Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 20, Number 18, August 2015 ISSN 1531-7714

Practical Issues in Estimating Classification Accuracy and Consistency with R Package cacIRT

Quinn N. Lathrop, *Northwest Evaluation Association*

There are two main lines of research in estimating classification accuracy (CA) and classification consistency (CC) under Item Response Theory (IRT). The R package cacIRT provides computer implementations of both approaches in an accessible and unified framework. Even with available implementations, there remains decisions a researcher faces when choosing and applying the best approach for the situation. This paper identifies and discusses the practical issues that researchers may face when estimating CA and CC. To exemplify the analytic decisions, both approaches are applied to a common dataset with discussion. In addition to generalizable guidance, the demonstration provides R code for the cacIRT package.

For both reporting and inferential purposes, a primary outcome of many assessments is to classify examinees into categories. For example, examinees can be classified into and reported as Advanced, Proficient, Basic, and Below Basic based on their test scores. The classification of each examinee is based on some estimate of his or her ability. The estimate of the examinee's ability contains some amount of measurement error, and this measurement error propagates to the classification decision. Classification Accuracy (CA) and Classification Consistency (CC) are two indices that provide a simple way to communicate the quality of the classification decision. CA estimates the rate at which the classification is correct, and so has a strong relationship to the validity of the classification. CC estimates the rate at which the classification decision will be the same on two identical and independent administrations of the test, and so has a strong relationship to the reliability of the classification. Both indices have a maximum value of 1. CA and CC are widely reported in educational assessment technical reports.

Under Item Response Theory (IRT) there are two main approaches to estimate CA and CC. The first is called the Rudner approach (Rudner, 2005) and the second is called the Lee approach (Lee, 2010). Both approaches can estimate CA and CC and both have been shown to perform well in simulation studies (Lathrop & Cheng, 2013). While they do have many similarities, their differences are enough that researchers wanting to estimate CA and CC should be aware of their differences and choose the most appropriate approach for their situation. But their differences, and how those differences affect the analytic decision making of the researcher, have been less discussed in the literature (although see Lathrop & Cheng, 2013 for a mathematical comparison). This paper briefly explains both approaches and identifies and discusses practical implementation issues that might influence the decision of the researcher. To do so, both approaches are applied to a common dataset using the freely available R package cacIRT (Lathrop, 2011).

Prior to the growth of IRT, methods to estimate CA and CC were rooted in Classical Test Theory (CTT). Most involved specifying a parametric form, such as the multinomial or beta-binomial models, to represent the distribution of test scores. A summary of the CTT methods can be found in Han & Rudner (2012). For tests that do not follow IRT, some methods

may still be of importance today. In particular, the method of Livingston & Lewis (1995) appears often in technical reports.

Estimating CA and CC

To estimate CA and CC, the researcher needs the calibrated item parameters, the cut score(s), and the IRT-based ability estimates. The cut score(s) defines the score required to be classified into a certain category. There can be one or more cut scores, but this paper uses the simplest case of a single cut score that classifies examinees into Pass/Fail groups. To make a classification, an examinee's ability estimate is compared to the cut score. If the examinee's ability is equal to or greater than the cut score, he or she passes the test.

CA and CC measure the quality of the classification by quantifying the measurement error around the examinee's ability estimate. Both the Rudner approach and the Lee approach use IRT models to construct probability distributions for each examinee's ability estimate. These distributions reflect the uncertainty about the ability estimate. How the two approaches form these distributions, however, is quite different.

To demonstrate the approaches and their implementation, an empirical dataset was used of over 2,800 students responding to a 46-item test as part of an on-line undergraduate course. The data were graciously provided for use in this article by L. M. Rudner (personal communication, September 3, 2014). The test is assumed to follow the 3PL IRT model and item parameters have been previously calibrated. The cut score is given as a total score of 27. Because the cut score is given as a total score, it is also transformed to the latent ability scale. According to the Test Characteristic Curve (TCC), an examinee with ability of .0245 has an expected total score of 26.9995. In practice, cut scores can be given on the total score or the latent scale, and transforming between the two scales may introduce error as well as the issue of rounding when moving from the (continuous) latent scale to the (discrete) total score scale.

The Rudner Approach

The Rudner approach relies heavily on IRT to estimate CA and CC. Notably, the Rudner approach uses an examinee's latent ability estimate $\hat{\theta}$ and a cut score that is on the same latent scale. The Rudner approach assumes that the ability estimate and its standard error form a normal distribution (which is a very reasonable assumption and has been examined in Guo, 2006).

The top panel of Figure 1 shows this distribution for a single examinee. The examinee's $\hat{\theta}$ is above the cut score of .0245, and so they pass the test. The proportion of the area shaded in red represents the probability that the examinee is misclassified. This single examinee's CA is the proportion of area under the curve that is not red. His or her CC is the proportion of the unshaded area squared plus the proportion of the red area squared (which represents the probability of being classified in the same category on two independent tests). To calculate CA and CC for a sample (or group) of examinees, distributions are formed and CA and CC are estimated for each individual, and then the individual CA and CC estimates are averaged to arrive at the marginal CA and CC.

The Lee Approach

The major difference in the Lee approach is that the classifications occur on the total score scale. The examinee's ability estimate is his or her total score x and the cut score is also given on the total score scale. Even with this emphasis on the total score scale, the Lee approach uses the IRT model to create a distribution that reflects the uncertainty about the examinee's total score. This is done with a well-known recursive algorithm (Lord & Wingersky, 1984). The resulting distribution gives the probabilities of each total score for the examinee.

The bottom panel of Figure 1 shows this total score distribution for the same examinee as the top panel. The examinee's total score is 28 which is higher than the cut score of 27. Note that the total scores range from 0 to 46, but only the probable total scores are included to aid the comparison with the top panel. Just as with the Rudner approach, the proportion in red represents the probability of a misclassification. The individual CA and CC, as well as the marginal indices, are computed in the same manner as described above.

cacIRT R Code

The R package cacIRT provides implementations of both the Lee and Rudner approach in a unified framework and code syntax. First a note on the following notation. Anything following a > is a

command that can be typed into the R console. Anything following a # is a comment and is only provided for information. After opening R, the first step is to install and then load the cacIRT package; the installation only occurs the very first time the package is used:

```
> install.packages("cacIRT")#install
    package
> library(cacIRT) #load package
```

The response data is in a matrix named `resp.data`, in which each row represents an examinee and each column an item. The item parameters are in a matrix named `item.params`, with columns for the discrimination, difficulty, and guessing parameter respectively.

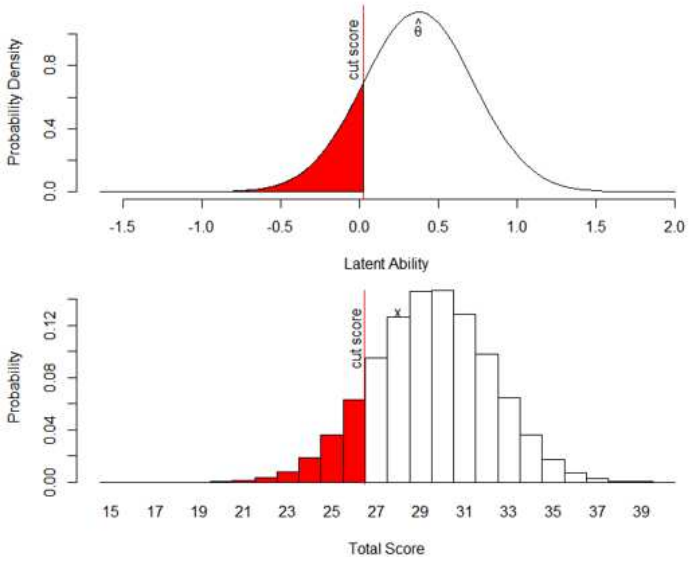

Figure 1: Conditional Probability Distributions for a Single Examinee with Ability Estimate of 0.37 and Total Score of 28. Top Panel is from The Rudner Approach. Bottom Panel is from The Lee Approach.

The two main functions in cacIRT are class.Lee and class.Rud, and both have help documentation and examples that can be accessed by typing in the R console:

```
> ?class.Lee
> ?class.Rud
```

To estimate CA and CC with the Rudner approach, the following code is used (the output is stored in an object):

```
> outR <- class.Rud(cutscore = .0245,
    ip = item.params, rdm = resp.data)
```

Because the response data matrix is given, the function will calculate the MLE ability estimates internally and their associated standard errors. The marginal CA and CC estimates are accessed by typing:

```
> outR$Marginal
                 Accuracy Consistency
cut at 0.0245 0.8791142   0.8327838
```

It is often helpful to translate the results into percentages depending on the audience. For example, a randomly selected examinee will be accurately classified 87.9% of the time.

Estimating CA and CC with the Lee approach requires only slight changes to the code

```
> outL <- class.Lee(cutscore = 27, ip =
    item.params, rdm = resp.data)
```

which results in a marginal CA estimate of 0.875 and a marginal CC estimate of 0.831. With this data and cut score, the estimates for CA and CC from the Rudner and Lee approach are almost identical. Note that the functions `class.Lee` and `class.Rud` can also accept the calculated IRT-based ability estimates and their standard errors, or a theoretical (or simulated) distribution of examinees instead of the response data matrix and example syntax can be found within the R package.

Discussion

Recall that both the Rudner approach and the Lee approach use the same IRT model and item parameters, both construct probability distributions for each examinee, and both manipulate and aggregate those distributions in the same way. The major two differences are the scale on which the classification occurs and in how the examinee uncertainty distributions are created. But in the above demonstration with a long test of 46 items, both approaches produce very similar estimates of CA and CC. Also, returning to Figure 1, both approaches have a similar understanding of a single examinee's ability or total score regardless of if the uncertainty is estimated by the Rudner or Lee approach. Because of the length of the test, the distribution of total scores under the Lee approach approaches normality thanks to the central limit theorem. For shorter tests, the total score distributions might be quite non-normal (but this is not a problem for the Lee approach). Also, if there is misfit or misspecification of the IRT model, both approaches might be affected in differential ways. So while similar

in this demonstration, meaningful differences can and will arise depending on the situation.

In general, the latent ability estimates will differ from the total scores. Thus, classification decisions based on the ability estimates will differ, to some extent, from the classification decisions based on total scores. This is why the choice of method is so important; it can change some examinees from the Pass to the Fail category and vice versa. With the demonstration dataset, the overall agreement between the two classifications is quite high at 93.3%. Even still, there are 192 examinees whose classification decision (of Pass or Fail) changes depending on the method of classification. Thus, the choice regarding whether to use the 3PL ability estimates or the total score affects the outcome of the test dramatically for these students.

It may be an uncomfortable fact for some that decisions made during an analysis can impact specific examinees in differential ways. But for simple IRT models such as the 1PL model, the total score is a sufficient statistic for the examinee's ability, and so there should not be any differences in CA and CC. But for more complex models, such as the 2PL and 3PL models, if the data fits, the latent ability estimate can provide more information and therefore make more accurate and more consistent classifications (Lathrop & Cheng, 2013). If there is evidence that the data fits an IRT model beyond the 1PL model, the latent trait should be used for the classification.

So when deciding between the Rudner approach and the Lee approach to estimate CA and CC, probably the simplest indicator is to consider the scale on which the cut score is given and the scale on which the classification occurs. If the cut score is given as a total score, the Lee approach can be used. If it is given on the latent ability scale (or some transformation of it) the Rudner approach can be used. But importantly, the scale of classification must make sense for the problem at hand. For example, in a computerized adaptive test where examinees respond to different items and possibly different numbers of items, it does not make sense to classify examinees based on their total score.

In short, by providing the above demonstration and discussion, hopefully a deeper understanding of these methods is possible. By being empowered to address the practical issues in these methods, as well as having access to the computer implementations provided by cacIRT and the above code, researchers may have a clearer path to estimate and communicate the accuracy and consistency of their classifications.

References

Han, K. T., & Rudner, L. M. (2012). Decision consistency. Paper presented at the Ronference Honoring Professor Ronald Hambleton. Amherst, MA, November, 2012. To appear in Wells, C. S. & Faulkner-Bond, M. (Eds.) (in press). *Educational measurement: From foundations to future*. New York, NY: Guilford.

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

Lathrop, Q. N. (2011). cacIRT: Classification accuracy and consistency under Item Response Theory. R package version 1.1. http://CRAN.R-project.org/package=cacIRT

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.

Lord, F. M., & Wingersky, M. S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings." *Applied Psychological Measurement*, 8(4), 453–461.

Lee, W. (2010). Classification consistency and accuracy for complex assessments using Item Response Theory. *Journal of Educational Measurement*, 47, 1-17.

Rudner, L. M. (2005) Expected classification accuracy. *Practical Assessment Research & Evaluation*, 10(13), 1–4.

Citation:

Lathrop, Quinn (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

Author

Quinn N. Lathrop
Northwest Evaluation Association
121 NW Everett St.
Portland, OR 97209

email: Quinn.lathrop [at] gmail.com