Knewton Machine Learning Challenge

Dan Kenefick

April 14, 2015

1 Introduction

I am a visiting professor at Mars University teaching Astrometrics. Because of limited resources, the midterm will take the form of five multiple choice questions, which are to be selected at random from a question bank. I have been tasked with choosing the questions that will comprise the question bank based on a dataset of prior questions and responses. Specifically, I should select those questions such that the exam results will provide the most meaningful ranking of the students. To do this, I follow an intuitive approach that selects questions that discriminate well among students in terms of their knowledge of Astrometrics. These questions are selected at varying levels of difficulty that reflect the underlying distribution of student knowledge. The questions I select should be able to rank the students taking the midterm in terms of their knowledge of Astrometrics.

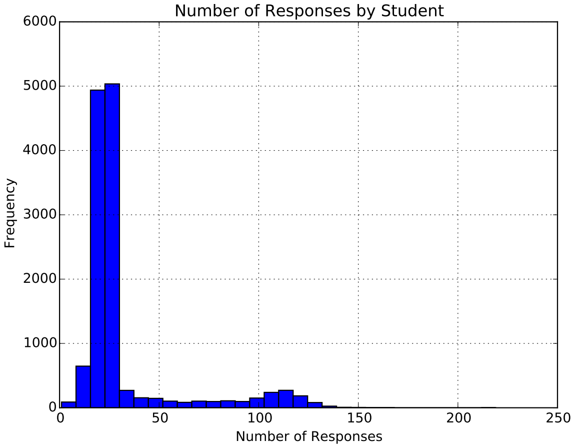
2 Motivation

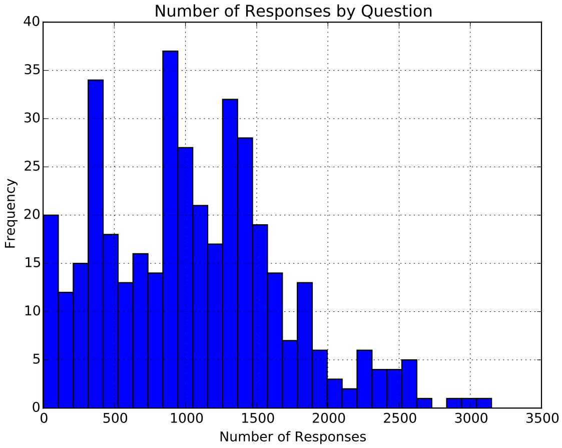
First, we need to decide what our criteria should be for identifying those questions that will give a meaningful ranking of students. A meaningful ranking is one that orders students according to their knowledge of Astrometrics. Therefore, in order to rank students effectively, our questions should be able to differentiate good students (one with a mastery of Astrometrics) from bad students (one with little knowledge of Astrometrics). In other words, we would like to choose questions that good students will get right, and bad students will get wrong. We will call this desirable characteristic discrimination.

Furthermore, we need questions can distinguish students of varying levels of knowledge. It is not enough to be able to distinguish good students from bad, but we must also discriminate bad students from terrible students, and great students from good students. Therefore, we will need to choose questions that are highly discriminating at different levels of student knowledge in order to supply a meaningful ranking. Intuitively, this means we should select questions of varying difficulty. We should also be relatively sure about the difficulty of the questions we choose. Otherwise, we may select a question that we believe to be highly discriminating for one kind of student, but is in fact highly discriminating for another kind. For example, we may think that a question is discriminating against good and great students, when in fact it is discriminating against good and average students. Any such confusion will be problematic for the ultimate ranking of the students who take the mid-term.

3 The Data

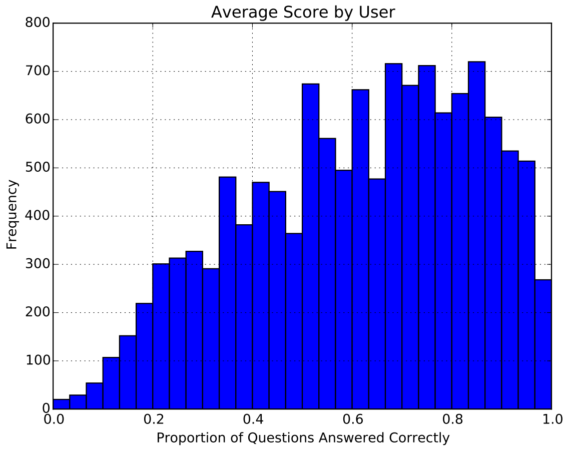
The data consist of 409,519 observations, representing the responses of 12,839 students to 391 questions. The number of responses is bimodal: the larger peak indicates that about 10,000 students answered a total of 25 questions, and the smaller peak indicates that about 250 students answered about 110 questions. The number of responses by question is highly variable, with notable peaks at just under 500, 1,000, and 1,500 responses. Histograms of both number of responses by student (user) and number of responses by question can be seen below.



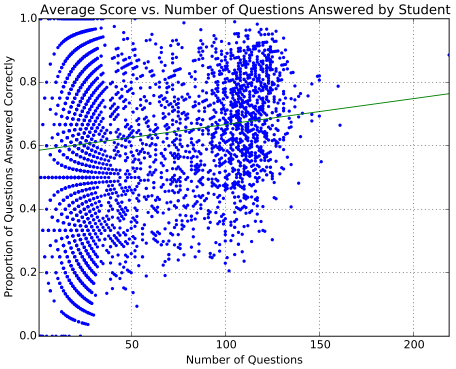


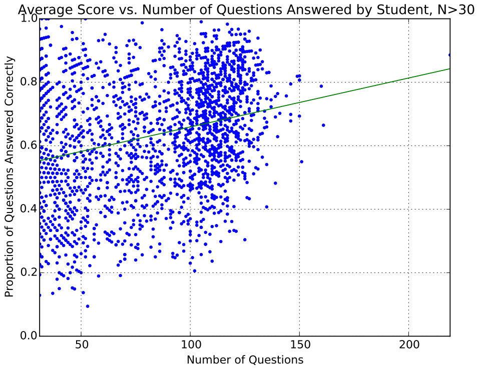
Similarly, there is a wide range of average scores by student. The scores range from 0 percent correct to 100 percent correct, where the 5th percentile represents an average score of 21 percent, and the 95th percentile represents an average score of 95 percent. This 5th percentile suggests that there may be 5 potential responses per multiple choice question, and that students who guess on every item may get a minimum of 20 percent correct. There are also a small number of duplicate questions per student, which are accounted for in the analysis that follows.[1]

[1] A question was answered at most twice by a given student, but a question may have more than one student with duplicate answers. There are 67 duplicate values. If a student answers a question twice, and answers the question correctly or incorrectly both times, then one copy of the question is removed from the data. Otherwise, these questions may appear more consistent than they actually are. If, however, the student answers the question once correctly and once incorrectly, then both copies are removed. This occurs 13 times, so 26 observations are dropped. Without a prior as to whether these instances are human error, or a student taking a second guess at a question that they do not know the answer to (which have different implications for the difficulty of the question) it is safer to drop these limited instances.



There is also a slight (but statistically significant)[2] relationship between the number of questions a student has answered and that student's average score. This is possibly worrisome because it may indicate that the question items are not independent, which is a key assumption in most Item Response Theory (IRT) models. Scatter-plots showing the number of questions answered vs. average score with lines of best fit are shown below.





4 Methodology

The methodology consist of two parts. The first appeals to intuition of how we should select these questions, touched on briefly earlier. This approach is then validated with

[2] The statistical significance comes from an OLS model, estimated as Student Score$_i = \beta_0 + \beta_1 *$Number of Questions Asked$_i + \epsilon$ where i indexes the student. Estimates for both β_1 and β_2 are statistically significant.

a more formal IRT model, which will be developed further below.

4.1 Intuitive Approach

First, we would like to determine which questions discriminate good students from bad students, or more generally, different classes of students from one another. To develop what we mean by discrimination, consider a question (item) i, a good student with mastery of Astrometrics z_g, and a bad student with mastery of Astrometrics z_b, where $z_g > z_b$. If this item could discriminate good students from bad students, then the probability that a good student would get the question right should be higher than the probability that the bad student gets it right, or

$$P(right_i|z_g) > P(right_i|z_b)$$

Unfortunately, we do not observe student knowledge (the zs) or the probabilities of getting a question right (the Ps) directly, but we can generate proxies for both from the data. We can proxy student knowledge by a student's average score on all of the questions they answered. This method is potentially problematic[3], but it should be a useful measure of student knowledge if there are a large number of students in the data students and the ratio of good questions to bad questions is high. Using the average score, we could then put students into a good group with a high average score and a bad group with a low average score. Then, the probability of getting a given question right for good students is the good students' average score for that question, and the probability that bad students get the question right is the bad students' average score. Then, the above becomes:

$$P(right_i|z_g) \approx P(right_i|\text{High Overall Score}) \approx \text{Average Score}_{i,\text{High Overall Score}}$$

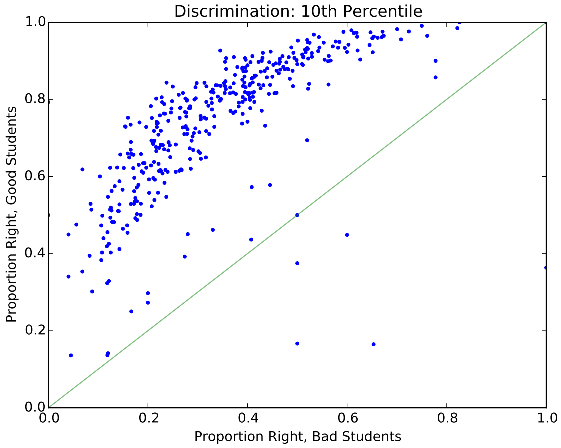
$$P(right_i|z_b) \approx P(right_i|\text{Low Overall Score}) \approx \text{Average Score}_{i,\text{Low Overall Score}}$$

Where i is the question, z_g and z_b indicate the unobserved good and bad students' knowledge, respectively, and the indexing on Average Score means that the average of answers to question i is taken for those students with a High or Low Overall Score.

Using these proxies, we can then determine how well each question discriminates between classes of students. To illustrate, let's see if a given question i discriminates the worst students from the rest of the students. To do this, we compare how those students with the bottom 10 percent of scores did on question i to how the students

[3] A student may, for example, only answer difficult questions, or answer questions that do not discriminate well between good and bad students. In both cases the average score may not be indicative of student knowledge

in the top 90 percent of scores did on question i. We would first determine each student's average score, and break the students into two groups: those in the bottom 10 percent of scores, and those in the top 90 percent. We would then compare how the students in each group did on question i. If the students in the top 90 percent of scores did better on question i than the students in the bottom 10 percent, then we can say that question i discriminates between these two groups. The bigger the difference in performance between the two groups, the better the question is at discriminating. If we repeat this exercise for all questions, we can see which questions are best at discriminating these two groups of students. A scatter-plot of the average scores of good students vs. the average scores of bad students for each question, where good and bad students are above and below the 10th percentile, respectively, is shown below.[4]



Anything above the diagonal represents a question that the good students scored higher on than the bad students. We can therefore see some easy candidates for rejection: anything below the diagonal, where bad students scored better than good students, are not good questions. Anything above the line, but close to it, is worse at discriminating between these groups of students than those farther away from the line.

We'd like to repeat this exercise for each class of students, but first we must determine what those classes should be. Above we divided students into classes based on how well they did on the questions they answered, which, as discussed, is

[4] For each question, the average scores for each user are computed the sum of all correct answers supplied by a student divided by the total number of questions answered, excluding the question being analyzed.

a proxy for their knowledge z. This is intuitively appealing: if the students who take the upcoming midterm are similar to the students last year (i.e., each student's knowledge, or z, is taken from the same distribution) then the questions that are best at discriminating the bottom 10 percent of scores in the data should be able to discriminate the bottom 10 percent of students in terms of ability in the upcoming mid-term. Similarly, the questions that can best discriminate the bottom 20 percent of scores in the data should be able to discriminate the bottom 20 percent of students in terms of ability, and so on. Accordingly, we would like to select questions that discriminate the various percentiles of scores in the data. Then, any mid-term consisting of 5 randomly selected questions will contain questions that are best suited to order students' knowledge relative to the underlying distribution, and therefore relative to their fellow students. The deciles of average students are given below[5]:

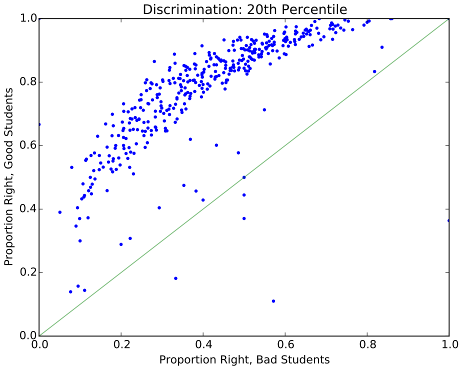
Decile	Average Score
1st	28%
2nd	39%
3rd	50%
4th	57%
5th	65%
6th	72%
7th	78%
8th	84%
9th	90%

Now that we have the classes that we would like to discriminate between, we can return to selecting questions. Performing the same exercise for the 20th through 90th deciles and each question yields the scatter-plots on the next page.[6]

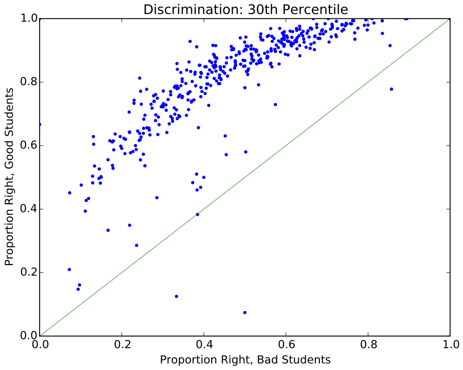
[5] I have removed the bottom 5 percent of users in terms of number of responses when creating these percentiles, to avoid potentially skewing the results. The choice to use deciles is somewhat arbitrary, I could have just as easily used quintiles (to reflect that the midterm consists of 5 questions or to break students into an A, B, C, D, F grading scale). There are some trade-offs of precision vs. being demanding of the data when choosing percentile length, however, and 10 seemed reasonable.

[6] As above, average scores are computed excluding the question being analyzed.

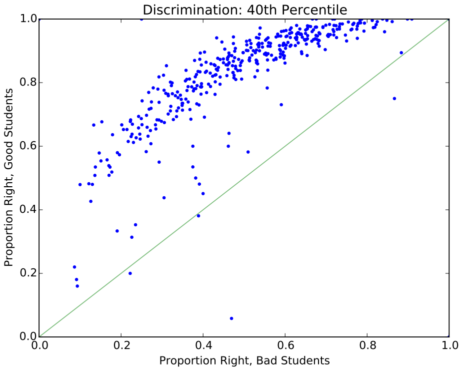
Discrimination: 20th Percentile



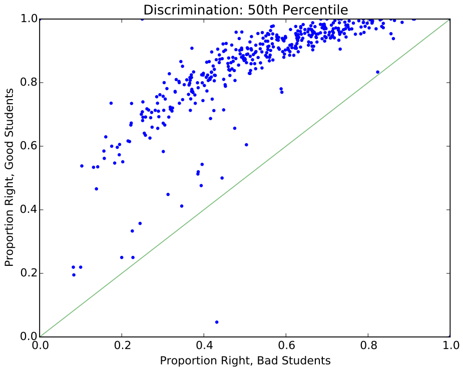
Discrimination: 30th Percentile



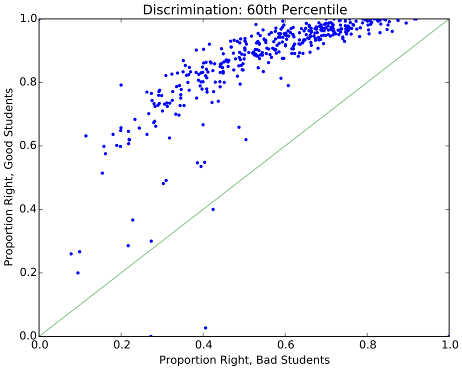
Discrimination: 40th Percentile



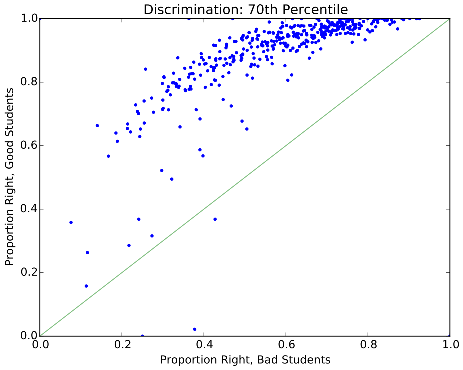
Discrimination: 50th Percentile



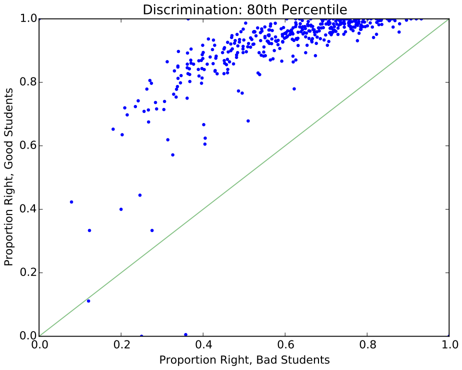
Discrimination: 60th Percentile



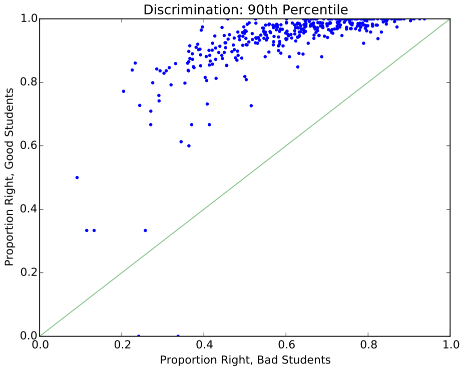
Discrimination: 70th Percentile



Discrimination: 80th Percentile



Discrimination: 90th Percentile



However, these plots do not in and of themselves select the most discriminating questions. We need to come up with a numerical measure of how well the questions discriminate at each percentile. One way to compare the probability of outcomes between groups is the odds ratio, which in this case would be:

$$\frac{P(right_i|z_g)/P(wrong_i|z_g)}{P(right_i|z_b)/P(wrong_i|z_b)}$$

Where i indicates the question, and z_g and z_b indicate the unobserved good and bad students' knowledge, respectively. Using our proxies for these values from the data, the above becomes:

$$\frac{\text{Average Score}_{i,\text{Overall Score}>p}/(1-\text{Average Score}_{i,\text{Overall Score}>p})}{\text{Average Score}_{i,\text{Overall Score}<p}/(1-\text{Average Score}_{i,\text{Overall Score}<p})}$$

Where i is the question, and p is the percentile of overall scores we would like to discriminate on. The indexing on Average Score means that, for students with an overall score $>$ or $< p$, the Average is taken over answers to question i. The above value, or odds ratio, is computed for each question, decile combination. The top 30 percent of questions in terms of the odds ratio are then selected for each decile.[7]

We would also like to be sure about the difficulty of each question, or the decile that it would be discriminating. To do this, we calculate the exact binomial confidence interval at the 95 percent confidence level for each question using every answer to the that question.[8] If the question has a confidence interval of wider that 10 percentage points, then we drop this question, even if it was selected via the odds ratio analysis above. This mostly has the result of dropping questions with few observations in the data. [9]

[7] Note that, since we are selecting the top 30 percent of questions in terms of odds ratio at each decile, we may select duplicate questions at different deciles (e.g., a question may be in the top 30 percent of questions in terms of odds ratio for the fourth and fifth decile).

[8] The lower bound for the binomial exact confidence interval at the 95th percent confidence level is calculated as $\frac{1}{1+\frac{n-x+1}{x}F_{2(n-x+1),2x,.975}}$ and the upper bound is calculated as $\frac{\frac{x+1}{n-x}F_{2(x+1),2(n-x),.975}}{1+\frac{x+1}{n-x}F_{2(x+1),2(n-x),.975}}$ where n is the number of responses to the question, x is the number of correct answers, and $F_{a,b,.975}$ is the 97.5th percentile from an F distribution with a and b degrees of freedom.

[9] As a possible extension, we may require that confidence intervals for questions be more narrow at difficulties where there are many students. For example, if many students are clustered around 80 percent in terms of score (as they are in the data), we may require questions with difficulties close to 80 percent to have narrower confidence intervals, so that we can be more sure about our ranking of these students. As of right now, we require questions that discriminate the 10th and 70th deciles to both be below the same width, even though the 10th percentile is 28 percentage points wide and the 70th percentile is 6 percentage points wide.

4.2 IRT Model

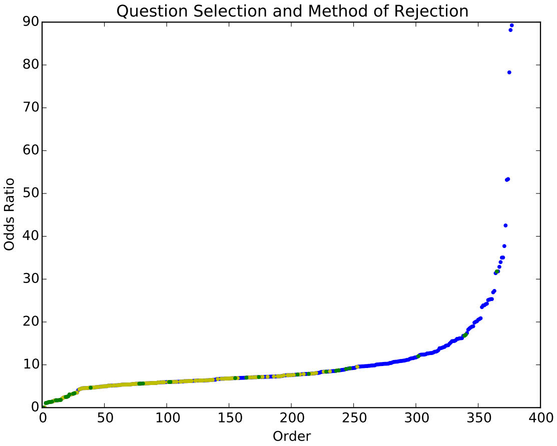
To validate the above analysis, we run two IRT models over the data. Briefly, an IRT model uses a maximum-likelihood framework (or some other method) to fit a curve to each item (question). This curve, called an item response function, shows the relationship between the probability of getting a question right and a students unobserved knowledge (referred to as zs above). The curve is typically fit via question-specific parameters that correspond closely to our understanding of how to select questions. For example, a two parameter model may have a parameter for the difficulty of the question, and another for the discrimination of the question[10]. We fit such a model to the data, and then select questions on the basis of maximum discrimination.[11] The results of this model is compared to the approach outlined in 4.1 (the intuitive approach) in the next section.

5 Results

Using the intuitive approach, we select 211 out of a possible 391 questions on the basis of a high odds ratio at some percentile and a sufficient confidence interval. To examine which ones we select, consider the following ordered plot of the odds ratio for each question computed at the 50th percentile. In this chart, blue questions are those that are selected, yellow questions are those that are rejected on the basis of a low odds ratio at some percentile (not necessarily the 50th percentile, which is shown), and green questions are those that are rejected on the basis of having a confidence interval that is too wide.

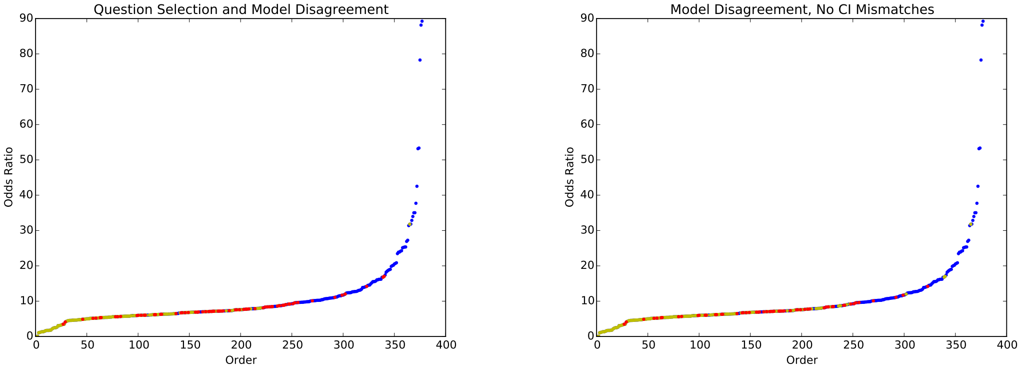
[10] If the item response function is logistic, then the difficulty corresponds to the value where the slope of the function is maximised, and the discrimination is the maximised slope

[11] We also fit a model with three parameters (an additional question-specific parameter for ease of guessing - the higher this value the easier the question is to guess). However, the three parameter model was rejected in favor of the two parameter model on the basis of ANOVA.



For the most part, we reject the questions with the lowest discrimination. In the chart above, you can see that there is a clear break around 225 on the x-axis where we reject most questions to the left (lower Odds Ratio) and accept those to the right (higher odds ratio). There are a few exceptions to this pattern: we sometimes reject questions with a very high Odds Ratio on the basis that its confidence interval is too high, and we sometimes keep a question with a very low Odds Ratio (which is examined in more detail later)

Now, we compare these results to the results of the IRT model. From the IRT model, we select the same number of questions as the intuitive approach in terms of the highest discrimination value. Of the 218 questions selected by each approach, 39 are selected differently. Of the 39 differences, 21 differences are due to the rejection of wide confidence intervals in the intuitive approach. Below, you can see mismatched questions in red. The chart on the left shows all mismatches, where the chart on the right only marks those mismatches that are not related to the confidence interval.



Most of the mismatches occur on the curve where the Odds Ratio is relatively flat, or where there is little difference in discrimination among the questions. These mismatches are relatively inconsequential: a small change in the measure of discrimination (the measure from the model vs. the measure from the intuitive approach, for example) may order these questions very differently, but because they are very similar in terms of their discriminatory power, any mismatches will have a relatively small impact.

Other mismatches are caused by the requirement in the intuitive approach that we be relatively certain about the difficulty of the question. Requiring a narrow confidence interval is very close to requiring more observations per question, which is why the intuitive approach rejects several questions which the model keeps, even questions which have a high discrimination. You can see this in the charts above: the rightmost red dots in the left chart, which shows all disagreements, are not red in the right chart, which does not show disagreements due to confidence interval rejections. Therefore, these questions were rejected due to having confidence intervals that were too wide. Although these questions appear to have a high degree of discrimination, without a sufficient number of answers to establish how difficult the question is we cannot be sure which group the question is discriminating. This would be problematic in our ultimate ranking of students who take the midterm.

There are a few questions which appear to have very low discriminatory power that the model rejects, but the intuitive approach does not. The leftmost red dot in both of the above charts is one such question. Although this question (question 13274) has low discriminatory power at the median and several other deciles, it actually has very high discriminatory power at the 90th percentile. Questions like this are desirable, so this mismatch is not concerning.

In the end, we use the intuitive approach to select the questions for the mid-term over the IRT model. We do this because the intuitive approach is more tailored to our needs (precision in question difficulty, and questions that are discriminating at many different percentiles). We also have reason to believe that the IRT model may be inappropriate: as discussed in 3, there is a positive relationship between the number of questions answered by a student and the number he answers correctly, suggesting that the question items may not be independent. Independence of question items is a key assumption of IRT models, and with this assumption violated we have less of a reason to trust the exact ordering produced by the model.[12]

[12] Furthermore, I am warned when the model is fit that the question ranking may be unstable. This is possibly due to the sparseness of the user / question response matrix: no students answer all of the questions and usually very few, so the matrix of 13,000 students and 400 questions will be mostly missing values. This may also contribute to how long the model takes to fit: with sparse

In summary, there are a few differences between the intuitive approach and an IRT model in terms of which questions are selected. However, many of these differences occur on questions with similar discriminatory power, and other differences result from the requirement of the intuitive approach that questions have narrow confidence intervals in terms of difficulty. Furthermore, there may be reason to believe that the assumptions underlying the IRT model are violated.

6 Conclusion

We select, using an intuitive approach, 218 questions that are able to discriminate students in terms of their knowledge. We select questions with high Odds Ratios at various percentiles of student performance in the data that have sufficiently narrow confidence intervals in terms of question difficulty. Questions selected on this basis will allow us to sort the students taking the midterm in terms of their knowledge of Astrometrics: we will be sure that each question asked is highly discriminatory; we will be able to tell with precision what we are discriminating with each question; and we will have highly discriminating questions at every decile of student performance.

In addition to those mentioned elsewhere in this report, there are a number of possible extensions to this analysis that could increase the ultimate ranking of students taking the mid-term.

First, it would be interesting to see if a three parameter IRT model with a fixed guessing parameter performs any better than the other models considered. There is evidence in the data that the students can guess correctly about 20 percent of the time, so it may be worthwhile including this observation in the model.[13]

Second, it might be useful to score the results from the IRT models in a way other than selecting those questions with the maximum estimated discrimination parameter. As discussed, we also care about the difficulty that those questions discriminate on. It might be better, for example, to select the questions of maximum discrimination at different levels of difficulty (ideally the difficulties we select at would in some way correspond with the underlying distribution of students, analogous to the use of student score deciles in the intuitive approach). It would then be interesting to see if that model agreed more or less with the intuitive model in terms of the questions selected.

Third, there may be a more precise way to reject questions than the exact confi-

data it may have difficulty converging.

[13] As further detailed in the next footnote, these models took several hours to run on my machine, so I was unfortunately limited in the number I could try.

dence interval being above a certain width. The confidence interval has the effect of rejecting questions with very few observations, with no consideration to how well the question discriminates. We could instead, for example, test the individual likelihood ratios using a Chi-square test to see if they are significantly different from zero, and keep those that are. This method may have its own issues, but it would be interesting to see how the results change.

Fourth, it would be interesting to examine whether any questions included in the final set are very similar in terms of their ability to discriminate students at a given difficulty. If a pair of questions were very similar in this regard, then including both in the final set would not add much to the final ranking. Furthermore, keeping both questions increases the likelihood that some student will get both questions on his midterm exam, which would give us less information about that student's knowledge than a set of five unrelated questions. One could examine this by looking at questions that have very similar discrimination and difficulties, or (somewhat equivalently) looking at questions that are answered very similarly by sets of students.

Lastly, and most importantly, it would be informative to validate the chosen questions with the data, or to perform the analysis presented in this paper in a way that lends itself to validation. For example, I could randomly break the data into a training and test set, perform the analysis on the training set, and then validate the results on the test set by seeing how well the chosen questions predict student scores (possibly removing the influence of a few obviously bad questions). Using this methodology, I could compare the questions selected by the IRT model and the intuitive approach in terms of how well those questions predicted students' scores in the test set. By repeating this procedure many times, I could determine whether the model or the intuitive approach predicted student scores better, and use this model to select the final list of questions. Furthermore, this methodology lends itself to the comparison of the effectiveness of other approaches: if I wanted to asses the effect of any of the other extensions discussed earlier, or even some of the arbitrary components of the intuitive approach, I could see how well those variations performed relative to the original.[14]

These considerations aside, I am confident that these questions would create a midterm that would result in a meaningful ranking of students.

[14] Unfortunately, I was somewhat limited by computing power: If I selected a training proportion to be 70 percent of the data, then running the IRT model, which took about 5 hours on all of the data, would take roughly three hours per run. If I wanted to draw 100 training and test samples to get a good measure of which model performed better, this approach would likely take about 300 hours without any sensitivities on the IRT model. Given the prompt says the mid-term is only a few days away, this does not seem like a viable approach.

7 Materials Used

7.1 References

1. Data and assignment provided by Knewton.

2. Manual pages for R and Python Packages

3. Morisette, J. and S. Khorram (1998), ”Exact Binomial Confidence Interval for Proportions,” Photogrammetric Engineering and remote Sensing, April, pp. 281-283.

4. Rizopoulos, D. (2006), ”ltm: An R Package for Latent Variable Modeling and Item Response Theory Analysis,” Journal of Statistical Software, Volume 17, Issue 5.

7.2 Python and R packages Used

- R

  1. ltm and dependencies.

- Python

  1. MatPlotLib
  2. OS
  3. Pandas
  4. Math
  5. Numpy
  6. SciPy
  7. Statsmodels