Minimizing Classification Errors with Unknown True Cut Scores With Software for Standard-Setting

Jesse R. Pace
Irina Grabovsky

Predicting Classification Error

- Classification errors are never good
  - Knowing what kind of errors might be made is valuable information to standard setting committees
- Estimating error BEFORE test administration is possible
  - E.g., Rudner (2001)
  - Grabovsky & Wainer (2017)
- Knowing estimated error rates at various potential cut scores might be valuable information to standard setting committees

Errors come in two basic forms

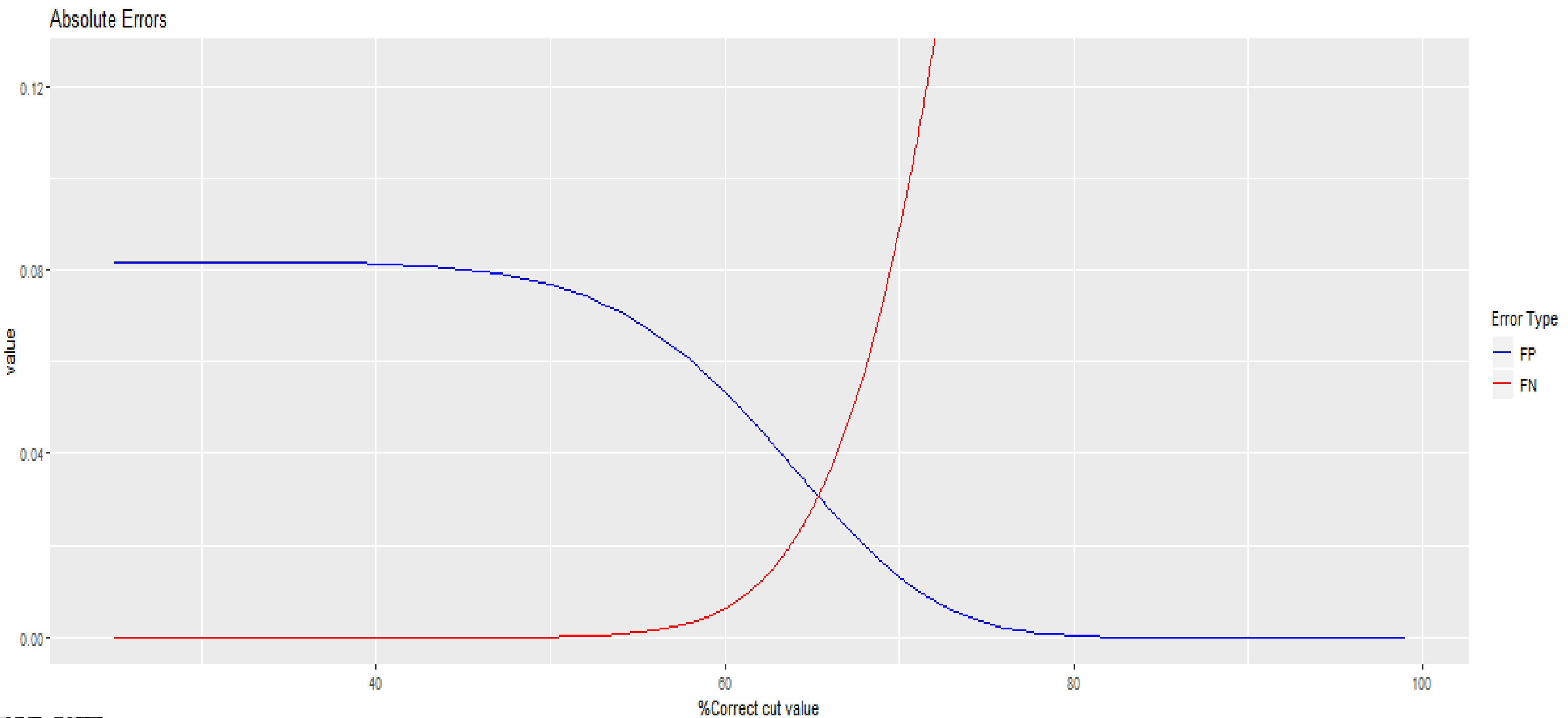
- False Positives FP) and False Negatives (FN)
- FP are examinees who should have failed but are given a passing score
- FN are examinees who should have passed, but are given a failing score

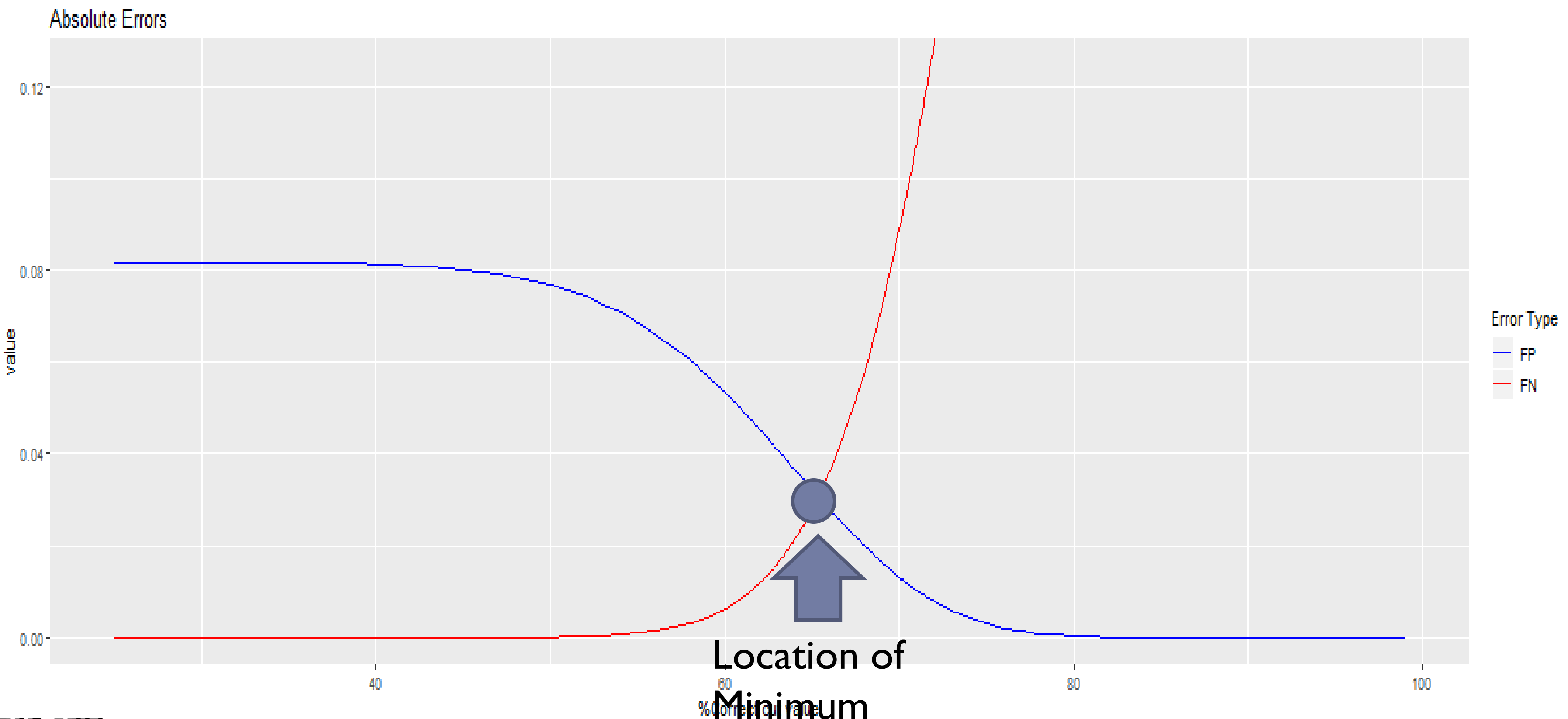
Optimal Cut Score Location

- Using a combination of FN and FP, it is possible to find the point that minimizes that combination.
- We focus on two such combinations
  - Absolute Error and Total Error

Absolute error minimum

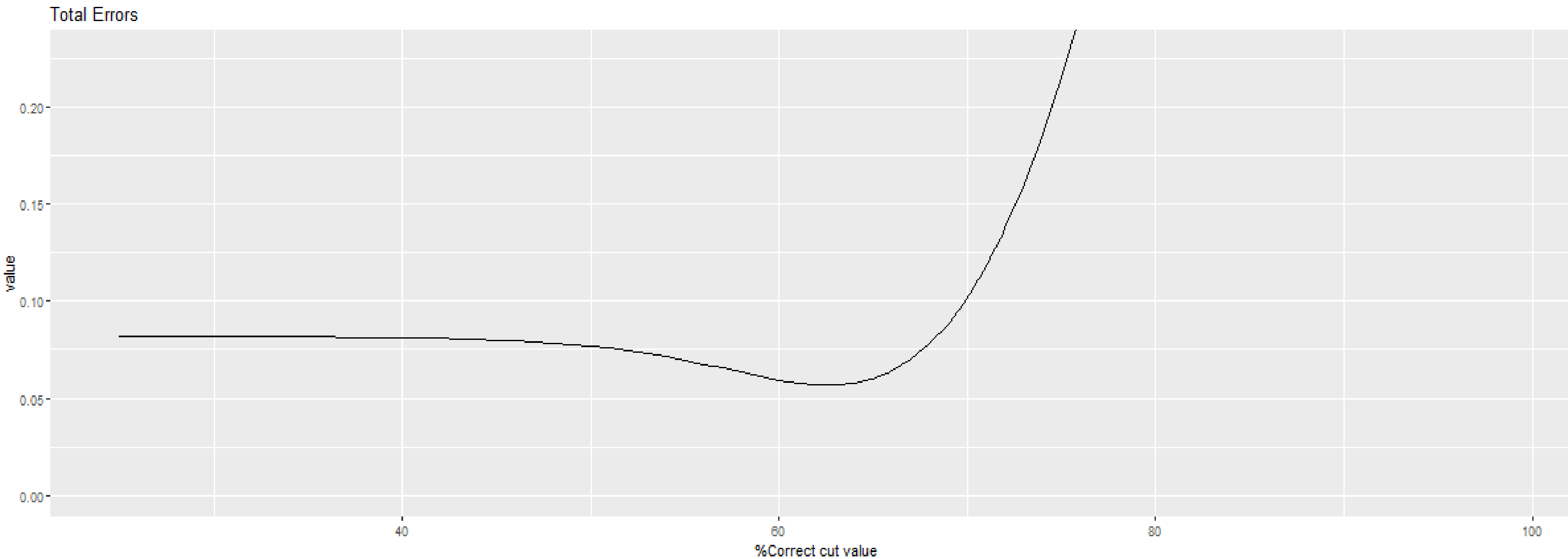
- The intersection of FP and FN is the point of min{max(FP,FN)}




Absolute Errors
0.12
0.08
0.04
0.00
value
40
60
80
100
%Correct cut value
Error Type
FP
FN
Location of Minimum

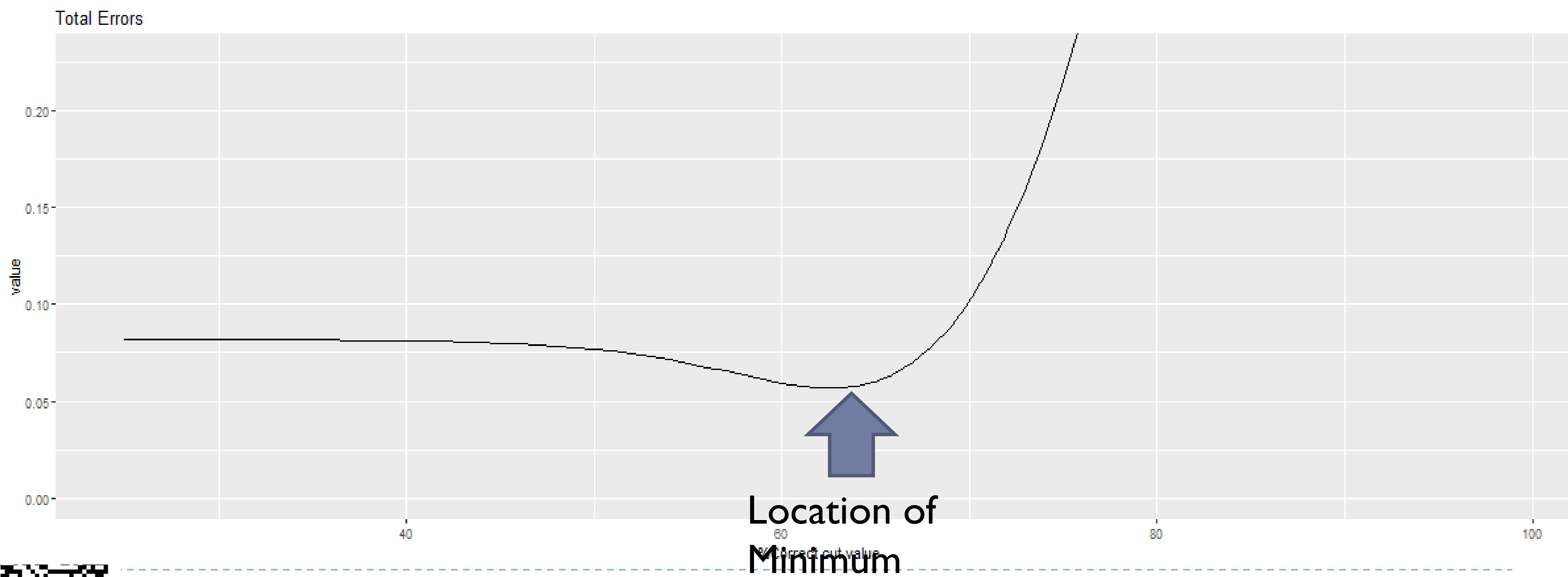
Total error minimum

- The sum of FP and FP



Total error minimum

- The sum of FP and FP
- Note, may be different than the absolute method.
  - I.e.,The minimum of the sum of the errors may be different than the minimum of the maximum of both.



Penalty Based Error

- If we so choose, we could penalize extreme errors more harshly
  - That is, situations where an examinee's true ability is far from the cut score are penalized greater than those whose true abilities are closer
  - Imagine this in medical testing, for instance.
    - A licensure test serves to protect the public from non-competent individuals practicing medicine.
    - Competence likely exists on a continuum.
    - It follows that the public is put at greater harm when a particularly low competence examinee is allowed to pass relative to when an almost minimally competent examinee is allowed to pass.
    - Thus, penalizing such extreme errors more heavily seems to be a safer decision

Penalty Function

- The penalty error function method involves adds a weight to the formula, and then finding the minimum value of the resulting function of cut score (c)
  - The penalty function chosen for this procedure was:
  - $e^{|\tau^* - \tau|/\sigma_A} \;\; - 1$

- Within the penalty function, we can calculate absolute and total error, just like in the marginal probability case

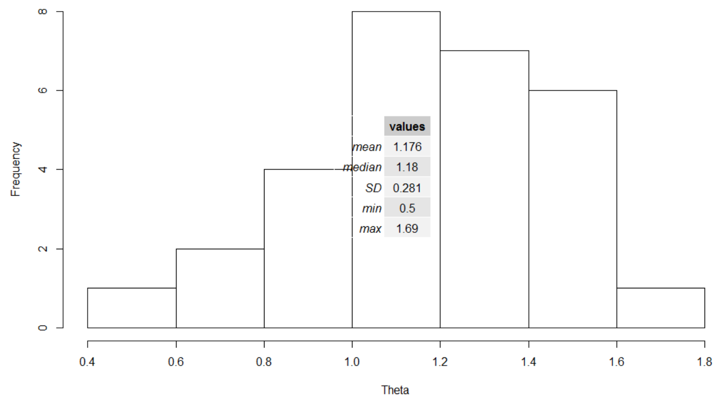
Estimation

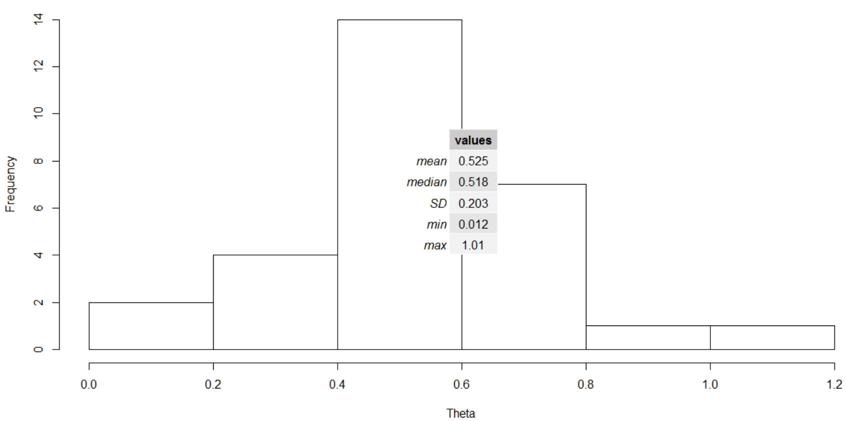
- We can estimate the FP and FN, and the location of minimum error, using a mathematical model
  - Such a model was published by Grabovsky and Wainer (2017)
- We have since worked to incorporate uncertainty about standard setting results

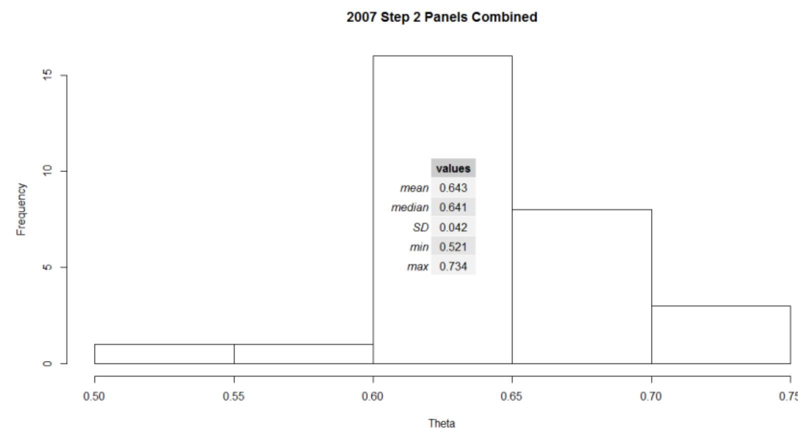
Standard Setting Variance

- Judges rarely all agree on given cut score
- Different judge panels are likely to produce different mean cut scores
- We have worked this uncertainty into our mathematical model

Empirical Angoff Panel Distributions







- Looking at multiple years of standard setting data, we believed a normal distribution was a reasonable approximation

- We assume that the distribution of the cut score from standard setting to be normal
  - We call this τ^* hereafter
  - We use unbiased estimators for a normal random variable

Mean = $\frac{\sum_1^n \tau_i{}^*}{n} = \mu_A$ (where the A subscript denotes that this comes from the Angoff ratings)

SD = $\sqrt{\frac{\sum_1^n (\tau_i{}^* - \mu_A)^2}{n-1}} = \sigma_A$

  - Thus, we say that $\tau^* \sim N(\mu_A, \sigma_A{}^2)$

- The random variable, τ^* enters in the calculation of false positive and false negative errors

E.g.,

p(false negative) = p(observed score < cut score $\cap$ true ability $>\tau^*$)

Using central limit theorem, and deriving some equations (see handout) we get the following form via independence

$$p(FN)= p(z<\frac{c - E[\text{observed}]}{\sigma_{observed}}) * p(z<\frac{\text{true ability} - \mu_A}{\sigma_A})$$

And p(FP)=p(observed score > cut score $\cap$ true ability $<\tau^*$)

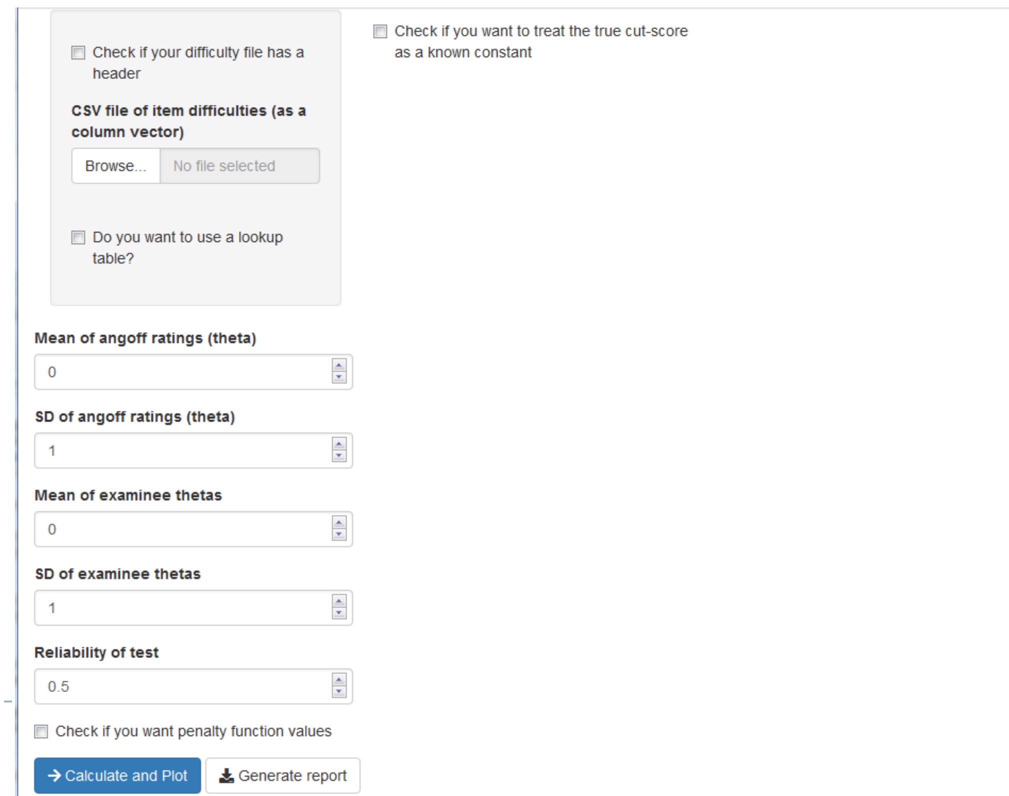
$$= [1- p(z<\frac{c - E[\text{observed}]}{\sigma_{observed}})]*[1-p(z<\frac{\text{true ability} - \mu_A}{\sigma_A})]$$

Intended Use and Software

- The ultimate goal of this work is to provide standard setting committees with additional information in order to aid their process of setting cut-scores.
- To this end, software which implements the mathematical model for the user has been developed

Software Interface

- Windows software has been developed for standard setting committees



If we believe cut scores are constant (not random variables)

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

Check if your difficulty file has a header

CSV file of item difficulties (as a column vector)

Browse... No file selected

Do you want to use a lookup table?

Mean of angoff ratings (theta)

0

SD of angoff ratings (theta)

1

Mean of examinee thetas

0

SD of examinee thetas

1

Reliability of test

0.5

Check if you want penalty function values

Calculate and Plot | Generate report

Selecting penalty output will yield penalty based optimal cutscores

☐ Check if your difficulty file has a header

CSV file of item difficulties (as a column vector)

Browse... No file selected

☐ Do you want to use a lookup table?

Mean of examinee thetas

0

SD of examinee thetas

1

Reliability of test

0.5

☐ Check if you want penalty function values

→ Calculate and Plot | Generate report

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

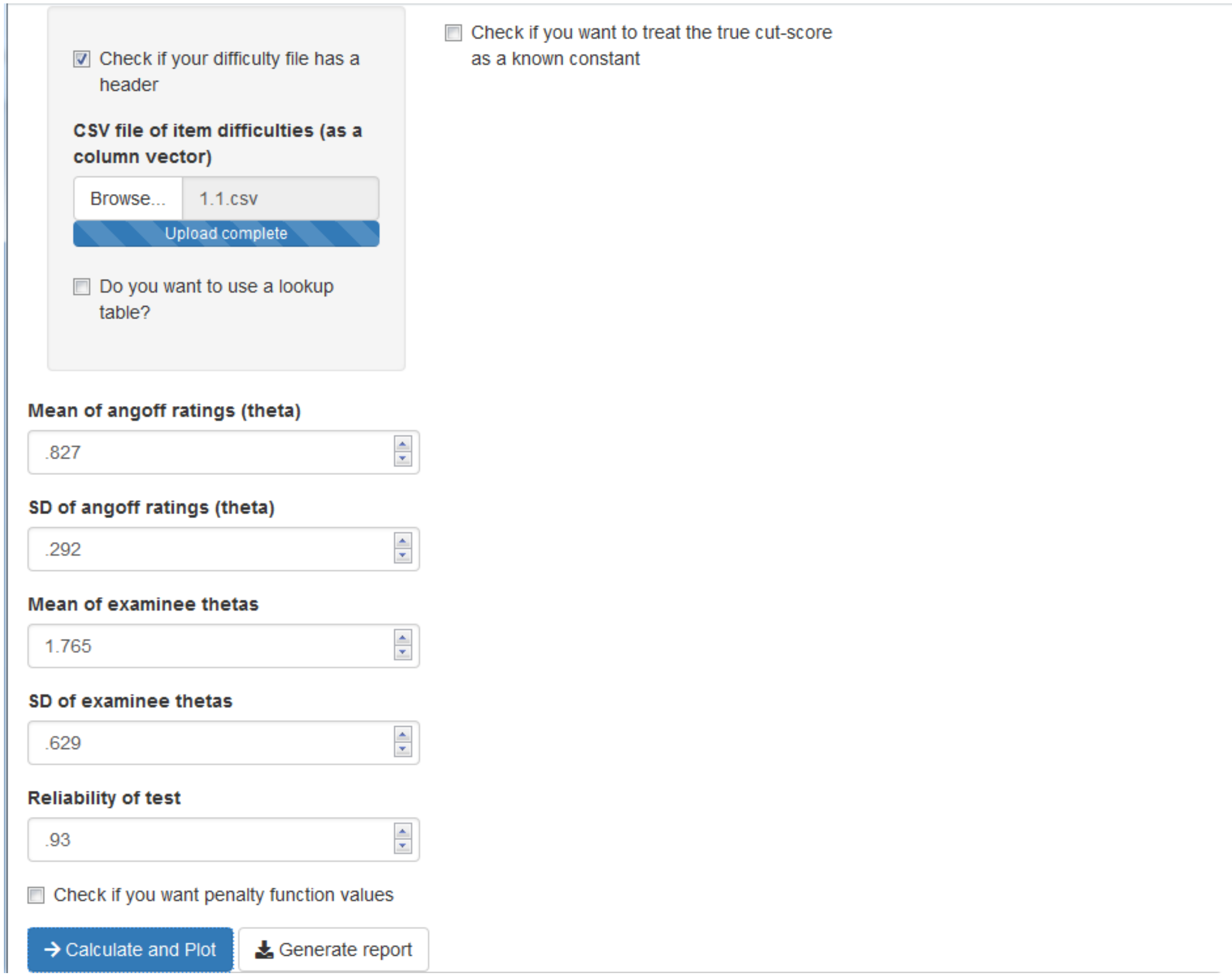
True cut-score value (theta scale)

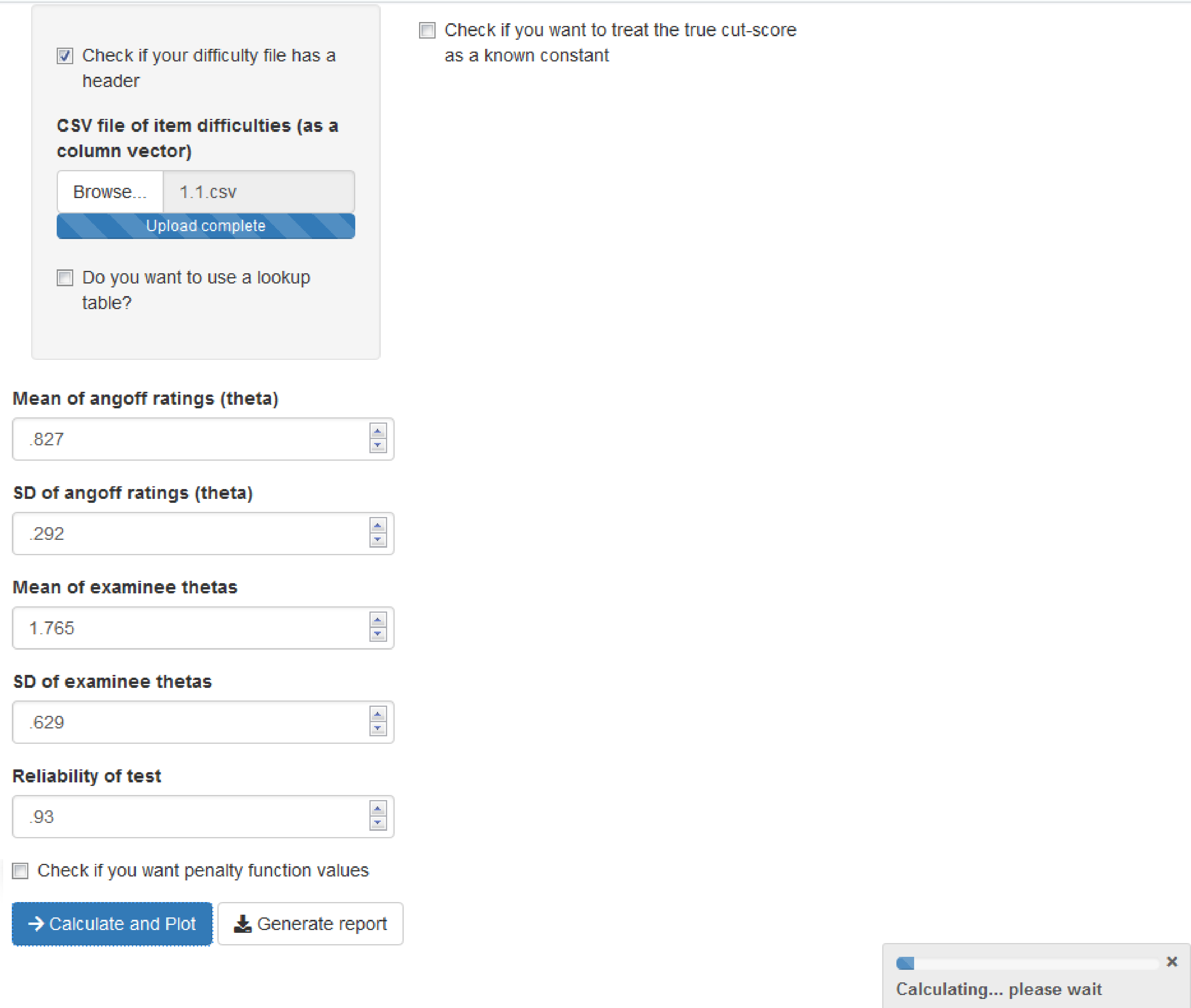
0

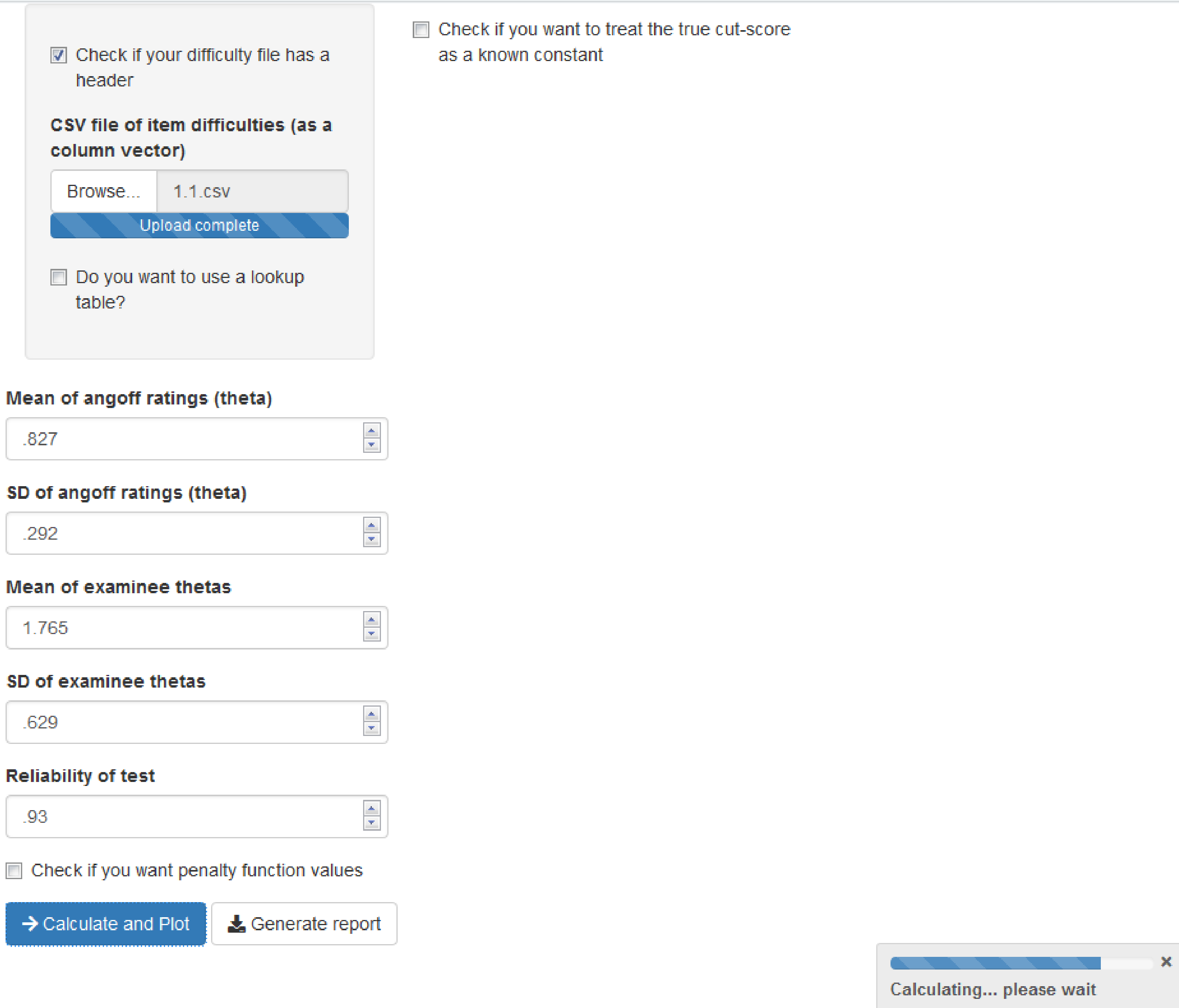
- Indicating that cut scores are known constants reduces the input variables
- No longer a need for variance of the cut scores

Software Interface

- When Supplied with Inputs…

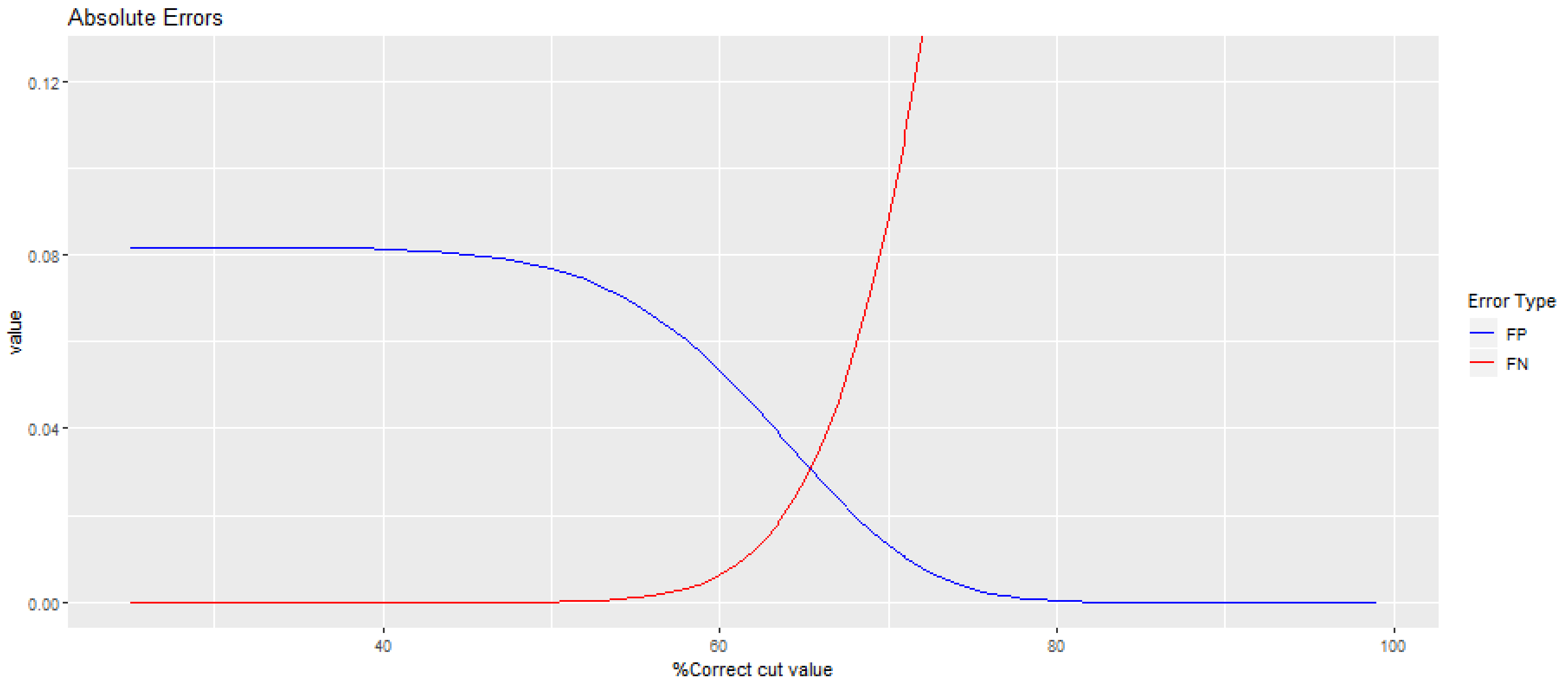



Check if your difficulty file has a header
CSV file of item difficulties (as a column vector)
Browse...
1.1.csv
Upload complete
Do you want to use a lookup table?
Check if you want to treat the true cut-score as a known constant
Mean of angoff ratings (theta)
.827
SD of angoff ratings (theta)
.292
Mean of examinee thetas
1.765
SD of examinee thetas
.629
Reliability of test
.93
Check if you want penalty function values
Calculate and Plot
Generate report
Calculating... please wait

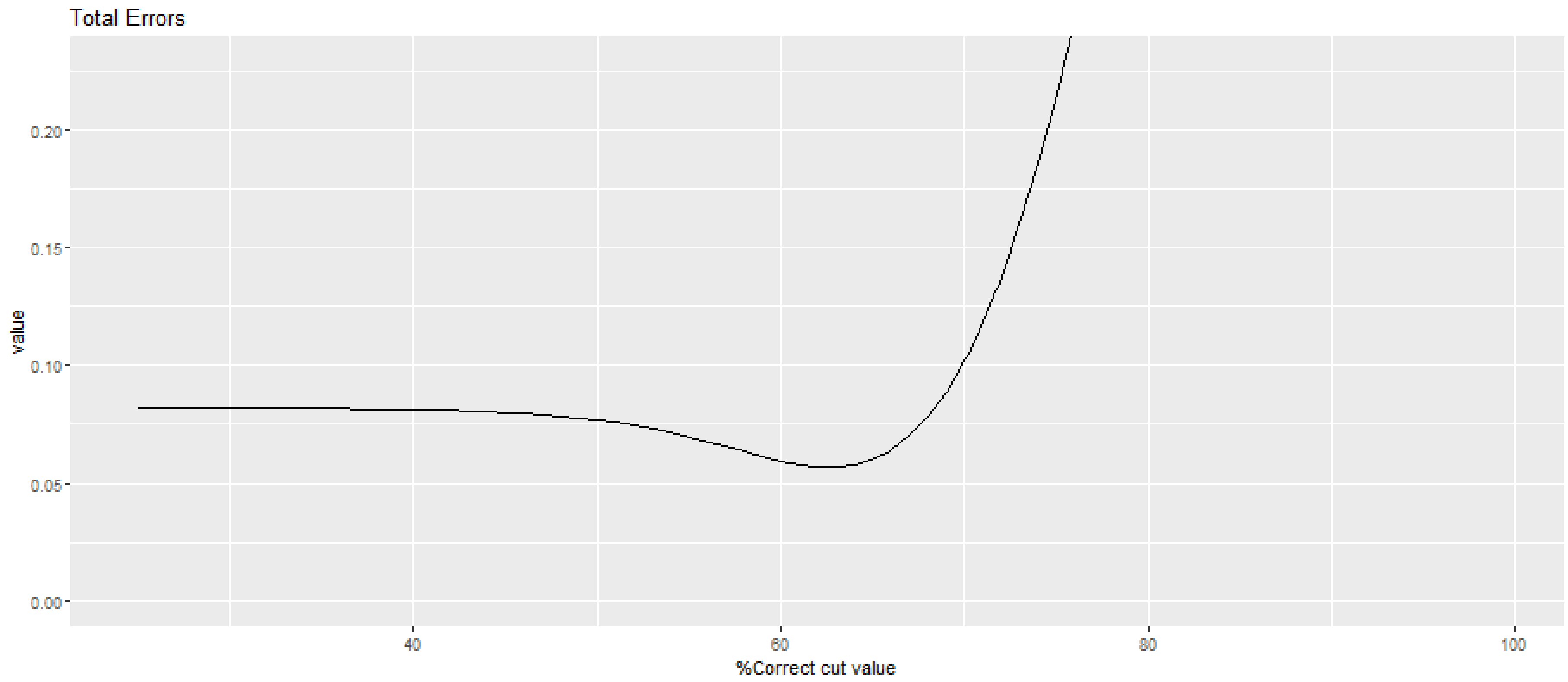

Check if your difficulty file has a header
CSV file of item difficulties (as a column vector)
Browse...
1.1.csv
Upload complete
Do you want to use a lookup table?
Check if you want to treat the true cut-score as a known constant
Mean of angoff ratings (theta)
.827
SD of angoff ratings (theta)
.292
Mean of examinee thetas
1.765
SD of examinee thetas
.629
Reliability of test
.93
Check if you want penalty function values
Calculate and Plot
Generate report
Calculating... please wait

Probability Output

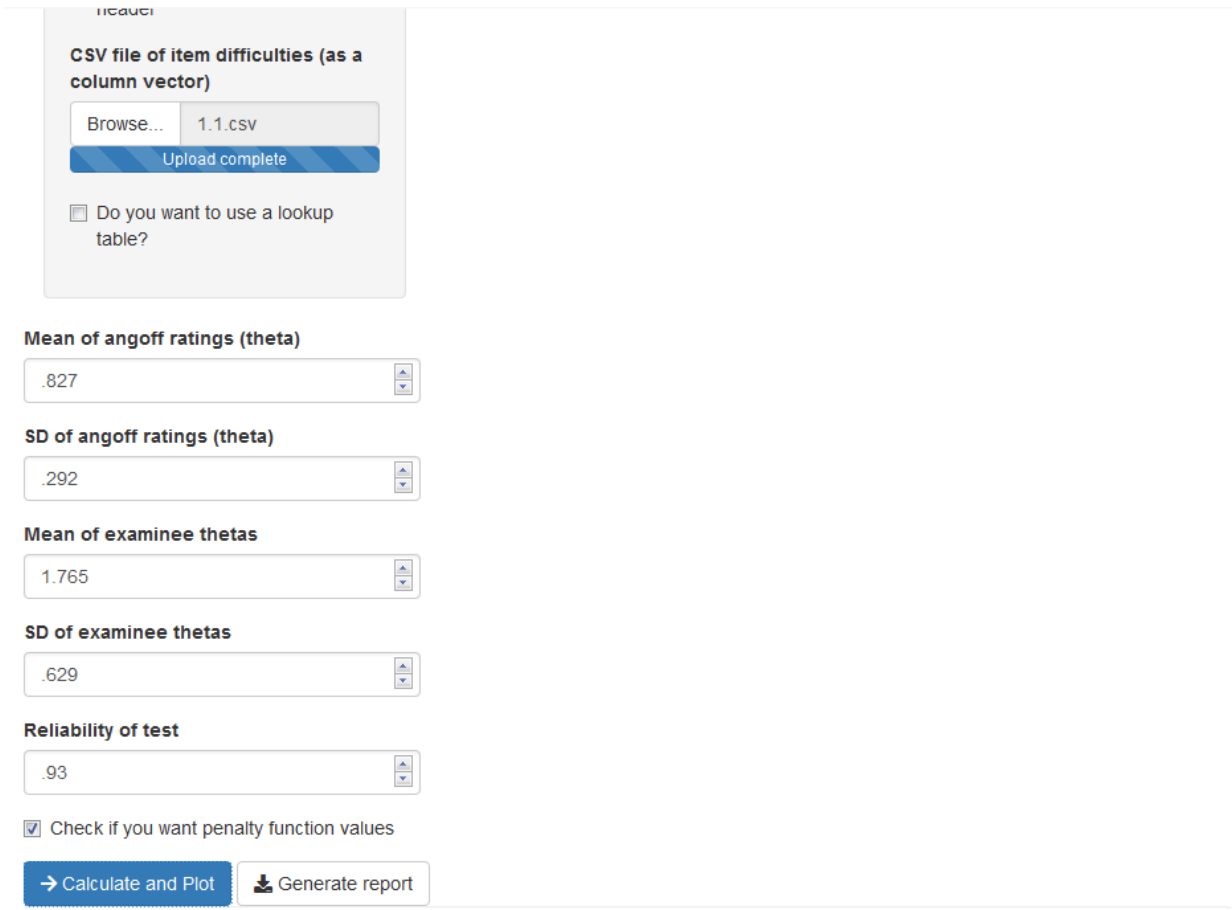
```
the optimal absolute error value is 0.031 {fp = fn = 0.031 }
%C at min is 65.3 , no conversion indicated
```



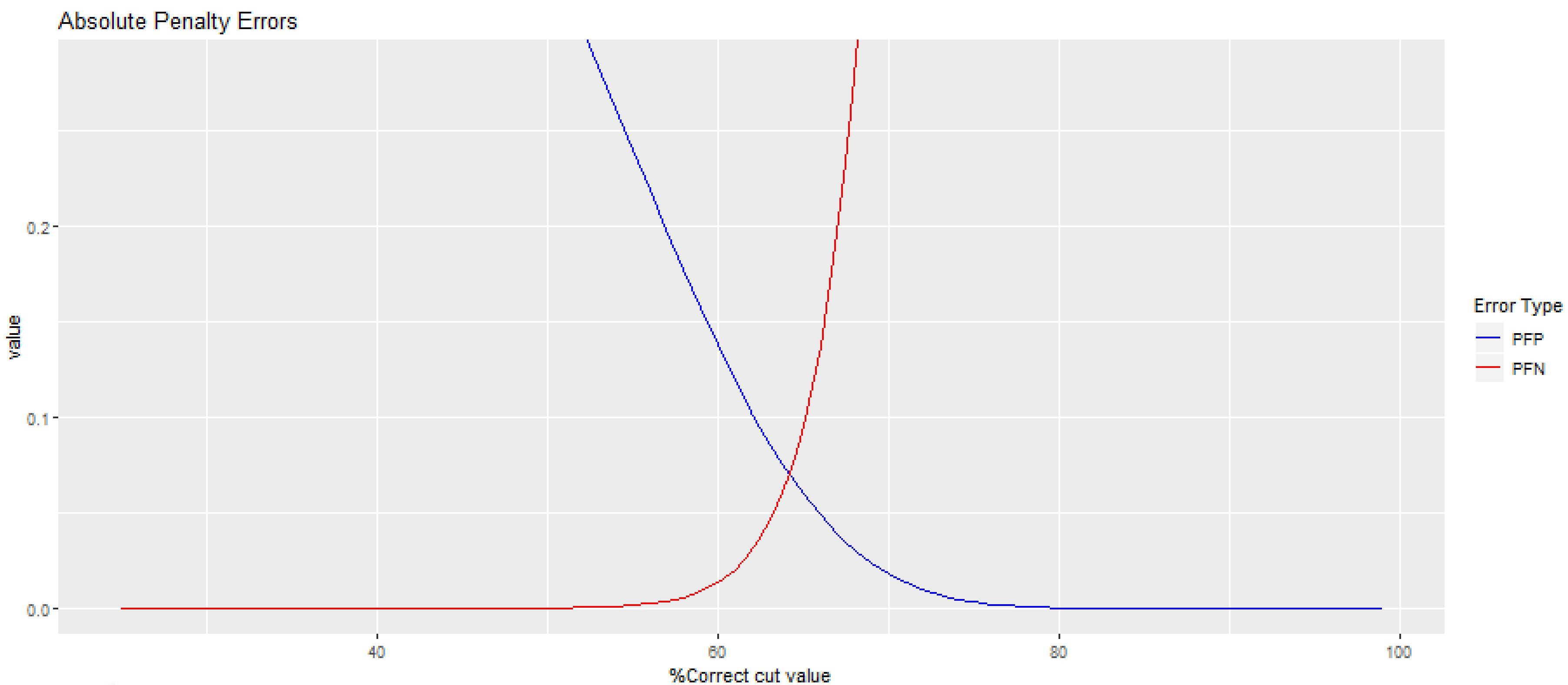
```
the optimal total error value is 0.057 {fp =  0.043 fn = 0.014 }
%C at min is 62.6 , no conversion indicated
```



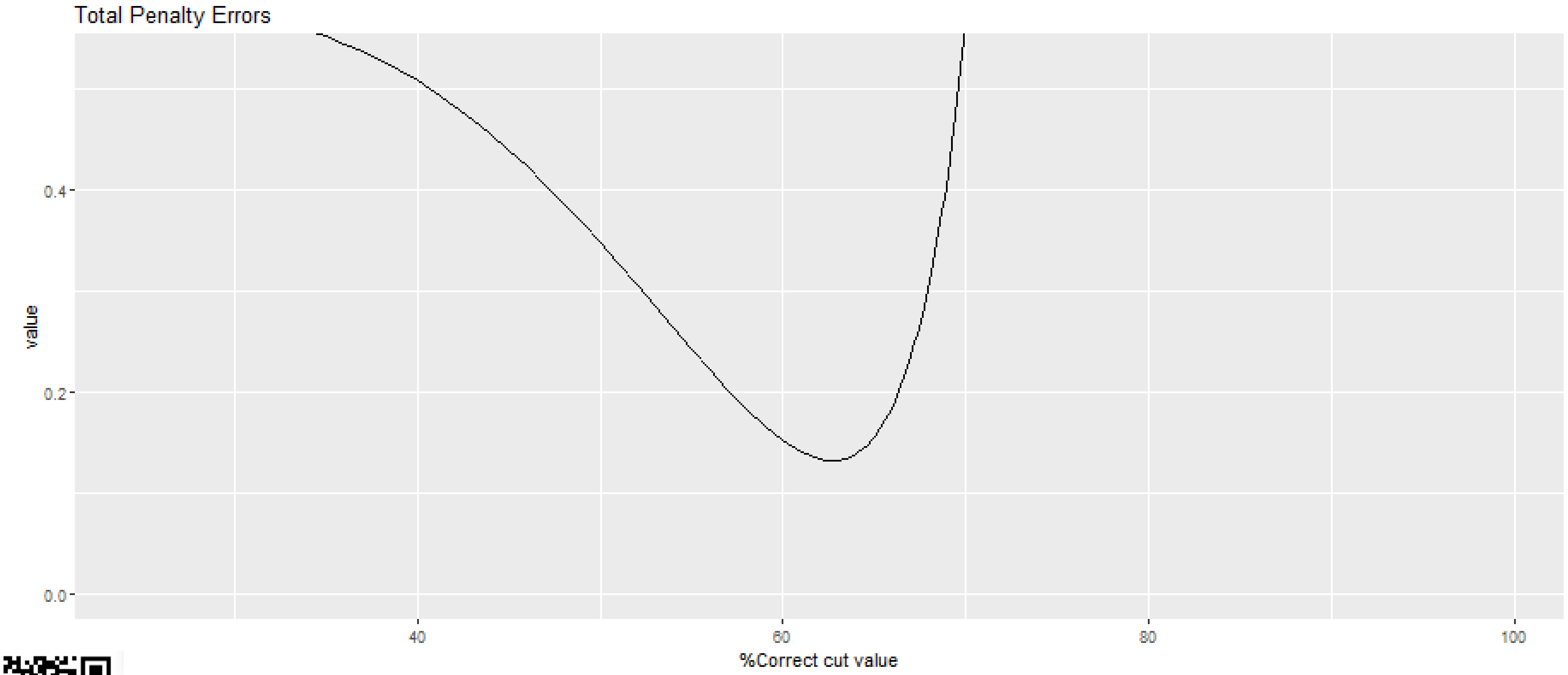
Penalty Based Error



Penalty Based Error Output



```
the optimal penalty total error value is 0.132 {fp =  0.092 fn = 0.04 }
%C at min is 62.7 , no conversion indicated
```



Conclusion

- Standard setting panels can use information about examinees and the exam to predict classification error
- This information may help inform increasing or lowering a the cut score
- Standard setting committees can choose to treat the estimated true cut score as known or as a random variable
- Software makes this process approachable to all
  - Software located at: https://drive.google.com/drive/folders/1qB3vMXqJ8PE3m9Y_MXehYil_osObICbW?usp=sharing

Ongoing Research

- Improvements to App (including UI improvements thanks to our colleague Christopher Runyon)
  - Look for updates here https://github.com/runyoncr/
  - Or here https://github.com/reypace
- Simulation studies to investigate robustness of violations to assumptions, and accuracy in various manipulations