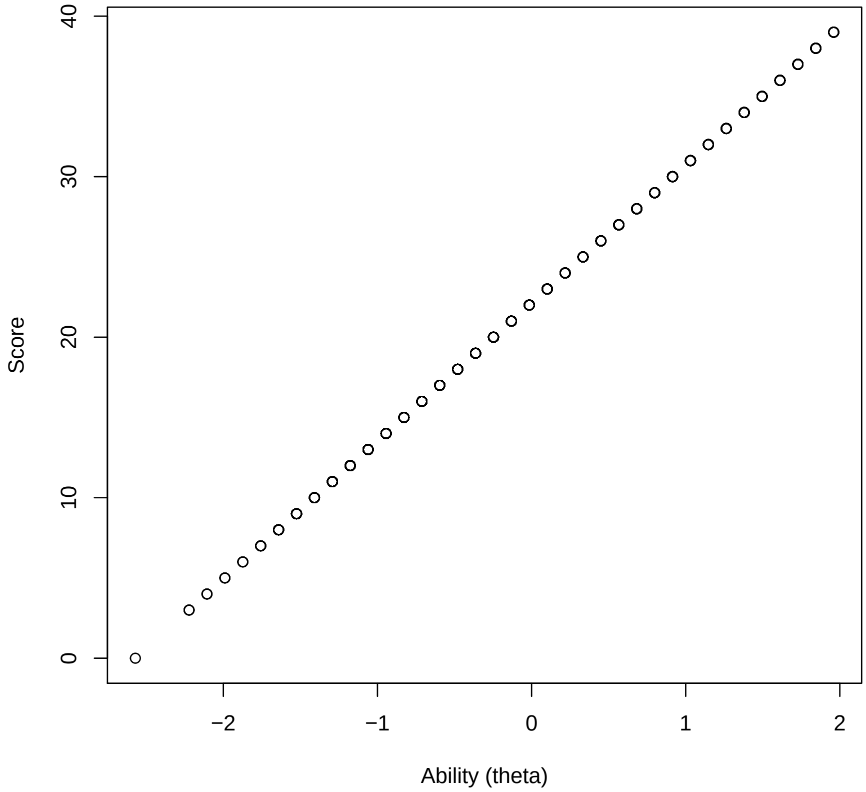
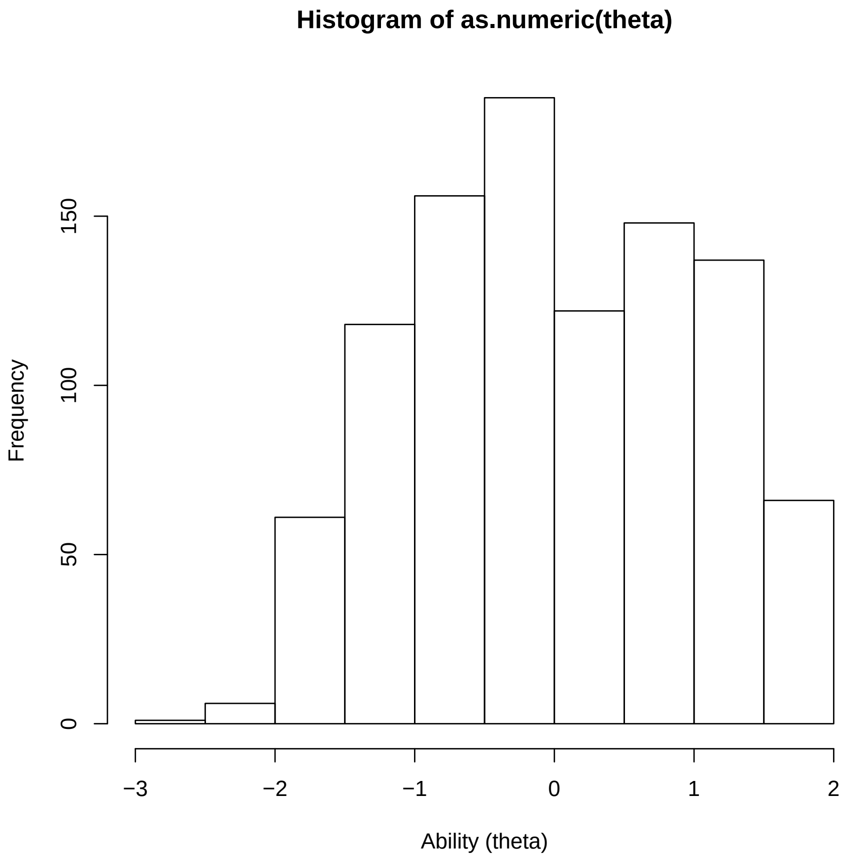
A CTT approach would not be a responsible choice for the NDE to use with a high-stakes test. It would be easier and cheaper but much less meaningful. CTT provides nothing more than total scores. With CTT, ability is related only to total score. The only information we can get from a set of test scores is the total score of each of the test-takers, as seen below, based on the data of 1000 test-takers:

```r
resp<-read.table("https://raw.githubusercontent.com/ben-domingue/252L_winter2018/master/data
rowSums(resp)->scoresNDE
scoresNDE
scoresNDE[5]
mu = mean(scoresNDE)
s = sd(scoresNDE)
theta = double()
(scoresNDE[65]-mu)/s
scoresNDE[54]
for (i in 1:1000) {
  theta[i] = (scoresNDE[i]-mu)/s
}
plot(unlist(theta), unlist(scoresNDE), xlab = 'Ability (theta)', ylab = 'Score' )
```



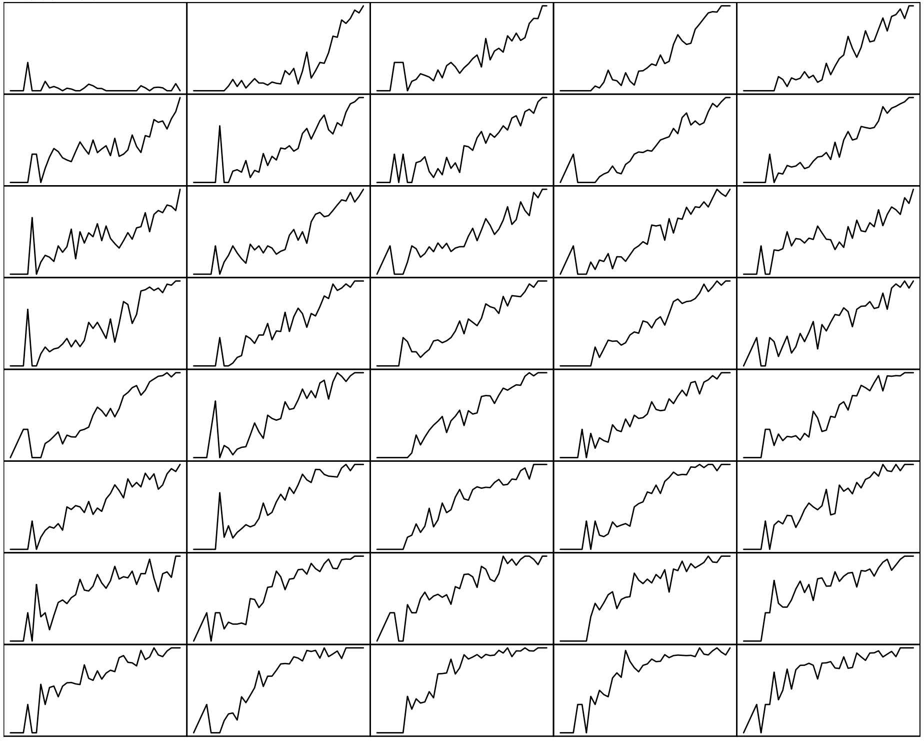
```r
hist(as.numeric(theta), xlab = 'Ability (theta)', ylab = 'Frequency')
```



For high-stakes situations where we are using our analyses for student grade placement, teacher and school evaluation, and school funding decisions, this CTT information is not nearly as meaningful as the analyses we will get with other models. Most notably, in these graphs, there is no information about the test itself. We can not even make claims about the validity of the test results or calculate useful standard errors. With IRT, we can understand and make strong claims about the test itself. This is even more the case when using the 3PL model.

For instance, with IRT, we can take a look at individual test items to determine how useful they are and whether or not we should throw some out or add more of certain types to make more meaningful claims.

In the following chart, for each test item, test scores are compared to the percentage of students who answered that item correctly. This can give us information about which items are easier and which are harder. This chart is organized from the hardest items in the upper left, where almost no one answered correctly, to the easiest in the lower right, where at least some students at nearly all score levels answered correctly. This kind of analysis would tell us that we should consider dropping the first item because none of our students are answering it correctly, so it could be throwing off our analyses. This kind of analysis could also show us if some items are behaving in such a way that it does not follow an increasing pattern across student scores, in which case we should review whether or not that item has an element construct irrelevance, for instance.



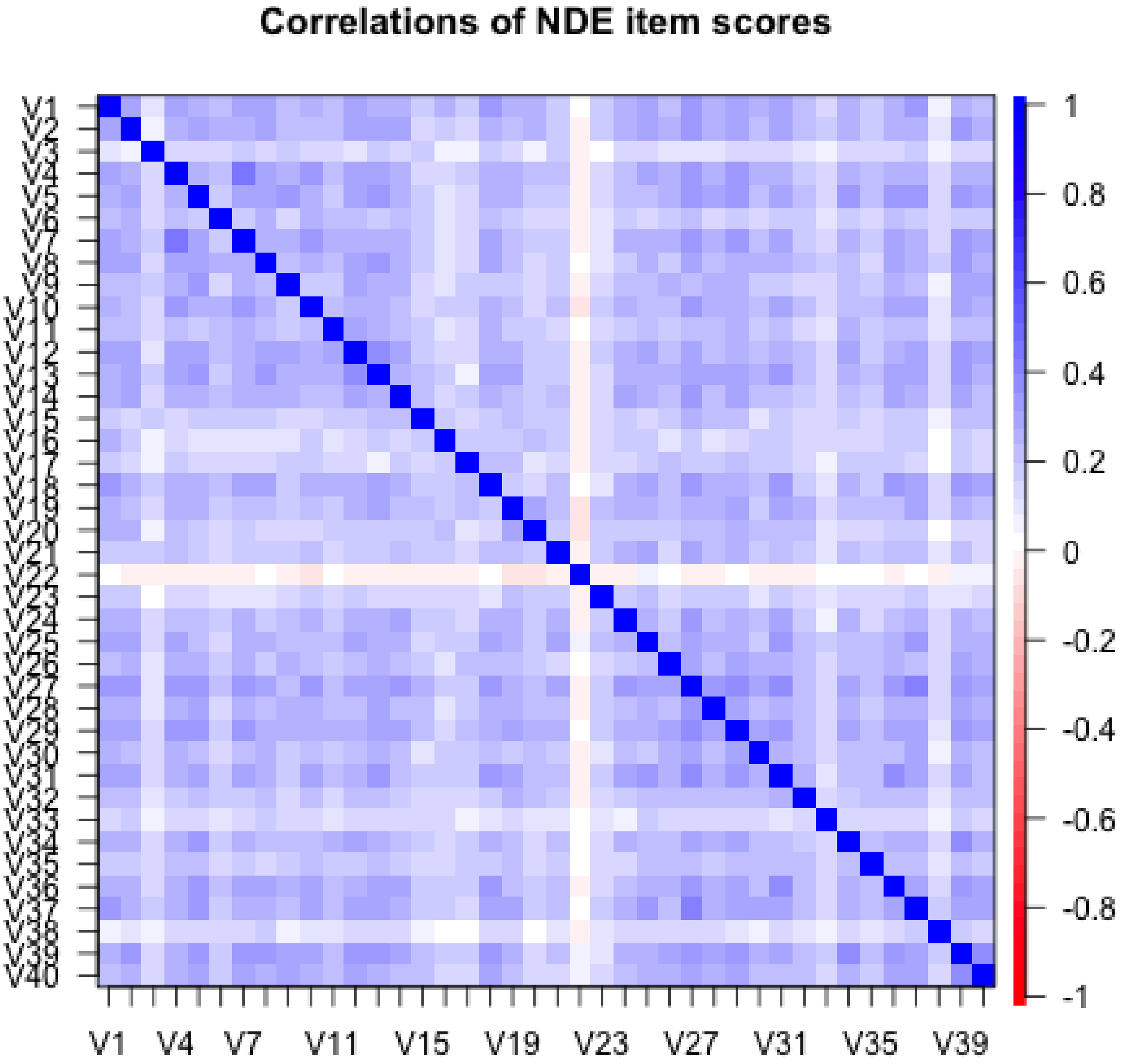
Now, the Rasch model only gives us information about item difficulty, but 2PL adds information about item discrimination. Furthermore, the 3PL model includes information about the contribution of guessing to the overall score. These analyses reduce error in our overall scores.

In IRT, we can decide if we have items that are not useful for the construct we are measuring by analyzing whether or not items are acting similarly among students who are at similar levels. One way IRT lets us do this is by looking at item correlations. For instance, given the dataset, we can look at the correlations between scores on different test items. In this chart, white means no correlation, red means negative correlation, and blue means positive correlation. We can see that a few items are qeakly correlated, and one item is negatively correlated with the other items. The line that is showing up with some red is the line associated with the item in the upper left corner of our previous chart. By showing that the trends in item responses to this item are not correlated with the trends in other items, it gives us further evidence that this item may have construct irrelevant variance. The lighter rows or columns also indicate some items that may be weaker and should be revisited. On the other hand, if items are showing up with very high correlations or dark blue squares, we will be alerted that we may have mirroring items that we need to cull. Using analyses like the ones above, we can also detect cheating across schools or within classrooms. If trends in student responses to certain items are widely variant from the rest of the population, we can flag a set of tests for further review.

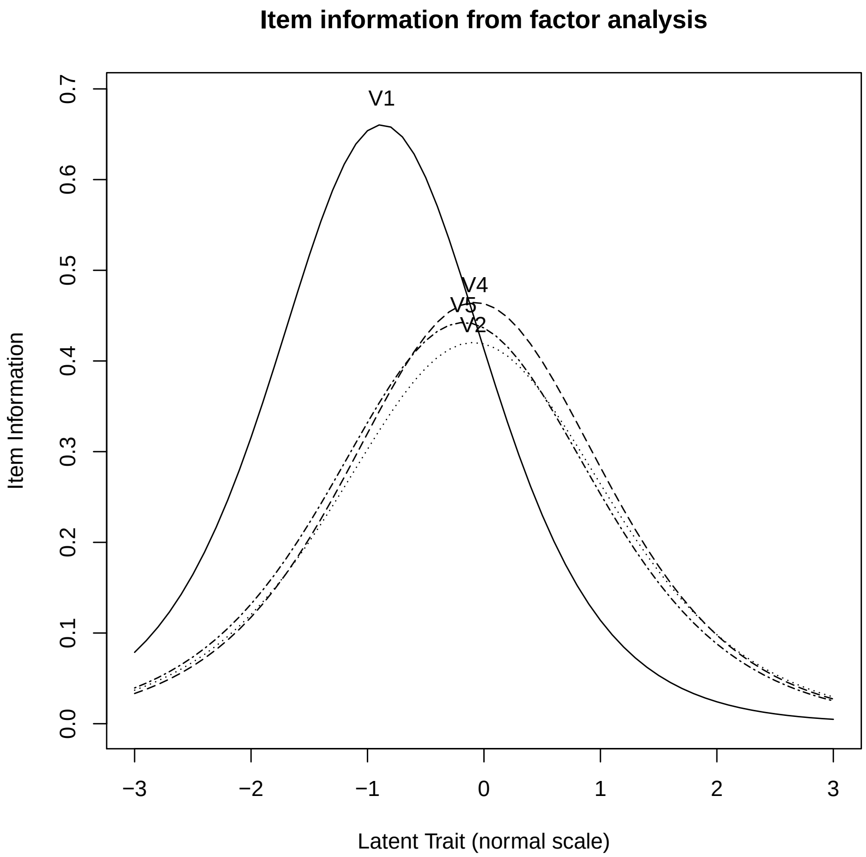
With 3PL, you can also test whether or not a test is biased for or against a group of students by running DIF analyses across demographics or if a test is vertically scaled well. You can test if items are dependent or independent. It gives you information regarding how much guessing plays a factor in test scores and wheter the test is unidemensional or multi-dimensional. All of this is lost in switching to CTT.

Over the Rasch Model, I recommend 3PL because it also takes into account guessing, which is a major contributor to error in multiple choice test scoring, especially at lower levels of ability. Like 2PL, it also takes into account item discrimination. This greatly reduces error because it gives us more accurate information curves. Items with high discrimination give us more information at the ability associated with their difficulty level, and items with low discrimination will give less. Analyzing all test items for the amount of information they provide helps us determine whether the test we are using is going to be useful for us. If we can not get a lot of information across the whole spectrum of abilities, or if there are major gaps in information, we will want to know that before we administer a deeply consequential high-stakes test.

For example, in the following chart, we look at the information curves forthe first five items in our sample.

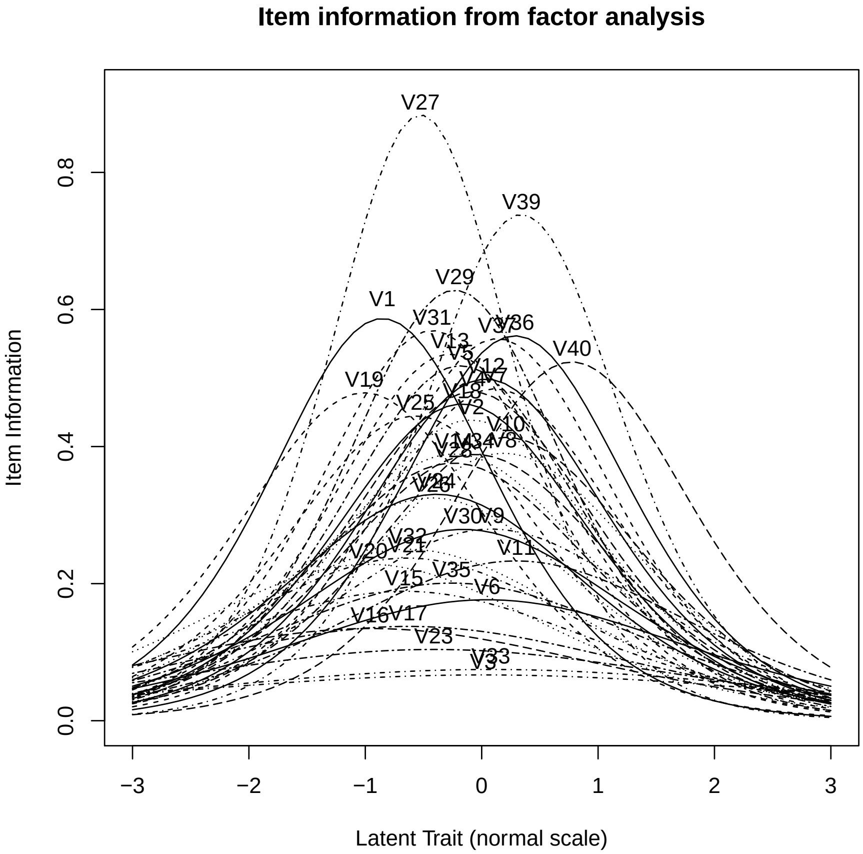


Correlations of NDE item scores



This shows that the first item offers the most information around -1, and the rest offer the most information around 0. If all of our items looked like one of these, we would want to know because these items do not give us much information about the students 1 standard deviation above the mean or 2 standard deviations below.

If we look at all 40 items, we can analyze the whole sample test.



As you can see, we do not have a lot of information at the low or mid-high ranges, which tells us that we will have higher errors in our scores at these ranges.

This also shows us that we have some items that are so poorly discriminating, they may be worth replacing as they are not contributing greatly to our overall measurement goal.