PS3
Edu 252L

- Due 2-22 at noon
- Instead of submitting via canvas, please email ONE (1) group copy to bdomingue@stanford.edu. This will make it easier for me to share feedback with entire group.
- Word or pdf are fine. If you submit a pdf, please be sure to include line #s.

Shortish Answer

1. Suppose that we have a test scaled with the Rasch model whose first 3 items have known difficulties -1, 0, and 1.5. An examinee with ability theta got the first item right, the second item right, and the third item wrong. Can you write the likelihood of observing this sequence of item responses as a function of theta?
2. Can you plot this as a function of theta?
   If you want to cheat:

```
th<-seq(-3,3,length.out=1000)
p<-function(b) exp(th-b)/(1+exp(th-b))
plot(th,p(-1)*p(0)*(1-p(-1.5)))
```

3. If theta=0.5, what is the likelihood of that response sequence?
4. If theta=0.5, what is the most likely response sequence given the known item difficulties?
5. At what value of theta does a response sequence of 1-1-0 (that is: they got the first and second items right and the third item wrong) become more likely than a response sequence of 1-0-0?
6. Returning to questions 1 and 2, can you plot the “test information” as a function of theta (see Eqn 2-6 in Lord).
7. Where is the function in #6 maximized? What do you think this implies?

8. For an item response dataset of your choosing, consider the relationship between theta and the SE across the three IRT models for dichotomous items. How much of a difference does the choice of model have on the size of the error estimate?

Consulting Exercise

A large-scale standardized test in mathematics is administered every year to all grade 3-8 students in Nebraska. Each grade-specific test consists of 40 multiple-choice items, and the test is used for summative purposes: to evaluate whether Nebraskan students are making "adequate yearly progress" (in the NCLB sense) in mathematics. In the past the NDE's testing contractor (Pearson) had used the Rasch Model to calibrate items and place students onto a score scale before these scores were categorized according to discrete proficiency levels. But recently, NDE has switched contractors (CTB McGraw-Hill), and the new one is suggesting a switch to the 3PL model.

You have been hired by the NDE as a consultant. They have two specific questions for you.

A. Some of the NDE leadership are wondering whether we wouldn't be better off abandoning the use of IRT altogether for our assessment program. What's wrong with taking a CTT approach and reporting student scores solely in a percent of total metric? Wouldn't this be a lot cheaper to do and easier to communicate to people?
B. If we are to continue using an IRT model, what do you see as they key pros and cons of using the 3PL relative to the Rasch Model? Should we follow the advice of our new contractor?

The NDE has provided you a randomly selected sample of item responses from 1,000 students in the state of took the grade 8 test form in 2009 [data]. In addressing the questions above, the NDE would like you to use this data to support your arguments. In making your arguments you should consider the following as necessary:

- Estimate the statistical models that you deem appropriate.
- Compare the interpretation of item "quality" across different measurement models (this should include comparisons within IRT models but also between IRT and CTT approaches).
- Discuss model fit.
- Compare the estimated proficiency of respondents across models.
- Compare plots of standard errors and information as a function of estimated proficiency across models.

Criteria

-You need to make an actionable recommendation to NDE.
-This recommendation needs to be supported by evidence.
-The relevant technical details are presented in a manner that is both technically correct but also interpretable by an intelligent layperson.
-The text of the main report should be < 2000 words.