Examining Alignment of Classification Quality to High-Stakes Test Decisions in Florida

Lissette A. Tolentino
Anne Corinne Huggins-Manley
University of Florida

Abstract

High-stakes testing in education often requires the use of cut scores to report achievement. In Florida, cut scores are used to establish different levels of proficiency. Although the Florida Standards Assessments (FSA) reports the accuracy rates for cut scores, it does not report classification consistency, nor does it report information on the alignment between the high-stakes cut scores and variations in classification quality across a range of possible cut scores. Our purpose is to perform a case study evaluating the alignment between marginal classification accuracy and consistency rates across the ability continuum to cut point locations for high-stakes cut scores, and to demonstrate the practical utility of this cut score evaluation method that was proposed by Wyse and Babcock (2016). We achieved this purpose through the use of a large set of simulated test data samples generated from FSA item and person parameter estimates.

Keywords: cut scores, high-stakes testing, education, classification, accuracy, consistency

Background

Increases in accountability in K–12 education have been a primary educational policy goal in the United States since the enactment of the Elementary and Secondary Education Act of 1965 (2018). While the use of high-stakes testing is multifaceted, its use in accountability as a way to measure achievement in accordance to standards mandated by the governing area is often taken as its most integral purpose. In using well-thought-out assessments as a way to measure progress, one can provide vital and pertinent information to the interested public (U.S. Department of Education, n.d.). High-stakes tests also serve as a way to measure performance in comparison to the benchmarks that are established by the state (Stockard, 2011). Often, a critical component of measuring student success on high-stakes tests is the process of placing students into performance categories based on cut scores.

Many states, including Florida, use cut scores, which can be defined as specific scores along the test score continuum that are used to classify students into performance categories, such that students below the cut score are placed into one category and students at or above the cut score are placed into another category. The process by which a student is categorized into different performance categories based on their test score is known as classification (Lathrop & Cheng, 2013), whereas the probability or rate at which a test taker is classified correctly into the appropriate category based on their ability level is called classification accuracy (Lathrop & Cheng, 2014). Similarly, classification consistency is the degree to which a student is classified into the same performance level if given the same test on more than one occasion (Lee, 2010). When high-stakes decisions are made based on the performance categories, it is expected that the classification accuracy and consistency rates are observed at a level that is adequate for such decisions.

In addition, there are several factors that influence the impact of classification accuracy and consistency indices including test length; test information and standard error of measurement; and

the relationship between the cut score location and the density of the test score distribution. While these are certainly not the only factors that influence classification quality, the impacts of these particular factors have been studied by researchers (e.g., Wyse, 2011; Wyse & Hao, 2012; Lathrop & Cheng, 2013). Test length was found to influence classification accuracy given that shorter length tests have worse rates of classification, and higher lengths demonstrate the opposite effect (Lathrop & Cheng, 2013). Furthermore, it is important to remember that classification accuracy and consistency are at least partially a function of the amount of test information and the standard error of measurement (Wyse & Hao, 2012). Lastly, when considering the cut score location in relation to the test score distribution, when the distribution is the densest there are more individual test takers at that score point and, hence, more chances for error in classifying individuals. Taking all of this information together, it is clear that many factors can contribute to the rates of classification accuracy and consistency.

Measurement validity of test use and reliability of test scores are two primary reasons that classification accuracy and consistency are of vital importance when discussing cut scores and performance categories on high-stakes tests. Ensuring that test use aligns with the psychometric properties of a test is critical (Kane, 2013), as a mismatch between test use and test properties could undermine the purpose of the test. There are many negative consequences that may occur if a high-stakes test has its cut scores located at an area where there is a low rate of classification accuracy and/or consistency. Broadly speaking, when scores are unreliable, cut scores cannot be expected to produce adequate levels of classification accuracy and consistency. Validity evidence of test use, and the high reliability that is a prerequisite for such validity evidence (Kane, 2013), forms the psychometric bases for the importance of having high accuracy and consistency of student classifications used for high-stakes test decisions.

Given that the classification accuracy and consistency indices are based on probabilities, it is important to understand that the results are always associated with misclassification probabilities (Florida Department of Education, 2016b). While this article does not focus on this specific aspect, it is nonetheless important to discuss its definitions and implications. The Florida Department of Education (FDOE) defines false positive rate (FPR) as the rate at which individuals are classified into a higher category than their true performance category (Florida Department of Education, 2016b; Lee, 2010). False negative rate (FNR) is defined as the rate at which individuals are classified into a lower category than their true performance category (Florida Department of Education, 2016b; Lee, 2010). FPR and FNR are two specific forms of classification inaccuracy, and hence would threaten the validity of using performance categories to make decisions about test takers.

Take for example the state of Florida, which uses the FSA (Florida Standards Assessments) as a way to evaluate student achievement based on standards set up by the State for a multitude of purposes (Florida Department of Education, 2016a), several of which are high stakes for individual students. The main goal of these assessments is to determine if the learning standards are being met and to determine if students are ready to move on to the next stage, whether it be college, a career, or a subsequent grade level (Florida Department of Education, 2016a). FPR, FNR, and, more generally, any form of classification inaccuracy or classification inconsistency would threaten such determinations about student readiness.

The FSA provides public reports on evaluations of classification accuracy at predetermined cut points for each performance category (Florida Department of Education, 2016b). However, the FSA does not report the marginal classification accuracy and consistency associated with all possible cut points along the ability continuum. Wyse and Babcock (2016) demonstrate the psychometric utility of evaluating such classification quality indices along the full ability continuum, rather than solely at pre-established cut point locations. Utilizing methods from Wyse and Babcock (2016) allows us to directly compare cut score locations to the classification

accuracy and consistency rates at various points along the ability continuum, allowing us in turn to evaluate the alignment between test properties and their use for high-stakes decisions based on student classifications. Strong alignment between these features can be seen as partial evidence for validity and reliability in high-stakes test uses, whereas weak alignment might indicate a need to revisit the test blueprints to ensure that they lend themselves to the high-stakes decisions that test users will draw from the test scores.

Study Purpose and the Need for the Case Study

The purpose of this study is to perform a case study evaluating the alignment between marginal classification accuracy and consistency rates across the ability continuum to cut point locations for high-stakes cut scores, utilizing methods introduced in Wyse and Babcock (2016). This was achieved using a large set of simulated data samples in order to further examine the relationships between the cut point locations and the classification quality indices. For this case study, we focus on the fifth-grade mathematics test within the FSA testing program, which has five achievement levels ranging from Level 1 (lowest achievement) to Level 5 (highest achievement), where according to Florida Statute 1008.34(1)(a), Level 3 is considered as "satisfactory" or often termed "proficient" for the student. This grade level and these achievement levels were chosen for the case study because the necessary test information was publicly available and, as in many other grades, there are many test-based decisions that surround a student based on their performance on FSA mathematics tests.

In general, there are many implications that a Florida student may face if they score in a low achievement level on the FSA. Take for example the third-grade English language arts (ELA) assessment: If students score at a Level 1, they may be retained for the following school year (Florida Department of Education, 2016a). If a third-grade student is to be inaccurately categorized into the wrong achievement level, they may find themselves struggling with fourth-grade material, or given the opposite, they may find themselves held behind unnecessarily. Referring back to the study subject of fifth-grade mathematics, given that students need extra support at the lowest achievement level, it is important then that they are being correctly classified into this level. In addition, such students are more likely to be placed into remedial education or retained in a grade level, both of which can be time consuming for the student's education and costly for the education system. Furthermore, when looking across the entire state, many students are located at the lower ends of the achievement scales. If these large numbers of students were to be often miscategorized, what are the implications for their future academic careers and for the Florida education system as a whole? Psychometrically speaking, what would this mean in terms of the validity evidence for test-score-based decisions?

By using a fifth-grade test as an example, we can apply the methods used in Wyse and Babcock (2016) to relate psychometric properties of the test to the locations of two particular cut scores that form the basis for placing students into performance categories, a placement that has high-stakes implications for a variety of educational stakeholders, including the possibility of student grade level retention or remediation, according to Florida Statute 1008.25(3)(5)(b). While the case study is only an example, we aim for the example to highlight the utility of introducing Wyse and Babcock's (2016) method into evaluations of validity evidence for educational decisions based on high-stakes test scores in K–12 settings.

Literature Review

Cut Scores

Often, the goal or objective of testing is a certain action or usage of the test scores, which can involve making a decision about those persons who take an assessment. These decisions can be categorical. For example, examinees can be placed into "Proficient" or "Not Proficient"

categories based on the placement of their test score being above or below a cut score. While cut scores are used in many applied settings and have several implications within the measurement field, some have argued in the past that they have not been given the attention they deserve (Dwyer, 1996). Cut scores are used "to divide a score scale or other set of data into two or more categories" (Dwyer, 1996, p. 360). Thus, the location of the cut score is determined by intended classification inferences and are most appropriately arrived at by a variety of standard-setting methods (Dwyer, 1996).

Standard setting is used as a way to determine a set of competencies and/or behaviors that can ultimately culminate in a cut score, where they are used as a comparison with the observed test score to determine the performance level of a student (Cravens et al., 2013; Wyse & Hao, 2012). During this process, experts, judges, or panelists determine a score that is representative of a minimum level of performance after considering other levels of student performance (Caines & Engelhard, 2012). Two widely used methods for standard setting include the Angoff method (Angoff, 1985) and the Bookmark method (Karantonis & Sireci, 2006). The FSA uses the Bookmark method to create cut scores by way of using an ordered item booklet based on test items (Verges, 2016). These items are then ranked from easiest to hardest based on student performance, where the selection process includes several rounds of judging from a group of panelists, ultimately arriving at a benchmark suitable for test takers (Verges, 2016). For further information on standard setting, readers may refer to Cizek and Bunch's (2007) book.

Cut scores allow for the interpretation of achievement levels based on student performance (Florida Department of Education, 2015). Thus, these cut scores need to be statistically sound if they are being used repeatedly, as is the case of K–12 high-stakes testing in U.S. public schools. In doing so, this will allow for a group of examinees, such as students, to be classified into specific groups such as proficient, or non-proficient. Therefore, establishing proficiency guidelines by way of standard setting followed by cut score determination adds to the credibility of decisions based on performance categories (Cravens et al., 2013).

Classification Accuracy and Consistency

Along with the need for classification comes the reality that errors are expected for any cut-score-based groupings of students. In making sure that the cut scores being used are accurate and consistent, inconsistencies can be avoided, and the cut scores that are implemented should more accurately reflect progress and achievement in a subject area (Lewis & Haug, 2005).

In educational measurement literature, changes and improvements have been made over time in the methods for estimating classification accuracy and consistency. For example, methods for evaluating classification accuracy and consistency were initially proposed within single administrations of tests that consisted of dichotomous items, but as assessments became more complex and included polytomous or mixed items, new and improved procedures for calculating classification accuracy and consistency have been used (Lee, 2010; Rudner, 2005).

In addition, some classification accuracy and consistency indices in the literature are based on total summated scores of examinees, while others are based on latent, item response theory (IRT) scoring systems. For example, any Rudner-based (2001, 2005) indices are by definition based on IRT principles, while the Livingston and Lewis (1995) indices are based on classical test theory scoring principles. As Wyse and Babcock (2016) state, the Rudner-based indices outperform some of the methods that hinge on summated test scores (e.g., Livingston and Lewis method, 1995). Furthermore, as this study focuses on the FSA, which uses IRT-based scoring, we focus on the Rudner-based indices in this study.

In this study, we utilized marginal indices of classification accuracy and consistency because these indices evaluate classification quality over a population or observed sample (Lathrop &

Cheng, 2013), rather than the classification quality for only a subset of examinees. The Rudner-based marginal classification accuracy index (Wyse & Hao, 2012) is defined as

$$accuracy = \sum_{i=1}^{N} \frac{w_{i1}P_{i1} + w_{i2}P_{i2}}{N} \tag{1}$$

where

$$P_{i1} = 1 - \phi\left(\frac{\theta_i - \theta_c}{se(\theta_i)}\right) \tag{2}$$

and

$$P_{i2} = \phi\left(\frac{\theta_i - \theta_c}{se(\theta_i)}\right), \tag{3}$$

where i is a student, N is the total number of students, w_{i1} equals 1 if student i is above the cut score, w_{i1} equals 0 if student i is below the cut score, w_{i2} equals 1 if student i is below the cut score, w_{i2} equals 0 if student i is above the cut score, θ_i is student i's ability, θ_c is the ability location of the cut score, $se(\theta_i)$ is the conditional standard error at the location of student i's ability, and ϕ is the area under a Gaussian distribution. The Rudner-based marginal accuracy index (Wyse & Hao, 2012) is defined as

$$consistency = \sum_{i=1}^{N} \frac{P_{i1}^{\ 2} + P_{i2}^{\ 2}}{N}. \tag{4}$$

Test Use Validity and Reliability

The FDOE relies on Messick's (1989) seminal chapter and AERA, APA, and NCME's (2014) *Standards for Educational and Psychological Testing* for their definition of validity and as their framework for supporting the inferences stemming from the test use. Messick (1989) defined validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support" the various uses stemming from test scores (p. 13). Additionally, he states that when describing validity in terms of test scores, it is in essence a "summary" of the evidence with the associated outcomes related to a test score usage (Messick, 1989, p. 13). Therefore, the test or measure itself is not subjected to be validated; instead, the uses that are extrapolated from the test scores are the subject of validity (Messick, 1989). Due to the nature of validity, every facet surrounding a test score and its use must be evaluated to ensure proper use of the test score itself (Messick, 1989).

Hence, as an important aspect of validity evidence, it is critical that K–12 students in the United States are accurately being classified because many of the uses of scores from K–12 tests are based on the performance categories derived from student classifications. Therefore, having properly categorized students by using appropriate estimates such as classification accuracy and consistency, a test is then able to "communicate the quality of the classification decision" (Lathrop, 2015, p.1). Lathrop (2015) further states that since classification accuracy is an estimator of the rate of the classification precision, it is strongly related to the overall validity of the classification itself. The classification quality is a critical piece of evidence in supporting a claim, such as calling a student proficient, indicating that a student is ready (or not) to move on to the next grade level, or indicating that a student is in need (or not in need) of remedial education.

This "claim" is the interpretation and use of the test itself, and it needs to be highly supported by the most befitting evidence (Kane, 2013).

Case Study: Florida

According to the Florida Department of Education FSA Annual Technical Report Volume 1: "The primary purpose of Florida's K–12 assessment system is to measure students' achievement of Florida's education standards" (2016a, p. 1). The report goes on to say that the assessment is the backbone in supporting instruction and learning (Florida Department of Education, 2016a). Therefore, a focus of the FSA is to ensure that the state's educational goals are being met, as well as to determine if students are ready to move on to the next grade level or stage in life (e.g., graduation, college) (Florida Department of Education, 2016a). The importance and impact that these assessments can potentially have on students within the Florida education system is paramount to their success.

The FSA has five achievement levels (associated with four cut points), which range from Level 1 to Level 5, where Level 1 is considered inadequate, and Level 5 is considered mastery (Bureau of K–12 Student Assessment, 2016). The FSA also compares scores to other national and international benchmarks in order to make comparisons on standards and compare results (Verges, 2015). Furthermore, the FDOE emphasizes the importance of the cut score between Level 2 and Level 3 since this is the level in which a student is categorized as satisfactory or below satisfactory (Florida Department of Education, 2016b).

The FSA uses IRT to calibrate items on the assessments as well as for estimating student scores (Florida Department of Education, 2016a). A student's theta score on the test is the maximum likelihood estimate (MLE) derived from IRT modeling (Florida Department of Education, 2016a). Once estimated, a student's theta score is used to create a scale score that is then used to assign a student to a performance category (Florida Department of Education, 2016a). The theta to scale score for fifth-grade mathematics is calculated as follows (cited from: Florida Department of Education, 2016a):

$$Scale\ score\ =\ round(\theta * 22.05 + 321.80). \quad (5)$$

In order to ensure minimum acceptable accuracy rates of student classifications, the FSA uses two approaches: an observed score approach and a method based on IRT that is used to calculate the probabilities associated with a student being misclassified (Florida Department of Education, 2016b). While the FSA combines these approaches (Florida Department of Education, 2016b), our simulation focuses on the observed score approach since it is a Rudner-based marginal classification accuracy index (Wyse & Hao, 2012).

Method

We used simulation methods in RStudio (R Core Team, 2017) to mimic repeated administrations of the FSA mathematics fifth-grade assessment to populations of students, using item and student properties reported by the FDOE. Specifically, 3,000 true theta parameters were generated from a normal distribution with a mean and standard deviation set to 0.02 and 1.09, respectively, as these values were reported as the observed moments of theta from the spring 2016 administration of the test (Florida Department of Education, 2016b). IRT item parameters were generated based on a uniform distribution for a 50-item test. The choice of distribution forms the minimum and maximum values for each item parameter (shown in Table 1) and were obtained from personal communication with FDOE (Florida Department of Education) personnel (Binici, S., personal communication, November 17, 2017). The first 13 items were assumed to operate under a two-parameter logistic model (2PL; Birnbaum, 1968) and the remaining 37 were from a three-parameter logistic model (3PL; Birnbaum, 1968), which mimic the properties of the FSA fifth-

Table 1. *Descriptives for IRT Parameter Distribution: 5th Grade Mathematics*

IRT Parameter	*M*	*SD*	Min	Max
b	0.20	1.00	-2.00	3.00
a	0.90	0.30	0.50	1.80
c	0.10	0.10	0.00	0.30

(Binici, S., personal communication, November 17, 2017)

grade mathematics assessment (Binici, S., personal communication, November 17, 2017). After the items and persons were generated, 1,000 data sets of binary item responses for $N = 3{,}000$ simulees were sampled by comparing each simulee's probability of correct response (i.e., 1) to a random number from a uniform distribution ranging from 0 to 1. If the probability of correct response was greater than the random number, the simulee was assigned a score of 1 for the item, otherwise the simulee was assigned a score of 0.

Each simulated dataset was calibrated under a 3PL model using the "ltm" package (Rizopoulos, 2006). The first 13 items had the *c* parameter restricted to zero in order to be considered as 2PL. An issue that often occurs when running 3PL models is that the models may not converge due to the Hessian matrix not being positive definite (Rizopoulos, 2006). In order to remediate this issue, code was included that allowed for the "catching" of these convergence errors. As a result, the final simulation analysis and results are based on 842 iterations.

Once the models had been fit to all 842 datasets, the estimated theta was used to define 601 quadrature points that were then used to estimate marginal (i.e., across the entire sample of test takers) classification accuracy and consistency at each quadrature point. This allowed for the estimation of classification accuracy and consistency for every possible cut point that the FSA could have used along the theta continuum. Classification accuracy and consistency were calculated using the Rudner-based marginal classification accuracy index (Wyse & Hao, 2012 (see Equations 1–4) using the "cacIRT" (Lathrop, 2014) package. Once all iterations were complete, we calculated the average of the 842 results for each of the 601 quadrature points and created graphics of the results. The FSA cut points were labeled on the graphs to evaluate the alignment between each of the marginal classification indices and the high-stakes cut points used by the State of Florida. In order to obtain the cut scores on the theta scale, we used the publicly reported scale score cut points **(**Florida Department of Education, 2016b) and converted them to the theta scale via Equation 5. The alignment of the FSA reported cut points and the thetas we extracted from them are shown in Table 2.

Table 2. *Simulation Results*

FSA Mathematics 5th Grade	Scale Score	FSA Reported θ	Simulated Marginal Accuracy Rate	Simulated Marginal Consistency Rate
Cut Between Level 1 and Level 2	306	-0.71	0.93	0.90
Cut Between Level 2 and Level 3	320	-0.08	0.91	0.88

(Florida Department of Education, 2016b)

Results

Figures 1 and 2 display the "Marginal Rudner Accuracy" and the "Marginal Rudner Consistency" results, respectively, along with two vertical lines that indicate where the proficiency and retention cut points are located for the fifth-grade mathematics test. The theta (scores) locations are on the *x*-axis and range from -3.00 to 3.00, and the averaged marginal classification accuracy and consistency rates (i.e., averaged across 842 iterations) are on the *y*-axis that ranges from 0.50 to 1.00. The two points along which the high-stakes cut point lines intersect with the classification index averages are marked by red triangles. Table 2 lists the observed values of those two points.

Figures 1 and 2 show that the second cut point, the proficient cut point, is associated with a lower classification accuracy and consistency level than the first cut point, the retention cut point. Specifically, the plotted accuracy and consistency rates are quite high (i.e., near 1.00) at the lowest end of the ability scale, but they decrease as they approach the first cut point. The rates then further decrease as they approach the second cut point. The figures also show that the proficient cut point is located at a point along the ability scale that is associated with one of the lowest possible classification accuracy and consistency indices, as compared to other points along the ability scale at which cut points could theoretically be made.

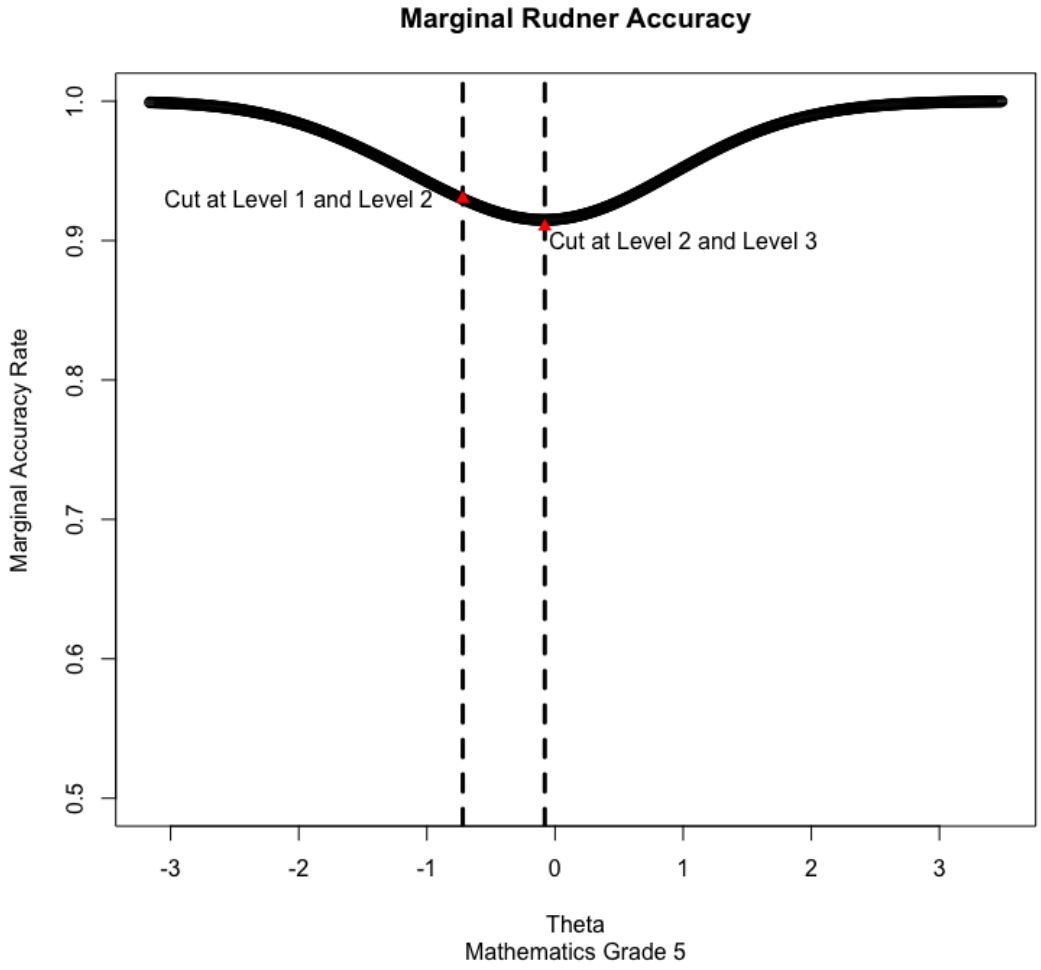

Figure 1. Marginal accuracy rate results, averaged over 842 iterations.

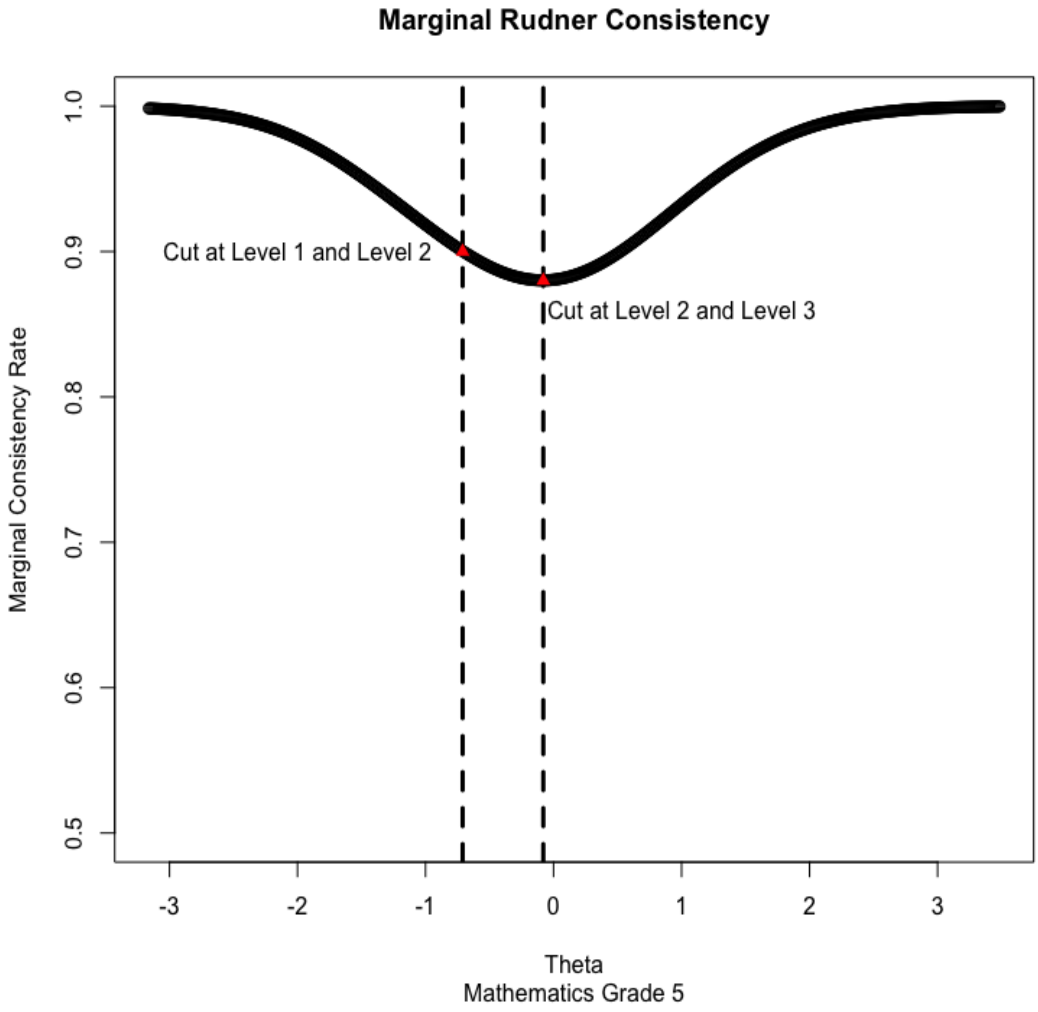

Figure 2. Marginal consistency rate results, averaged over 842 iterations.

The study results also indicate that while the fifth-grade mathematics FSA cut scores are at locations where marginal classification accuracy and consistency rates are lower on the test relative to other places on the continuum, the classification indices are relatively high across the ability continuum. Also, while not shown in the figures, readers should know that, generally speaking, points along the ability continuum associated with relatively lower marginal accuracy and consistency rates are aligned with dense points within the theta distribution. This is to be expected given the nature of the classification indices, and we elaborate on this result in the following section.

Discussion

The purpose of this study was to explore the alignment between the high-stakes cut points on a Florida fifth-grade mathematics assessment with variations in classification accuracy and consistency rates along the ability scale. The marginal accuracy and consistency results based on the cut scores are consistent with what was found in other studies (Wyse & Hao, 2012; Lathrop & Cheng, 2013; Wyse & Babcock, 2016). Generally speaking, points along the ability continuum associated with relatively lower marginal accuracy and consistency are located in places where

the theta distribution is densest. Hence, it may be misleading to say that the test item properties can be changed in such a way that results in higher classification accuracy and consistency near the high-stakes cut scores, given that the observed ability scores in the sample are distributed in a particular manner. Also, cut scores must be based on substantive alignment between the scale score and the curriculum standards, not solely based on psychometric outcomes.

However, knowing that in our case study one of the high-stakes cut points (i.e., the proficiency cut point) is located at an ability level that is associated with one of the lowest marginal accuracy and consistency levels may be considered a problem. One can easily imagine using the method shown in this study to consider some possible redesigns of the test itself. While cut scores are ultimately determined by substantive alignments to curricula standards (Karantonis & Sireci, 2006), the choice of how to balance content on a test can be informed by the degree to which one wants to have high marginal classification accuracy and consistency at such cut scores. Yet, in our case study even the lowest observed marginal accuracy rate was above 0.9, indicating that the issue in our particular case study might not be substantial. This does align with the fact that the reported accuracy from the FDOE along these cut points is considered high (i.e., above 0.9) (Florida Department of Education, 2016b). But from our simulation, one can see that there are other locations that the cut points could be made in which these classification indices would be higher. In addition, while 0.9 may be high for group level decisions, from an individual student perspective, this might not be sufficient.

We do want to be sure that readers are not misled by the very high rates of marginal classification indices at the tail ends of the ability continuum. A small proportion of the test-taking population has scores on those tail ends, so putting cut points at those tail ends automatically results in accurate and consistent classification of the vast majority of students, and hence the high rates of those classification indices. Our purpose in applying some of the methods in Wyse and Babcock (2016) to our case study simulation is not to imply that cut scores should be placed where these particular classification indices are maximized, as that would result in useless cut scores at the lowest and highest points on the ability continuum. But none of this negates the finding that one of the most widely used high-stakes cut points (i.e., the proficiency cut point) lies closely to the minimum marginal accuracy and marginal consistency rates that is possible along the ability continuum. We argue that simply having this information is important for test developers when evaluating the alignment of the psychometric properties of the test and the validity evidence for high-stakes test uses based on performance classifications of students.

Wyse and Babcock (2016) showed similar results in some of their applications, and they noted that as students got closer to the cut score, they became harder to classify. Hence, cut points located in dense portions of the ability distribution will suffer from more misclassification than cut points located at less dense portions of the ability distribution. But ultimately, what is important is how these performance levels stand up to the test use, and what they add to the overall validity of such test use. The study results indicated that while the fifth-grade mathematics FSA cut points are at locations where marginal classification accuracy and consistency are lower on the test relative to other places on the continuum, the classification indices are relatively high across the ability continuum, allowing for some evidence of validity of classification and generalization of test uses in general. Yet, for individual student level decisions, it may be important for the FSA to continue to build their tests in such a way that the most critical cut scores for lower performing students are not located at points along the ability scale that are associated with nearly the lowest marginal classification accuracy and consistency results.

High-stakes tests may not always be built in such a way that ensures alignment between where the cut scores are located and where the (marginal) classification accuracy and consistency rates are performing best. Thus, it may be misleading to say that the test item properties can be changed in such a way that results in higher classification accuracy and consistency, given that

the observed ability scores in the sample are distributed in a particular manner. However, our study demonstrated that one can use application and simulation methods shown in Wyse and Babcock (2016) and in our above case study to evaluate the variations in classification quality indices in relation to where high-stakes decisions are being made. We encourage researchers and practitioners to utilize these methods to evaluate the validity of their own high-stakes decisions based on performance classification derived from cut scores.

References

Corresponding Author: Lissette A. Tolentino
Author Contact Information: ltolen@ufl.edu

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Angoff, W. H. (1985). *Scales, norms, and equivalent scores.* Princeton, NJ: Educational Testing Service.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–470). Reading, MA: Addison-Wesley.

Bureau of K–12 Student Assessment. (2016). *2015–16 FSA English language arts and mathematics fact sheet.* Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/ELA-MathFSAFS1516.pdf.

Caines, J., & Engelhard, G. J. (2012). How good is good enough? Educational standard setting and its effect on African American test takers. *Journal of Negro Education*, *81*(3), 228–240. doi:10.7709/jnegroeducation.81.3.0228

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage Publications.

Cravens, X. C., Goldring, E. B., Porter, A. C., Polikoff, M. S., Murphy, J., & Elliott, S. N. (2013). Setting proficiency standards for school leadership assessment: An examination of cut score decision making. *Educational Administration Quarterly*, *49*(1), 124–160. doi:10.1177/0013161X12455330

Dwyer, C. A. (1996). Cut scores and testing: Statistics, judgment, truth, and error. *Psychological Assessment*, *8*(4), 360–362. doi:10.1037/1040-3590.8.4.360

Elementary and Secondary Education Act of 1965, Pub. L. No. 115-224 (2018).

Florida Department of Education. (2015). *Florida standards assessments achievement level descriptions 2015.* Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/2015FSARangeSummary.pdf.

Florida Department of Education. (2016a). *Florida standards assessments 2015–2016 volume 1: Annual technical report.* Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/V1FSA1516TechRpt.pdf.

Florida Department of Education. (2016b). *Florida standards assessments 2015–2016 volume 4: Evidence of reliability and validity.* Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/V4FSA1516TechRpt.pdf.

Florida Statute § 1008.25(3)(5)(b) (2018).

Florida Statute § 1008.34(1a) (2018).

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi:10.1111/jedm.12000

Karantonis, A., & Sireci, S. G. (2006). The Bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4–12. doi:10.1111/j.1745-3992.2006.00047.x.

Lathrop, Q. (2015). Practical issues in estimating classification accuracy and consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation, 20*(18), 1–5. Retrieved from http://pareonline.net/getvn.asp?v=20&n=18

Lathrop, Q. N. (2014). R package cacIRT: Estimation of classification accuracy and consistency under item response theory. *Applied Psychological Measurement*, *38*(7), 581–582. doi:10.1177/0146621614536465

Lathrop, Q. N., & Cheng, Y. (2013). Two approaches to estimation of classification accuracy rate under item response theory. *Applied Psychological Measurement*, *37*(3), 226–241. doi:10.1177/0146621612471888

Lathrop, Q. N., & Cheng, Y. (2014). A Nonparametric approach to estimate classification accuracy and consistency. *Journal of Educational Measurement*, *51*(3), 318–334. doi:10.1111/jedm.12048

Lee, W. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, *47*(1), 1–17. doi:10.1111/j.1745-3984.2009.00096.x

Lewis, D. M., & Haug, C. A. (2005). Aligning policy and methodology to achieve consistent across-grade performance standards. *Applied Measurement In Education*, *18*(1), 11–34. https://doi.org/10.1207/s15324818ame1801_2

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. Retrieved from https://www.jstor.org/stable/1435147

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.

R Core Team (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. URL https://www.R-project.org/.

Rizopoulos, D. (2006). Ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*(5), *1–25*. doi:10.18637/jss.v017.i05

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, *7*(14), 1–5. Retrieved from http://PAREonline.net/getvn.asp?v=7&n=14.

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment, Research & Evaluation*, *10*(13), 1–4. Retrieved from https://pareonline.net/pdf/v10n13.pdf

Stockard, J. (2011). Merging the accountability and scientific research requirements of the No Child Left Behind Act: Using cohort control groups. *Quality & Quantity*, *47*(4), 2225–2257. doi:10.1007/s11135-011-9652-5

U.S. Department of Education (n.d.). *Every Student Succeeds Act (ESSA).* Retrieved from https://www.ed.gov/essa?src=rn

Verges, V. (2015). Rule 6A-1.09422: Establishing achievement level cut scores for Florida standards assessments. Rule development workshops. [PowerPoint slides]. Retrieved from http://www.fldoe.org/core/fileparse.php/5663/urlt/FSARuleDevelopmentWorkshopPres.pdf.

Wyse, A. E. (2011). The potential impact of not being able to create parallel tests on expected classification accuracy. *Applied Psychological Measurement*, *35*(2), 110–126. doi

10.1177/0146621610377449

Wyse, A. E., & Babcock, B. (2016). Does maximizing information at the cut score always maximize classification accuracy and consistency? *Journal of Educational Measurement*, *53*(1), 23–44. doi:10.1111/jedm.12099

Wyse, A. E., & Hao, S. (2012). An evaluation of item response theory classification accuracy and consistency indices. *Applied Psychological Measurement*, *36*(7), 602–624. doi:10.1177/0146621612451522