Minimizing classification errors when the true cut score is not known, with software for standard setting

Jesse R. Pace and Irina Grabovsky

Presentation Given at NCME on 04/06/2019

Abstract

We develop a method for finding an optimal cut-score for Pass/Fail examinations which incorporates uncertainty about the ‘true’ point separating proficient examinees from non-proficient ones. We derive false positive and false negative probabilities, introduce several classification metrics, and present software we have developed which performs these calculations for the user.

There are at least two different incarnations of classification accuracy metrics that one can find the research literature: actual and predicted. Those methods which we term 'actual' are conducted post-hoc, when data collection has been completed. These methods include the familiar Sensitivity and Specificity (Yerushalmy, 1947), Youden's J (Youden, 1950), and Kappa (Cohen, 1960), among others. The method we describe in the present article falls into the second category: predictive. We seek to explicate a method that can be used by researchers and standards setting panels to estimate what the rates of misclassification might be, thus providing crucial information to help judges make decisions.

We are not the first to propose a method of estimating classification accuracy. Livingston (1993) described a method for estimating accuracy that was found to be very close to actual values. Lee (2010) explicated a method that could be used for such estimation with complicated item-response theory based assessments. Other methods are described in the literature, and we refer the reader to Lee (2010), who provide a more complete list. A common feature of such estimation methods is that they typically require, as an input, the user's belief about where the cut-score should be. Rudner (2001) described a model for estimating accuracy, which also expected the operational cut-score to be known, but also went on to demonstrate how accuracy estimates change if the cut point is altered. The idea that a cut score might need to be adjusted has been the focus of the current authors preceding work on this topic.

Grabovsky and Wainer (2017a & 2017b) described a method for estimating the optimal cut score. This method included estimation of accuracy at the inputted value of the cut score, but went on to explicate a method for determining the classification accuracy of different potential cut points. The inputted cut-score value was considered to be the 'true' score, which could be informed by a standards setting panel (e.g., via Angoff method [Angoff, 1971]). As the authors showed, the point of the true cut-score is not necessarily the point of optimal classification. While we, and others, have treated the true cut-score as a known value for our estimates, this is a strong assumption.

That panelists vary in their estimates of the true cut-score is known to anyone who has participated in standard setting, and can be seen in any article which reports the results of standard setting (e.g., Buckendahl, Smith, Impara, and Plake, 2002). This variability, however, has not been included in previous methods attempting to estimate classification accuracy. In the present article, we explicate a method to include this variability in our estimates of accuracy, and our search for the optimal cut-score.

Methods

We begin by defining several measures of classification error. By searching over the values of these various measures we can identify the optimal cut-score.

The first measure we consider is defined as the larger of the probabilities of false-positive and false-negative classification for each possible cut-score point *C* on the observed score scale, or maximal classification error

When a standard setting is held, the Angoff method (Angoff, 1984) allows for a given judge to provide a possible value for the true cut score. If one were to select a content expert at random from the population of all relevant content experts, the value of the cut-score one would obtain would be a random variable. The distribution of this random variable is assumed herein to be normal. We estimate the mean and variance of this variable, which we subscript with A for

Angoff, using the standard unbiased estimates commonly called the sample mean and standard deviation,

$$\text{Mean Angoff} = \frac{\sum_1^n \tau^*{}_i}{n} = \mu_A$$

$$\text{SD Angoff} = \sqrt{\frac{\sum_1^n (\tau^*{}_i - \mu_A)^2}{n-1}} = \sigma_A$$

We proceed under the 1-PL/Rasch Item Response Theory model.

The probability that an examinee with ability τ answers a question of difficulty b correctly is:

$$p(\tau, b) = \frac{1}{1 + \mathrm{e}^{b-\tau}}$$

The proportion correct score, T_N,for an examinee on a test with N items is:

$$T_N = \frac{1}{N} \sum_1^N \xi_i(\tau)$$

Where $\xi_i(\tau)$ are independent Bernoulli variables (where a 1 is a correct response, and 0 is incorrect), with probability of success, $p(\tau, b)$. T_N, then, is a sum of independent Bernoulli random variables divided by N, hence its mean and variance are

$$\text{Mean} = \overline{p(\tau)} = \frac{1}{N} \sum_1^N p(\tau, b_i)$$

$$\text{Variance} = \sigma(\tau) = \frac{1}{N} \sqrt{\sum_1^N p(\tau, b_i)\,(1 - p(\tau, b_i))}$$

When the exam has many items, the probability that an examinee will fail the test, with a cut score of c, is estimatible using central limit theorem

$$p_{fail}(\tau, \mathrm{c}) = p(T_N < \mathrm{c}) \approx \Phi(\frac{\mathrm{c} - \overline{p(\tau)}}{\sigma(\tau)})$$

Where Φ is the cumulative distribution function of a standard normal variable.

The probability that an examinee has true ability above the true cut-score, τ^*, is

$$p(\tau > \tau^*) = \Phi(\frac{\tau - \mu_A}{\sigma_A})$$

The probability that a given examinee fails, while their ability is above the true cut-score, is defined to be the false negative probability, which, by independence, is

$$p_{\mathrm{fn}}(\tau, \mathrm{c}) = p(T_N < \mathrm{c} \text{ and } \tau > \tau^*) = \Phi(\frac{\mathrm{c} - \overline{p(\tau)}}{\sigma(\tau)}) \Phi(\frac{\tau - \mu_A}{\sigma_A})$$

The false negative probability is derived similarly, resulting in

$$p_{\mathrm{fp}}(\tau, \mathrm{c}) = p(T_N > \mathrm{c} \text{ and } \tau < \tau^*) = \left[1 - \Phi\left(\frac{\mathrm{c} - \overline{p(\tau)}}{\sigma(\tau)}\right)\right]\left[1 - \Phi\left(\frac{\tau - \mu_A}{\sigma_A}\right)\right]$$

Using test reliability, ρ, and the standard deviation of the examinee sample, σ_{Ex}, we can estimate true score variance, σ_0 (e.g., Harvill, 1991)

$$\sigma_0{}^2 = \sigma_{ex}{}^2 - (1 - \rho)\sigma_{ex}{}^2 = \sigma_{ex}{}^2\rho$$

If we select an examinee at random from the examinee population, then their ability, τ_x, is normally distributed, N($\tau_0, \sigma_0{}^2$) where τ_0 is the mean of true ability distribution. We can use the mean from the empirical theta distribution to estimate τ_0. The probability of making a false negative classification across a specified small interval of true score values is approximated by:

$$p(T_N < \mathrm{c} \text{ and } \tau > \tau^* \; and \; \tau_i < \tau_x < \tau_{i+1}) = p_{\mathrm{fn}}(\tau_i, \mathrm{c}) p(\tau_i < \tau_x < \tau_{i+1})$$

If we take the sum of all such intervals, where each interval is infinitely small, we have the integral which expresses the probability of making a false negative error on an exam:

$$p(\mathrm{FN}) = \int_{-\infty}^{\infty} p_{\mathrm{fn}}(\tau, \mathrm{c}) \frac{1}{\sigma_0} \phi\left(\frac{\tau - \tau_0}{\sigma_0}\right) d\tau$$

And the false positive error

$$p(\mathrm{FP}) = \int_{-\infty}^{\infty} p_{\mathrm{fp}}(\tau, \mathrm{c}) \frac{1}{\sigma_0} \phi\left(\frac{\tau - \tau_0}{\sigma_0}\right) d\tau$$

Where ϕ is the PDF of a standard normal variable.

The total probability of making an error is simply the sum of the two errors, i.e., FP+FN, and if we minimize this function we find the cut-score value of optimal classification (where total error is lowest).

We also derive penalty based measures which seek to penalize extreme misclassifications more harshly. That is, the resulting optimal point will bend further away from those points where extreme classifications are taking place. We use the function

$$w(x) = e^{|x|-a} - 1$$

We choose a to be σ_A. This leads to the function treating differences larger than σ_A as meaningful. Deriving the resulting equations follows the same proceedures outlined above, but with the inclusion of the penalty function. We arrive at:

$$Pen_{fn}(\tau, c) = \Phi\left(\frac{\mathrm{c} - \overline{p(\tau)}}{\sigma(\tau)}\right) \frac{1}{\sigma_A} \int_{-\infty}^{\tau} w(x - \tau)\, \phi\left(\frac{x - \tau_A}{\sigma_A}\right) dx$$

$$Pen_{fp}(\tau, c) = \Phi\left(\frac{\overline{p(\tau)} - \mathrm{c}}{\sigma(\tau)}\right)\frac{1}{\sigma_A}\int_{\tau}^{\infty} w(x - \tau)\,\phi\left(\frac{x - \tau_A}{\sigma_A}\right)dx$$

Finally, we have developed software which allows users to use either penalty or regular probability based measures to locate optimal cut scores.

Software

To aid standard setting members in using the method described here, we have developed easy-to-use software. The software is distributed as a .exe file which can be installed and run like any typical application. The software was developed using the R programming language (R Core Team, 2017), along with the packages Shiny (Chang, Cheng, Allaire, Xie, and McPherson, 2018) and RINNO (Hill et al., 2018). The user need only supply the proper inputs, which are (See figure 1):

1) A vector of item difficulties (1-PL)
2) (optional) Entry of a lookup table from %correct to another metric, in case the user would prefer output in another metric
3) The mean of the Angoff ratings
4) SD of Angoff ratings
5) Mean of Examinees abilities (1-PL)
6) SD of Examinees abilities(1-PL)
7) Test reliability

This software identifies the optimal cut-score location, and also provides graphical output indicating the error at each point of the potential cut-score (see Figure 2). The method provides output via 5 metrics: one is the total error, described above. The software also supplies the point of minimal classification error for all methods we introduced.

Discussion

In this paper we have presented a method to incorporate uncertainty in standard setting into predictions of classification error, introduced penalty based errors, and explained the use of software which allows for ready use of our model.

The penalty function provides potential utility to standard setting committees in that it penalizes extreme classification mistakes more heavily than small mistakes. This could have utility in certain situations. For example, imagine medical licensure testing, which serves to protect the public from non-competent individuals practicing medicine. Competence likely exists on a continuum. It seems to follow directly that the public is put at greater harm when a particularly low competence examinee is accidentally allowed to pass a certification test, relative to when an almost minimally competent examinee is allowed to pass. Thus, penalizing such extreme mistakes more heavily seems to be a safer decision than treating all errors equally, since it serves to better protect the public. In other scenarios, however, this may be unnecessary, or counterproductive. The decision will have to be weighed by the standard setting committee.

Decisions such as whether extreme misclassifications should be penalized greater, or whether or not false positives vs. false negatives should be weighted more heavily, or whether or not the true cut-score should be treated as a random variable, can all lead to different optimal cut scores. It behooves the standard setting committee to consider these questions when attempting to settle on a final cut score. The method explicated in this paper (and delivered via software) provides a mathematical method to incorporate these features in the final decision. Thus standard setting committees are provided an analytical way to consider these decisions.

Limitations and future directions

The purpose of this paper has been to explicate the mathematical model for an unknown true cut score, to introduce the penalty function, and to offer software for use of these methods in standard setting.

The mathematical model accounts for variability in the true cut score by treating it as a random variable with an estimatable variance. While we can say that this variation has now been accounted for in the model, the magnitude and direction of the impact of accounting for this variability remains unstudied. Future research which compares optimal cut scores with and without treatment of the true cut score as a random variable is warranted.

References

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

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.

Chang,w., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018) shiny: Web Application Framework for R. R package version 1.1.0. Retrieved from: CRAN.R-project.org/package=shiny

Cizek, G. J., & Bunch, M. B. (2007). The Bookmark Method. In *Standard setting: A guide to establishing and evaluating performance standards on tests*., 155-192, SAGE Publications Ltd.

Cizek, G. J. (2012). An Introduction to Contemporary Standard Setting: Concepts, Characteristics, and Concepts. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed., pp 3-13). New York, NY: Routledge.

Harvill, L. M. (1991). Standard Error of Measurement: an NCME Instructional Module on. *Educational Measurement: issues and practice*, 10, 33-41.

Hill, J. & Pang, W.L. (2018). RInno: An Installation Framework for Shiny Apps. R package version 0.2.1. Retrieved from: CRAN.R-project.org/package=RInno

Hurtz, G. M., & Auerbach, M. A. (2003). A meta-analysis of the effects of modifications to the Angoff method on cutoff scores and judgment consensus. *Educational and Psychological Measurement*, *63*(4), 584-601.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Grabovsky, I., & Wainer, H. (2017a). The cut-score operating function: A new tool to aid in standard setting. *Journal of Educational and Behavioral Statistics*, 42, 251-263.

Livingston, S. A., & Zieky, M. J. (1982). Passing scores: A manual for setting standards of performance on educational and occupational tests.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education*, *2*(2), 121-141.

Lee, W. C. (2010). Classification consistency and accuracy for complex assessments using item response theory. *Journal of Educational Measurement*, 47, 1-17.

Lewis, D. M., Mitzel, H. C., Mercado, R. L., & Schulz, E. M. (2012). The Bookmark Standard Setting Procedure. In G. J. Cizek (Ed.), Setting performance standards: Foundations, methods, and innovations (2nd ed., pp. 225 - 253). New York, NY: Routledge.

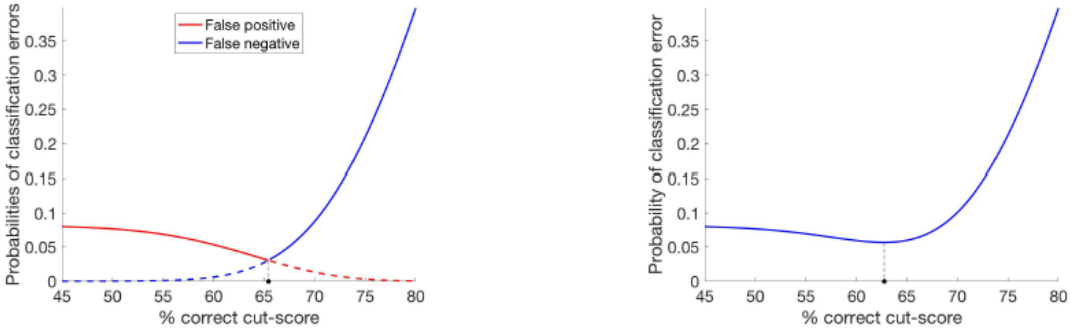
R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org/

Rudner, L. M. (2001). Computing the expected proportions of misclassified examinees. *Practical Assessment, Research & Evaluation*, 7, 1-8.

Yerushalmy, J. (1947). Statistical problems in assessing methods of medical diagnosis, with special reference to X-ray techniques. *Public Health Reports (1896-1970)*, 1432-1449.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*, 32-35.

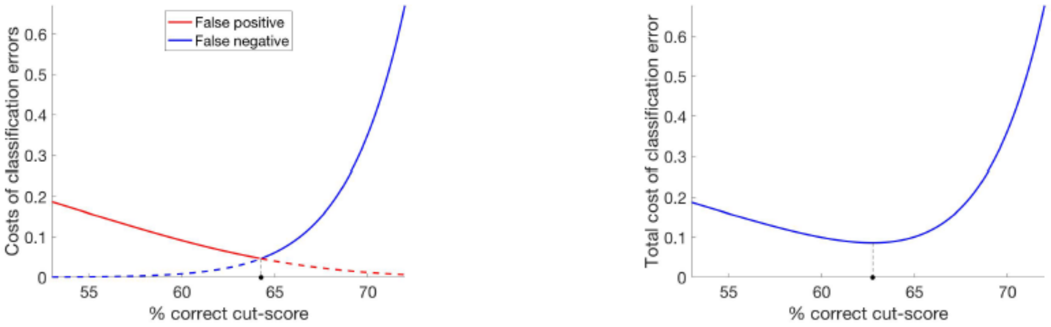
Figure 1



(a) Probabilities $FP(c)$ of false positive and $FN(c)$ of false negative misclassifications as a function of the cut-score c.

(b) Probability $TCE(c)$ of misclassification as a function of the cut-score c.

Figure 2.



(a) Costs $FP(c)$ of false positive and $FN(c)$ of false negative misclassifications as a function of the cut-score c.

(b) Total cost $TPCE(c)$ of misclassification as a function of the cut-score c.

Figure 3. Conditional probability of misclassification.

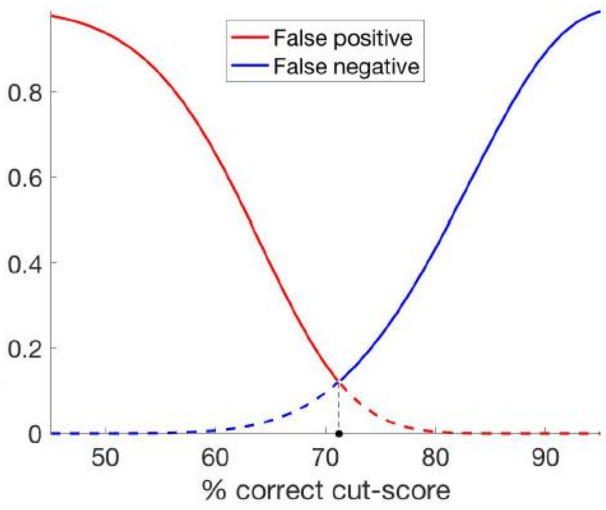
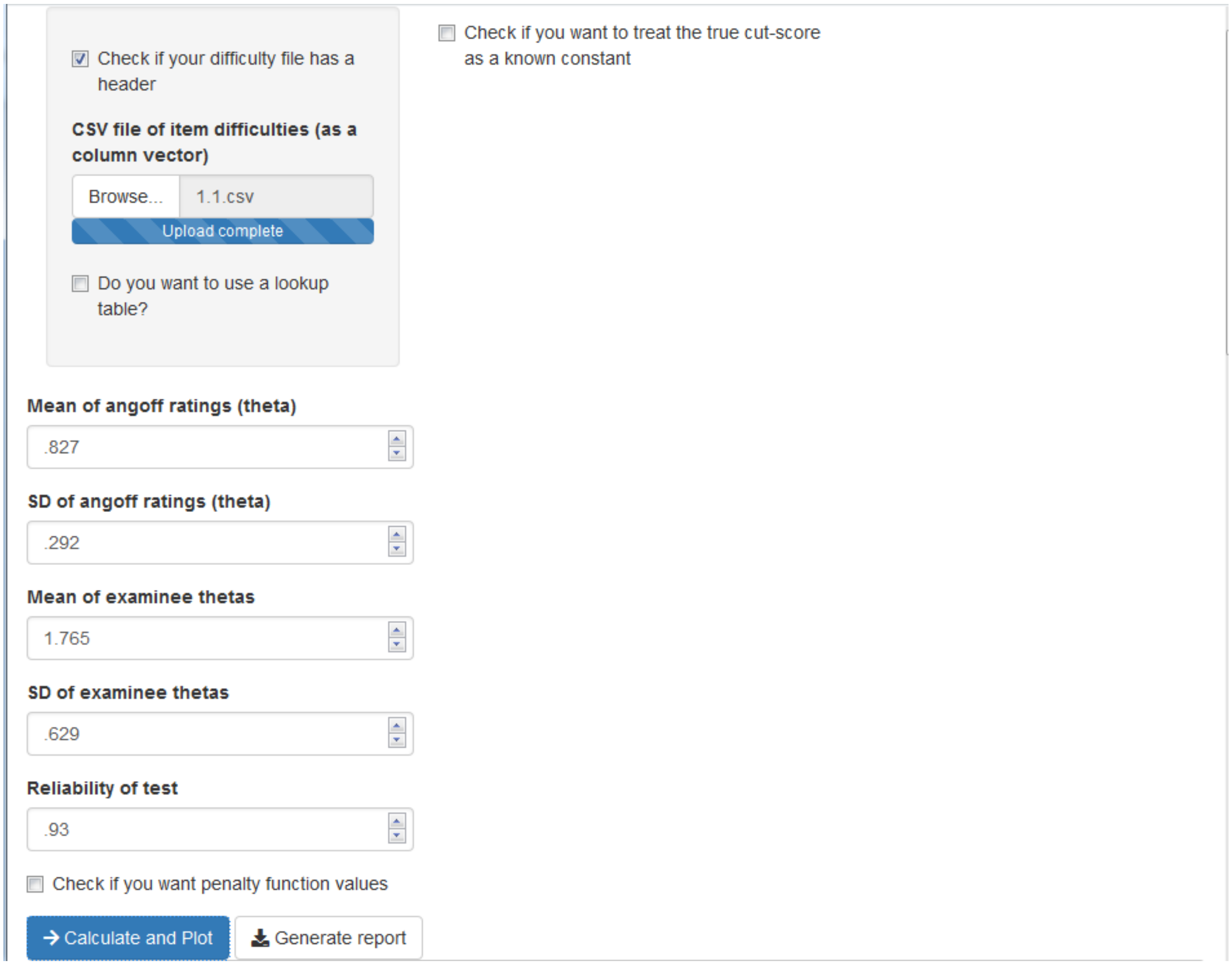


Figure 4.



☑ Check if your difficulty file has a header

CSV file of item difficulties (as a column vector)

1.1.csv

□ Do you want to use a lookup table?

Mean of angoff ratings (theta)

.827

SD of angoff ratings (theta)

.292

Mean of examinee thetas

1.765

SD of examinee thetas

.629

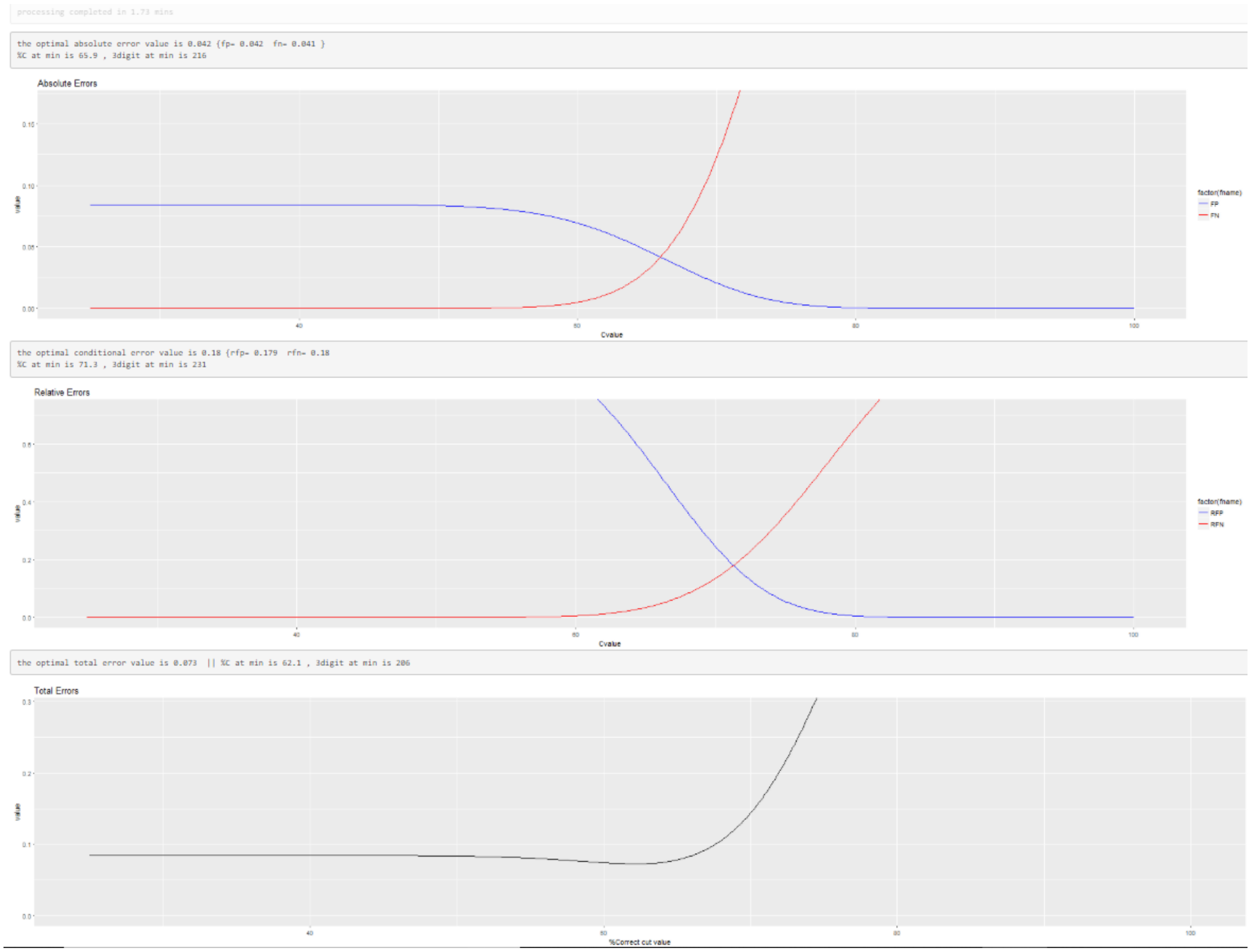
Reliability of test

.93

□ Check if you want penalty function values

□ Check if you want to treat the true cut-score as a known constant

Figure 5



```
processing completed in 1.73 mins
```

```
the optimal absolute error value is 0.042 {fp= 0.042  fn= 0.041 }
%C at min is 65.9 , 3digit at min is 216
```

Absolute Errors

```
the optimal conditional error value is 0.18 {rfp= 0.179  rfn= 0.18
%C at min is 71.3 , 3digit at min is 231
```

Relative Errors

```
the optimal total error value is 0.073  || %C at min is 62.1 , 3digit at min is 206
```

Total Errors