilford Press.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice, 25(1), 4-12.*

LaFond, L. J. (2014). *Decision consistency and accuracy indices for the bifactor and testlet response theory models. (*DOI: 10.17077/etd.ytivc04x) [Doctoral dissertation]. University of Iowa, 2014.

Lathrop, Q. N. (2011). cacIRT: Classification accuracy and consistency under Item Response Theory. http://CRAN.Rproject.org/package=cacIRT

Lathrop, Q. N. (2015). Practical issues in estimating classification accuracy and consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation*, 20(18). Available online: http://pareonline.net/getvn.asp?v=20&n=18.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996, June). *Standard setting: A Bookmark approach.* Symposium presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

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

Livingston S.A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement,* 32(2), 179-197.

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

Rudner, L.M. (1983). A Closer look at latent trait parameter invariance. *Educational and Psychological Measurement, 43(4): 951-955.*

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.

Rudner, Lawrence M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation, 14(8). Retrieved from* http://pareonline.net/getvn.asp?v=14&n=8.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*(1), 63-84. https://doi.org/10.1177/0013164404273942

US Department of Education, Office of Career, Technical, and Adult Education (2016). *Implementation Guidelines: Measures and Methods for the National Reporting System for Adult Education*.

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O'Donnell, F., Wells, C. S., Padellaro, F., Jung, H. J., Banda, E., Pham, D., Hong, S. E., Park, Y., Botha, S., Lee, M., & Garcia, A. (2018). *Massachusetts Adult Proficiency Tests for College and Career Readiness: Technical Manual*. Center for Educational Assessment research report No. 974. Amherst, MA: Center for Educational Assessment.