Item Response Theory Modeling Using the ASSISTments Math Dataset

Background: One of the issues that educational assessment companies face is the existence of unfair or biased items in their assessments; test-takers from different demographic groups may perform differently even though they possess the same abilities. Being able to identify and reduce such unfair items is important especially in high-stakes contexts like standardized testing for college admissions. Differential Item Functioning (DIF) analysis is designed to address this problem. (Martinková, 2017) observed that, in a DIF analysis of a physiology assessment, the men outperformed women in total scores but that there was no case of DIF items. Thus, looking at total scores alone is not enough to identify unfairness. In another case study, involving simulated data, they demonstrated that DIF items can be present even when the distributions of total scores between two groups are the same; again, this demonstrates that relying on total scores alone is not sufficient. Their findings highlight the importance of DIF analysis.

Another issue is that simply administering the same exam to all students and counting how many problems they got right isn't the most precise way to measure student ability. In addition, it isn't the most efficient way to measure student ability. Furthermore, the student experience of exam-taking may not be comfortable if the exam has too many problems, takes a long time to take, and contains many problems that are not at the right level for the particular student. Computer Adaptive Testing (CAT) helps with this issue by allowing problems to adapt to the student's ability as the student progresses through the exam thus reducing the number of problems, the time it takes to take the exam, and the number of ill-aligned problems. At the same time, CAT improves the measurement precision of student ability.

Item Response Theory (IRT) provides a foundation for DIF and CAT by allowing us to estimate the abilities of students and the difficulties of items. In our project, we will perform IRT modelling for one of the skills in the Assistments dataset.

Business Justification: IRT models allow us to estimate student ability and item characteristics (difficulty, discrimination) on a common scale. These estimates allow assessment companies to build shorter tests that measure ability with the same precision; instead of a fixed set of items for everyone, CAT selects items that are most informative for a test-taker's ability level, reducing test length while maintaining precision. Because IRT parameters are approximately sample-invariant, companies can maintain large item banks without throwing tests away yearly. IRT allows for more meaningful interpretation of student scores on assessments. DIF allows for the identification of unfair items. Thus, IRT, DIF, and CAT provide educational assessment companies with cost-savings by improving operational efficiency as well as with support for the legal defensibility and compliance standards of their assessments.

Overview of Data: The Assistments Skill Builder 2009-2010 Dataset (Skill Builder 2009-2010 Dataset) is a large, anonymized collection of data on student responses to math assessments at the K12 level. The dataset is comprised of 346,860 records of student responses to various math problems assessing different skills; the dataset spans students from 75 schools in the year 2009-2010. More information can be found in (Heffernan, N.T. et al., 2009). The data is collected from the Assistments platform which tutors students as it assesses them, providing hints and scaffolding questions as students work through problems. Some problems assess multiple skills and, for the purposes of this project, the dataset has been filtered to only those problems that assess single skills. Furthermore, we filter to only one skill—skill 311—which is called 'Equation Solving Two or Fewer Steps' because it was the most abundantly represented in the data. We also filter to only free-response problems as there were 'choose 1' and free-response problem types. The problems are generated from templates so that many different problems fall under the same template. We keep only the student's first attempt for a given template so that the data has one row per student-item; the columns are `student_id`, `item_id`, `is_correct`, and `school_id`. The final dataset, called `df_irt`, has 5,524 rows; there are 959 unique students and 15 unique template ids. The templates will be the items in our IRT modeling.

1PL Model: The first thing we do is fit a 1PL model; we maximize the log-likelihood of observed responses to estimate student abilities (theta's) and item difficulties (b's). The minimum difficulty is -1.9, the maximum difficulty is 0.89, and the mean difficulty is -0.38. The easiest items are items '46280' and '30833' while the most difficult items are '55508' and '55901'.

The minimum student ability is -2.48, the maximum student ability is 2.21, and the mean student ability is 0. The lowest ability students are '84122', '82676', '90010', and the highest ability students are '96239', '96215', and '78907'.

In Figure 1, we have the histogram of student abilities. The histogram looks relatively normal.

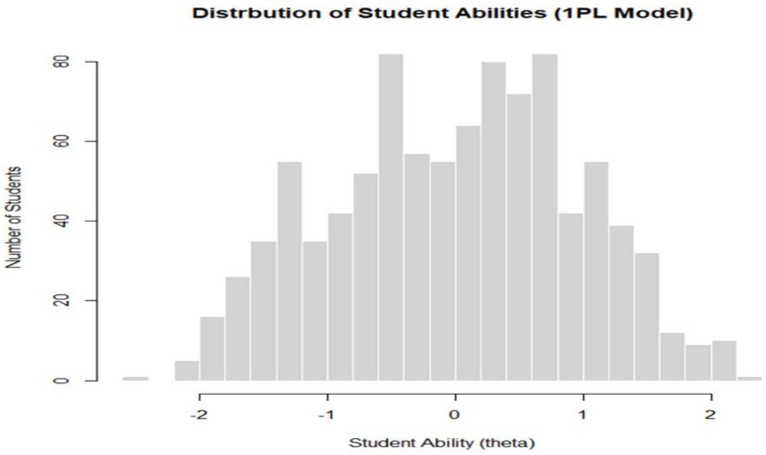

Figure 1

In Figure 2, we see the item-person map. The mean item difficulty is -0.38 and the max is 0.89 while the student ability distribution is roughly normal with mean 0 and spanning between roughly -2 and 2. This means that the items are easy relative to the student population. This item–person map indicates stronger measurement precision for lower-to-mid ability students and reduced information for higher-ability students.

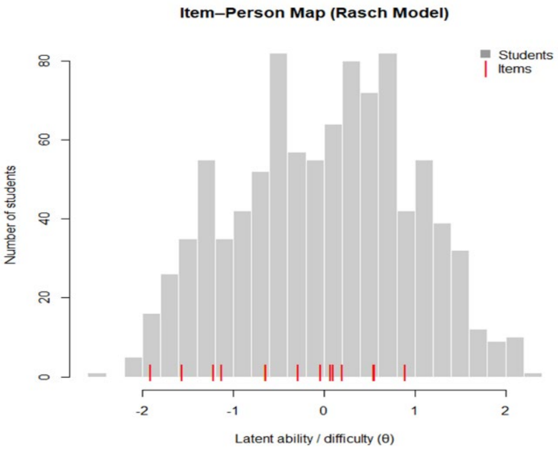

Figure 1

We can plot the item difficulty and p-value (from classical test theory) together, as seen in Figure 3. By plotting the classical item difficulty as proportion correct compared to Rasch difficulty estimates, we see a clear downward trend, as expected; items with lower p-values correspond to higher estimated difficulty parameters, confirming alignment between observed and latent difficulty.

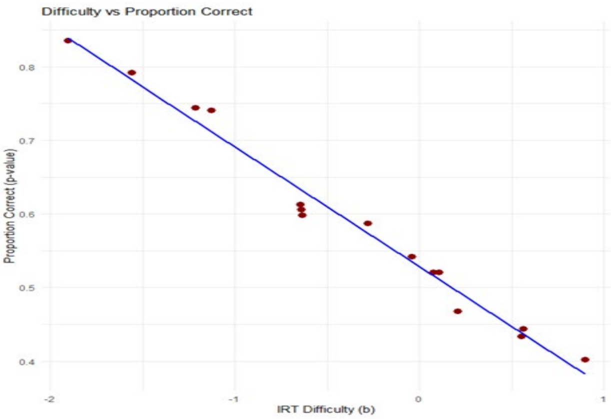

Figure 3

In Figure 4, we plot item characteristic curves for the hardest, medium, and easiest items. The easiest item ICC is shifted more to the left while the hardest item ICC is shifted more to the right. All curves have the same slope of 1 at their respective b values. In the Rasch model, we assume all items have the same discrimination.

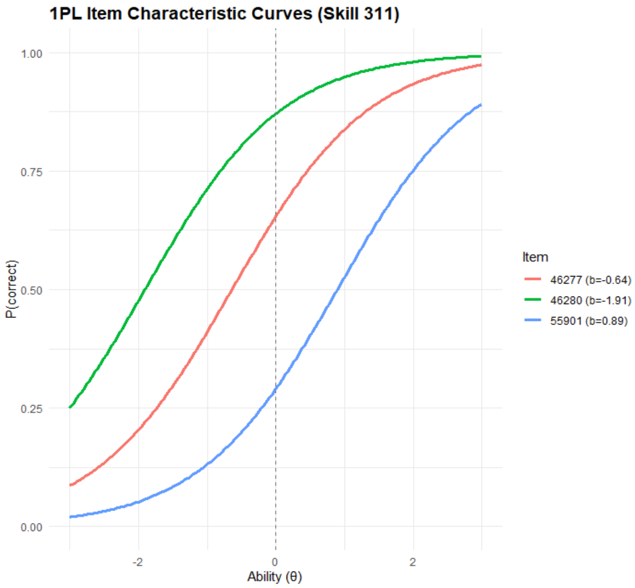

Figure 4

In Figure 5, we plot the test information curve. The test information function peaked at θ = -0.33, indicating that the assessment measures most precisely students who are slightly below-average ability. It is less accurate for high-ability or very low-ability students.

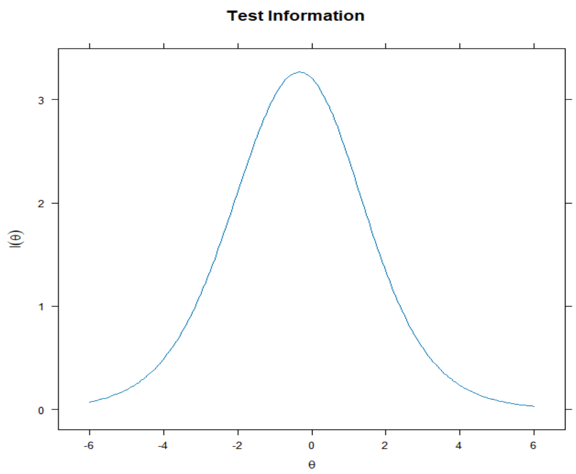

Figure 5

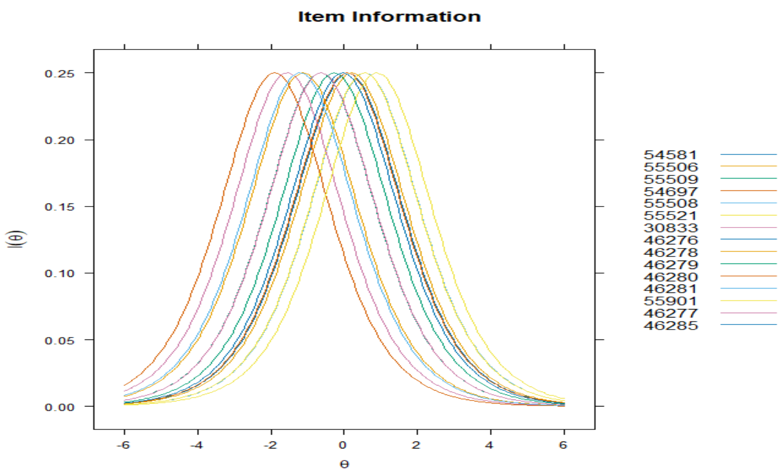

Figure 6

In Figure 6, we plot all 15 item information curves; notice that they all have the same peak height, and the only difference is the location of their peaks along the difficulty scale.

2PL Model:

The next thing we do is fit a 2PL model which take into account item discrimination as well as item difficulty. The minimum difficulty is -1.45, the maximum difficulty is 0.81, and the mean difficulty is -0.24. The easiest items are items '46280' and '55506' while the most difficult items are '55508' and '55901'.

The minimum student ability is -1.86, the maximum student ability is 1.68, and the mean student ability is 0. The lowest ability students are '84122', '82676', '90010', and the highest ability students are '96215', '96292', and '96239'.

In Figure 7, we have the histogram of student abilities. The histogram looks relatively normal.

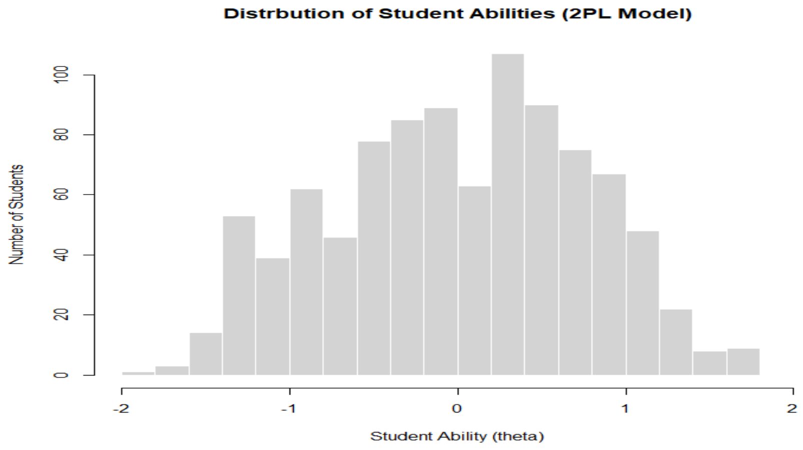

Figure 7

In Figure 8, we see the item-person map. The item-person map shows that student abilities are distributed approximately normally around θ = 0, with most students falling between –1.5 and +1. Item difficulties cluster between –1 and +1 on the same latent scale. This indicates good targeting for average-ability students, but limited measurement precision for individuals at the extremes (θ < –1.5 or θ > +1). Additional easy or hard items would improve coverage across the full ability range.

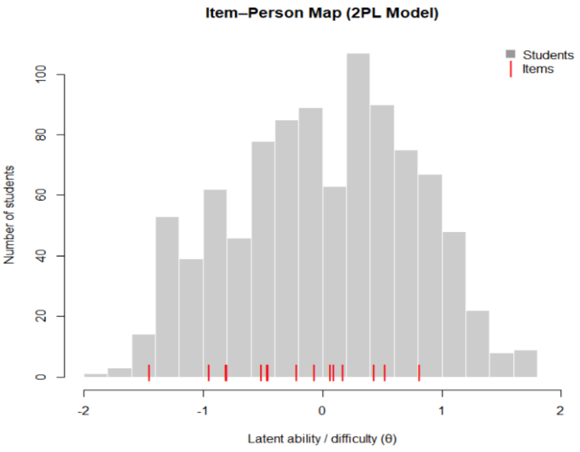

Figure 8

We can plot the item difficulty and p-value (from classical test theory) together, as seen in Figure 9. By plotting the classical item difficulty as proportion correct compared to 2PL difficulty estimates, we see a clear downward trend, as expected; items with lower p-values correspond to higher estimated difficulty parameters, confirming alignment between observed and latent difficulty.

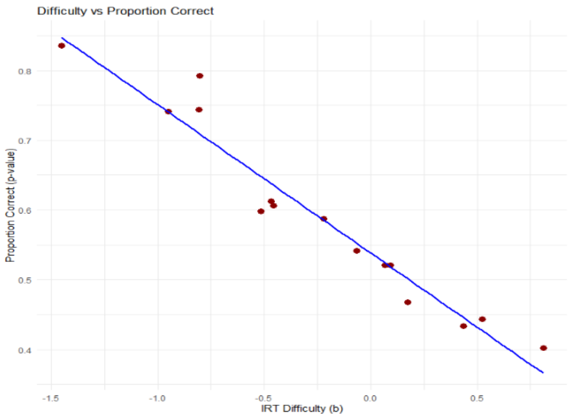

Figure 9

In Figure 10, we plot item characteristic curves for the hardest, medium, and easiest items. The easiest item ICC is shifted more to the left while the hardest item ICC is shifted more to the right. The easiest and medium curves have higher discrimination than the hardest curve. In the 2PL model, we allow items to have different discriminations.

In Figure 11, we plot all the item characteristic curves together. We notice that a couple items have relatively steep slopes while most of the curves have moderate slopes.

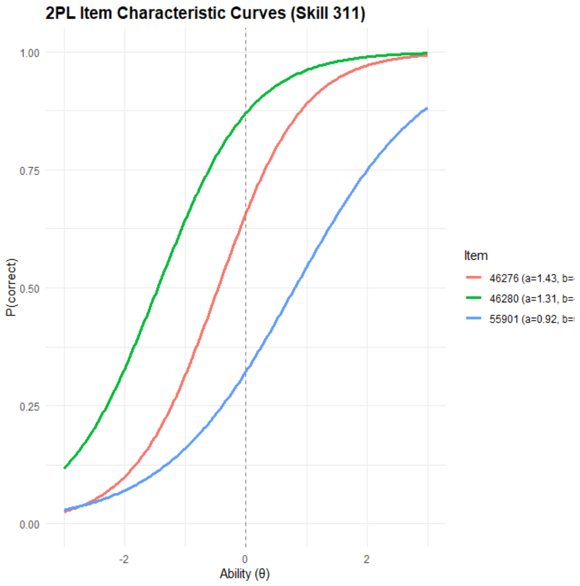

Figure 10

The mean discrimination is 1.38, the min discrimination is 0.79, and the max discrimination is 2.83. Most items discriminate reasonably well, and there are two items that have very high discrimination—one at low ability, and one at medium ability.

In Figure 12, we plot the test information curve. The test information function peaked at θ = -0.58, indicating that the assessment measures most precisely students who are slightly below-average ability. It is less accurate for high-ability or very low-ability students.

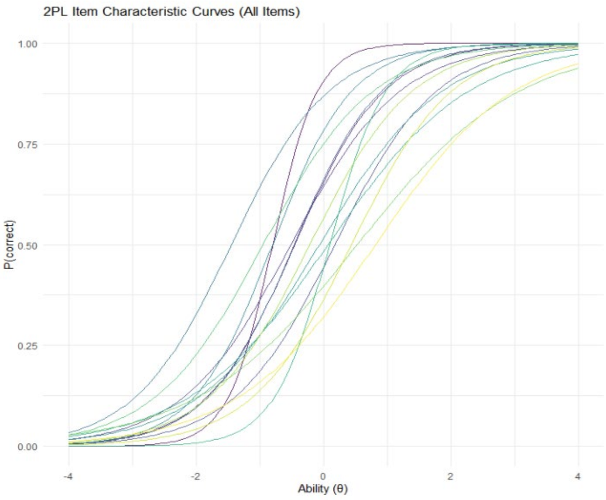

Figure 11

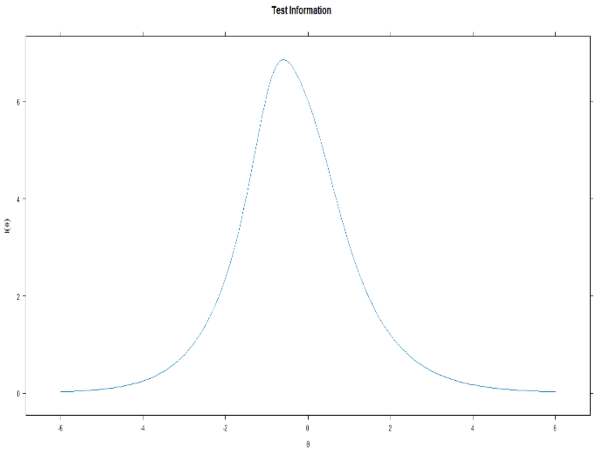

Figure 12

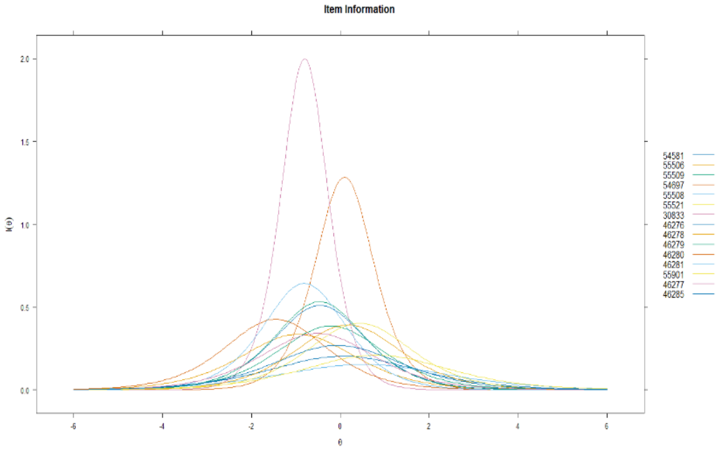

Figure 13

In Figure 13, we plot all 15 item information curves; notice that two of them have relatively high peaks. They correspond to the two items with highest discrimination.

Conclusion: We performed 1PL modeling and found that the items are easy relative to the student population and that there is stronger measurement precision for lower-to-mid ability students and reduced information for higher-ability students. The assessment measures most precisely students who are slightly below-average ability. It is less accurate for high-ability or very low-ability students. We also performed 2PL modeling and found good targeting for average-ability students, but limited measurement precision for individuals at the extremes. Additional easy or hard items would improve coverage across the full ability range. Most items discriminate reasonably well, and there are two items that have very high discrimination. Just as in the 1PL case, the assessment measures most precisely students who are slightly below-average ability. It is less accurate for high-ability or very low-ability students.

References

Feng, M., Heffernan, N.T., & Koedinger, K.R. (2009). *Addressing the assessment challenge in an Intelligent Tutoring System that tutors as it assesses. The Journal of User Modeling and User-Adapted Interaction, 19*, 243-266. https://doi.org/10.1007/s11257-009-9063-7

Martinková, P., Drabinová, A., Liaw, Y. L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). *Checking Equity: Why Differential Item Functioning Analysis Should Be a Routine Part of Developing Conceptual Assessments*. *CBE life sciences education, 16(2), rm2*. https://doi.org/10.1187/cbe.16-10-0307