Journal of Applied Testing Technology, *Vol 19(1), 20-25, 2018*

Item Response Theory-Based Methods for Estimating Classification Accuracy and Consistency

Hongyu Diao[*] and Stephen G. Sireci

University of Massachusetts Amherst, USA; hdiao@educ.umass.edu

Abstract

Whenever classification decisions are made on educational tests, such as pass/fail, or basic, proficient, or advanced, the consistency and accuracy of those decisions should be estimated and reported. Methods for estimating the reliability of classification decisions made on the basis of educational tests are well-established (e.g., Rudner, 2001; Rudner, 2005; Lee, 2010). However, they are not covered in most measurement textbooks and so they are not widely known. Moreover, few practitioners are aware of freely available software that can be used to implement current methods for evaluating decision consistency and decision accuracy that are appropriate for contemporary educational assessments. In this article, we describe current methods for estimating decision consistency and decision accuracy and provide descriptions of "freeware" software that can estimate these statistics. Similarities and differences across these software packages are discussed. We focus on methods based on item response theory, which are particularly well-suited to most 21st century assessments.

Keywords: Criterion-Referenced Testing, Classification Consistency, Classification Accuracy, Item Response Theory, Psychometric Software, Reliability

1. Introduction

Whenever classification decisions are made on educational tests, such as pass/fail, or basic, proficient, or advanced, the accuracy and consistency of those decisions should be estimated and reported (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Methods for estimating the accuracy and consistency of classification decisions made on the basis of educational tests are well-established (e.g., Lee, 2010; Livingston & Lewis, 1995; Rudner, 2001, 2005). However, they are not covered in most measurement textbooks, and so they are not widely known. In this article, we describe methods for estimating the accuracy and consistency of classification decisions. We focus on those methods based on Item Response Theory (IRT), because they are most appropriate for contemporary educational testing programs. In addition to describing the methods, we also describe freely available software for implementing the methods.

Classification decisions on educational tests are typically made by comparing a student's test score to one or more cut-scores derived from a standard setting procedure. By doing so, examinees are classified into two or more achievement levels (e.g., fail, pass; basic, proficient, advanced, etc.). There are two terms that describe the degree of reliability (stability) of these decisions: *classification accuracy and classification consistency*. These two terms are sometimes referred to as Decision Accuracy (DA) and Decision Consistency (DC), respectively.

Classification Accuracy (CA) is the extent to which the "true" classifications of examinees agree with the observed classifications. Classification Consistency (CC) refers to the degree to which examinees are classified into the same performance levels between independent, parallel forms of a test (Hambleton & Novick, 1973; Lee, 2010). In this article, we summarize five approaches for computing IRT-based estimates of CA and CC based on a single test administration. Two of these approaches,

[*] *Author for correspondence*

Rudner (2001, 2005) and Lee (2010), are widely used in statewide assessments (e.g., Connecticut State Department of Education, 2013, p.19; Pearson, 2015, p. 171). The third approach was developed by Guo (2006), which is an extension of Rudner's approach using a likelihood function. A fourth, the Hambleton and Han procedure (in Bourque, Goodman, Hambleton & Han, 2004) involves data simulation; while the fifth (Lathrop & Cheng, 2014) is a non-parametric approach proposed for situations where the assumptions of an IRT model are violated.

2. Descriptions of CA and CC Estimation Methods

In this section, we describe and summarize afore mentioned methods for estimating CA and CC using an IRT approach. These methods are Rudner (2001, 2005), Guo (2006), Lee (2010), Hambleton and Han (in Bourque, et al., 2004), and Lathrop and Cheng (2014). We highlight similarities and differences across the methods. Given that Rudner introduced the first IRT-based method, we begin with his approach.

2.1 Rudner's (2001, 2005) Method

Rudner (2001, 2005) proposed an approach to compute CA and CC when the cut scores are placed on the underlying IRT scale, referred to as the θ metric. The approach assumes that for any θ score, the corresponding observed $\hat{\theta}$ follows a normal distribution with mean of θ and standard deviation of SE ($\hat{\theta}$). For example, the distribution observed $\hat{\theta}$ for an individual is displayed in Figure 1 (Lathrop, 2015), with a given passing cut score θ_c = 0.0245. The probability of having observed $\hat{\theta}$ above the θ_c, is unshaded area on the right side of the cut score, assuming a normal distribution. This area represents the probability that an examinee has correctly classified (CA) as "pass."

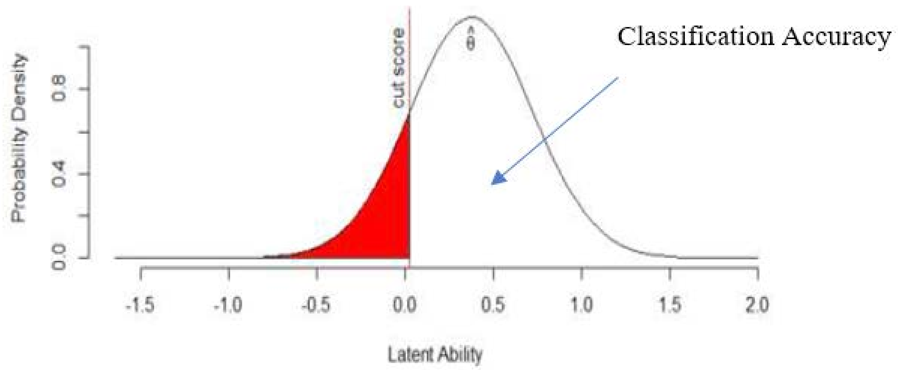
Because CC represents the probability of being classified into the same category on two independent tests, the value of CC is the proportion of the right area squared (both pass) plus the proportion of the left area squared (both fail). The CA and CC for one test administration are computed by summing up the "individual" CC and CA across all examinees.

2.2. Guo's (2006) Method

Guo (2006) introduced a latent distribution method to compute CA. This method relaxes the assumption of normality. In Guo's approach, the θ scale is discretized into multiple points based on $\hat{\theta}$ with equal distances. The likelihood function is computed for each examinee across all $\hat{\theta}$ points. Next, the likelihood function is normalized so that the sum of the likelihood is equal to 1 for each individual. Then, the likelihood functions across examinees are summed and the total area is equal to the number of examinees. The computation of CA and CC can be presented by Figure 1 by replacing the probability density on the y-axis with the likelihood function. The limitation of this approach is that the true item parameters need to be given. In practice, it is impossible to obtain the true parameters of items after calibration. Therefore, the estimated item parameters are required to be close to the true parameters before implementing Guo's method.

2.3 Lee's (2010) Method

The goal of Lee's (2010) approach is to find the distribution of all possible total scores conditional on a given θ.



Note. From Lathrop (2015).

Figure 1. Illustration of CA and CC under Rudner's Method.

Each possible total score is calculated by summing the probabilities of all possible response patterns. The estimated CA is considered as the probability of an examinee's observed score and true score falling into the same category conditional on a given θ. The CC is viewed as the probability of examinee's observed score classified in the same category on independent administrations of two parallel forms of a test. The logic is similar to the Rudner approach, but the procedure is performed with the observed score. In Figure 2, the area on the right of the cut-score denotes the CA if the examinee passes the test and the sum of the square of two areas represents the CC.

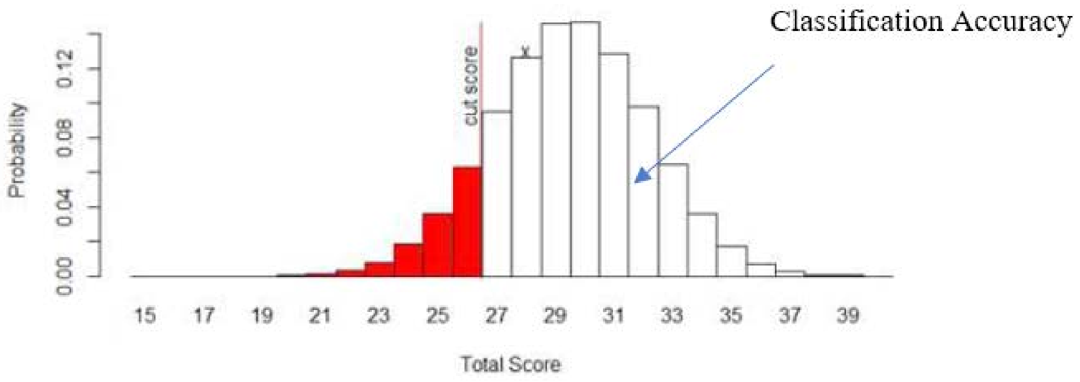
Both the Lee approach and the Rudner approach were developed under an IRT framework. The main difference between the two is the scale where the cut-scores are placed. For the Lee approach, the cut-score is on the observed-raw score scale. In the Rudner approach, the cut-score is placed on the θ metric. One method to put the cut-score on the θ score scale using the Lee approach is to map the θ score on raw score scale by means of the Test Characteristic Curve (TCC). Lathrop and Cheng (2013) compared the empirical CA results using the observed total score x and latent trait score θ as the cut-score and evaluated the performance of the Lee and the Rudner approach. They found both the Lee and the Rudner approach produced negligible bias if the assumption of model fit was satisfied. In addition, they found that using θ for conditioning resulted in slightly higher CA than using the raw score (x). The difference might be due to the method of classifying examinees: one classified based on θ (the Rudner approach) the other based on number-correct score x (the Lee approach). Although the Lee approach and the Rudner approach may produce similar classification indices, Lathrop (2015) recommended researchers use the Rudner approach for Computerized-Adaptive Tests (CAT) because examinees answer different items in CAT and so the TCCs may vary across tests.

2.4 Hambleton and Han Method

The approach proposed by Hambleton and Han in Bourque, et al., (2004) is appropriate when IRT item parameters are known and examinees' responses to those items can be simulated. Deng (2011) described the Hambleton and Han (HH) approach as a three-step procedure. First, two sets of response data for two parallel test forms are generated. Second, the cut scores are mapped from the observed score metric to the θ scale through the TCC (if necessary), and the examinees are classified into "true" performance categories based on cut scores. Third, the classification indices are calculated. In this step, CA is computed based on the classification of examinee's true ability and classification based on the observed classification. CC is calculated based on classification of parallel forms. The HH approach is easy to understand, calculate, and interpret; however, the values of classification may vary across simulations. To obtain CC and CA for one test, Bourque, et al., (2004) suggested averaging the classifications over multiple replications.

2.5 Lathrop and Cheng's (2014) Method

The methods reviewed thus far were all developed within an IRT framework and so they assume that IRT model that it fits the data. In addition, Rudner's (2001, 2005) and Lee's (2010) approaches assume θ is normally distributed. To address these limitations, Lathrop and Cheng (2014) developed a non-parametric approach for estimating CA and CC. Their method modifies Lee's method for computing classification



Note. From Lathrop (2015).

Figure 2. Illustration of Lee's Method Using the Distribution of Observed Scores for a Given $\hat{\theta}$.

indices by using Ramsay's (1991) kernel-smoothed item response functions. They found that their nonparametric approach produced less bias in performance classification than Lee's approach when the IRT model was mis-specified and the ability distribution was not normal. However, if the assumptions of the IRT model are satisfied, then the Lee approach is still appropriate.

3. Software for Computing CA and CC

Educational testing practitioners who need to estimate CA and CC will be pleased to know that all of the methods previously described have freely available software that can be used to implement the method. Perhaps the most popular software for computing classification indices is IRT-CLASS which was developed by Lee and Kolen (2008) and is available at: https://education.uiowa.edu/centers/center-advanced-studies-measurement-and assessment/computer-programs#class. The IRT-CLASS software implements Lee's (2010) method and can be used for tests containing dichotomously scored items, polytomously scored items, or both.

To implement Lee's method, Wheadon's (2014) "classify" *R* statistical package can be used to compute the classification indices. The package computes the probability of obtaining each possible score x with given $\hat{\theta}$ using the Wingersky-Lord recursive algorithm (Lord & Wingersky, 1984). One benefit of this package is that it can be applied to multiple types of IRT models such as the Rasch Model (Rasch, 1960), partial credit model (Masters, 1982), generalized partial credit model (Muraki, 1992) and so forth. In addition, the package allows users to examine model-fit by comparing the observed score distribution with the predicted score distribution and also provides an output including classification indices, consistency plots and Cohen's Kappa measure of consistency.

Lathrop (2015) developed a statistical *R* package "cacIRT" that is able to compute classification accuracy and consistency indices using the Lee approach, the Rudner approach, and the nonparametric approach (Lathrop & Cheng 2014).

4. Discussion

In Table 1, we list the features of the IRT-based approaches for estimating CA and CC that were summarized in this report. Table 1 also lists the relevant computer programs for calculating these indices using each approach. In evaluating which program may be best in a particular situation, the differences across the methods should be noted. The methods differ by whether the cut scores are set on the raw score or theta metric, whether examinee data are used or simulated, and whether parametric or nonparametric models are used.

If the cut-score is based on the raw score scale, Lee's (2010) method is certainly applicable. If based on the theta metric, Rudner's (2001, 2005) method may be preferable. If there is any concern about the fit of the IRT model used in calibrating items or scoring examinees, Wheadon's (2014) software, or any other approach (e.g., Liang, Han, & Hambleton, 2009), can be used to evaluate the fit of the IRT model to the data. If the fit is poor, Lathrop's *R* package could be used to apply the nonparametric approach. If the fit is good and the item parameters are considered to be well-estimated, Guo's (2006) method is a good option.

The HH method is the only one that can be used when there are no examinee response data. As a result, it can be used in situations where testing programs do not have access to examinee response data but do have the item parameters and test score distributions. In fact, it was that very situation that gave birth to that method.

According to the features of each approach, it is hard to conclude which method has the best performance under all conditions because each method was developed for a certain purpose. Thus, the evaluation of each IRT-based approach should take the specific features and the use of the assessment into account. If CA and CC are estimated only for simulation studies where the true item parameters can be easily obtained, the HH approach and Guo approach are reasonable choices. Between these two approaches, the HH approach is more straightforward and easier to program in most statistical programming packages. The Rudner (2001, 2005) and Lee (2010) methods are used in many large-scale assessment programs (e.g., Pearson, 2012, 2015; Sireci et al., 2008). As Lathrop (2015) pointed out, if the test is adaptive, Lee's approach can be difficult to apply because examinees answer different set of items (and so there will be many different TCCs and different cut-scores on the theta scale).

The nonparametric approach proposed by Lathrop and Cheng (2014) was developed for test conditions where IRT assumptions are not met. Their approach might be best in situations where the scores are not normally distributed or test length is relatively short (e.g.,

Table 1. Summary of IRT-based approaches to estimating CA and CC

Approach	Features	Computer/Statistical Program
Rudner (2001, 2005)	Cut-score on theta scale; mixture IRT models; need to meet assumptions of IRT models	*R* package "cacIRT (Wheadon, 2014)" Available at: https://cran.r-project.org/web/packages/cacIRT/index.html
Hambleton and Han (in Bourque, et al.,2004)	Cut-score on theta scale; mixture IRT models; need to meet assumptions of IRT models; developed for simulation data only	Easy to program in any programming software
Guo (2006)	Cut-score on theta scale; mixture IRT models; no normality assumptions; true item parameter or estimated item parameter close to the true parameters	SPSS code is provided in the appendix of the article (see https://pareonline.net/getvn.asp?v=11&n=6)
Lee (2010)	Cut-score on observed-score scale; mixture IRT models; need to meet assumptions of IRT models	*R* package "classify" Available at: https://cran.r-project.org/web/packages/classify/index.html R package "cacIRT" Available at: https://cran.r-project.org/web/packages/cacIRT/index.html **IRT-CLASS v2.0** for PC Available at: https://education.uiowa.edu/centers/center-advanced-studies-measurement-and-assessment/computer-programs#class
Lathrop & Cheng (2014)	Cut-score on observed-score scale; does not need to meet assumptions of IRT models	*R* package "cacIRT" Available at: https://cran.r-project.org/web/packages/cacIRT/index.html

10 items). Clearly, one area of future research is additional studies comparing the strengths and limitations of these methods.

5. Conclusion

Criterion-referenced testing is an increasingly common component of educational testing programs. Trend assessments such as the National Assessment of Educational Progress and the Trends in International Mathematics and Science Assessments, report percentages of students falling into specific achievement levels. In addition, the No Child Left Behind legislation, and its extensions (i.e., Race-to-the-Top, Every Student Succeeds Act) all require classifying students into achievement levels. By necessity, licensure and other credentialing exams classify students into "pass" and "fail." Indeed, in many modern testing programs it is the classification decision, rather than a total test score, that carries the highest "stakes." Thus, estimating CA and CC is one of the most important sources of information for evaluating the technical quality and utility of contemporary educational assessment.

Although the need to estimate CA and CC is important for supporting the use of a test for a particular purpose, methods for doing so are not widely discussed in the literature or in most modern measurement textbooks. We hope this brief article addresses this lack of discussion in a way that encourages practitioners to evaluate current testing programs with respect to the accuracy and consistency of the classification decisions they provide.

In closing, we would like to thank the authors of the methods for not only providing methods for estimating CA and CC, but also for providing freely available software to help us implement the methods. As new methods are proposed, we hope future authors do likewise.

6. References

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.

Bourque, M. L., Goodman, D., Hambleton, R. K., & Han, N. (2004). *Reliability estimates for the ABTE tests in elementary education, professional teaching knowledge, secondary mathematics and English/language arts (Final Report).*

Connecticut State Department of Education. (2013). *The Connecticut mastery test: Technicalreport.*

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.

Guo, F. (2006). Expected classification accuracy using the latent distribution. *Practical Assessment Research & Evaluatio*n, *11*(6). Available from http://pareonline.net/getvn.asp?v=11&n=6

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. https://doi.org/10.1111/j.1745-3984.1973.tb00793.x

Lathrop, Q. N., & Cheng, Y. (2014). A Nonparametric Approach to Estimate Classification Accuracy and Consistency. *Journal of Educational Measurement*, *51*(3), 318-334. https://doi.org/10.1111/jedm.12048

Lathrop, Quinn (2015). Practical Issues in Estimating Classification Accuracy and Consistency with R Package cacIRT. *Practical Assessment, Research & Evaluation*, *20*(18). Available from http://pareonline.net/getvn.asp?v=20&n=18.

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

Lee, W., & Kolen, M. J. (2008). IRT-CLASS: *A computer program for item response theory classification consistency and accuracy (Version 2.0)* [Computer software]. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa. Available at http://www.education.uiowa.edu/casma. https://doi.org/10.1017/S0009840X07002582

Liang, T., Han, K. T., & Hambleton, R.K. (2009). ResidPlots-2: Computer software for IRT graphical residual analyses. *Applied Psychological Measurement*, *33*(5), 411-412. [software package available at available at https://www.umass.edu/remp/main_software.html] https://doi.org/10.1177/0146621608329502

Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, *32,* 179-197. https://doi.org/10.1111/j.1745-3984.1995.tb00462.x

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*(4), 453-461. https://doi.org/10.1177/014662168400800409

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149-174. doi:10.1007/bf02296272 https://doi.org/10.1007/BF02296272

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, *16*, 159-176. https://doi.org/10.1177/014662169201600206

Pearson (2012). The Puerto Rico Pruebas Puertorrique-as de Aprovechamiento Académico (PPAA) technical manual. Austin: Author,

Pearson. (2015). *Technical Manual for Minnesota's Title I and Title III Assessments for the Academic Year 2014-2015.* Roseville, MN: Minnesota Department of Education. Available from: http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika*, *56*(4), 611-630. https://doi.org/10.1007/BF02294494

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Copenhagen: Danish Institute for Educational Research.

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

Rudner, L. M. (2005). Expected classification accuracy. *Practical Assessment Research & Evaluation, 10*(13). Available from: http://pareonline.net/pdf/v10n13.pdf

Sireci, S. G., Baldwin, P., Martone, A., Zenisky, A. L., Kaira, L., Lam, W., Shea, C. L., Han, K. T., Deng, N., Delton, J., & Hambleton, R. K. (2008, April). *Massachusetts Adult Proficiency Tests technical manual: Version 2.* Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.

Wheadon, C. (2014). Classification accuracy and consistency under item response theory models using the package classify. *Journal of Statistical Software*, *56*(10), 1-14. https://doi.org/10.18637/jss.v056.i10