FEWER QUESTIONS, SHARPER INSIGHTS

Using Item Response Theory to Improve Evaluation of College Access Programs

Holden Ellis, Travis Candieas

ABSTRACT

College preparation for high school students has been a major component of higher education policy research for several decades. Common college preparation program services include academic advising, presentations, assistance with financial aid, and support during the application process. However, previous research focuses heavily on self evaluation of student attitudes towards college and attempting to extract behaviors associated with college-going behavior, leaving student knowledge about higher education underexplored. The present study focuses on assessing student knowledge about college application requirements by randomly assigning students multiple choice questions from a bank of test questions designed by professional college counseling experts. These questions were written to measure student knowledge about a specific facet of college preparation, such as financial aid. This study leverages Item Response Theory (IRT), which links a continuous latent ability of knowledge to a probability of answering a question correctly based on the difficulty, discrimination, and guessability of the question. After estimating IRT models for the latent knowledge score, the procedure is continued by analyzing information curves and parameters. Findings support the removal of some questions, development of new questions, and stratified sampling for question assignment.

METHODS

This analysis uses quiz questions data from EAOP's *College, Making it Happen* survey (n=1114), which has 11 multiple choice questions with 4 options and 2 true or false questions. The full quiz being studied is shown in the section below. For each student, two questions were assigned from `qz_piq_num`, `qz_he_camp`, `qz_tf_ag`, `qz_csu_gpa`, `qz_uc_gpa`, and `qz_tf_relig` and one question was assigned from the remaining options. Responses were analyzed using a 3 parameter model with the guessing parameter restricted.[1]

$$P(x_i = 1|z_n) = c_i + (1 - c_i)g(\alpha_i(z_m - \beta_i))$$

In this model, x_i=1 is a correct answer, z_n is the latent ability of a respondent, which is assumed to follow a standard normal distribution. c_i is the probability that a respondent guesses the correct answer to a question which is 0.5 for the true/false questions and 0.25 for the others. a_i is the discrimination parameter for each question and β_i is the difficulty. g is the logit function that maps from the real numbers to a probability. The model is fit using iterative a expectation maximization procedure.

Analysis begins with checking the unidimensionality assumption required for IRT models. Due high missingness, CFA methods for testing the assumptions fail, so instead an IRT model with 2 latent factors is fit and compared against the unidimensional model.[2] Once unidimensionality is confirmed, we can evaluate the model estimates. A useful visualization of these parameters is the Item Information Curve (IIC). This is the derivative of the Item Characteristic Curve, which represents the probability of a correct answer at each ability level. Peaks in IIC can be interpreted as the latent variable level that a question, or the test as a whole, best discriminates around.

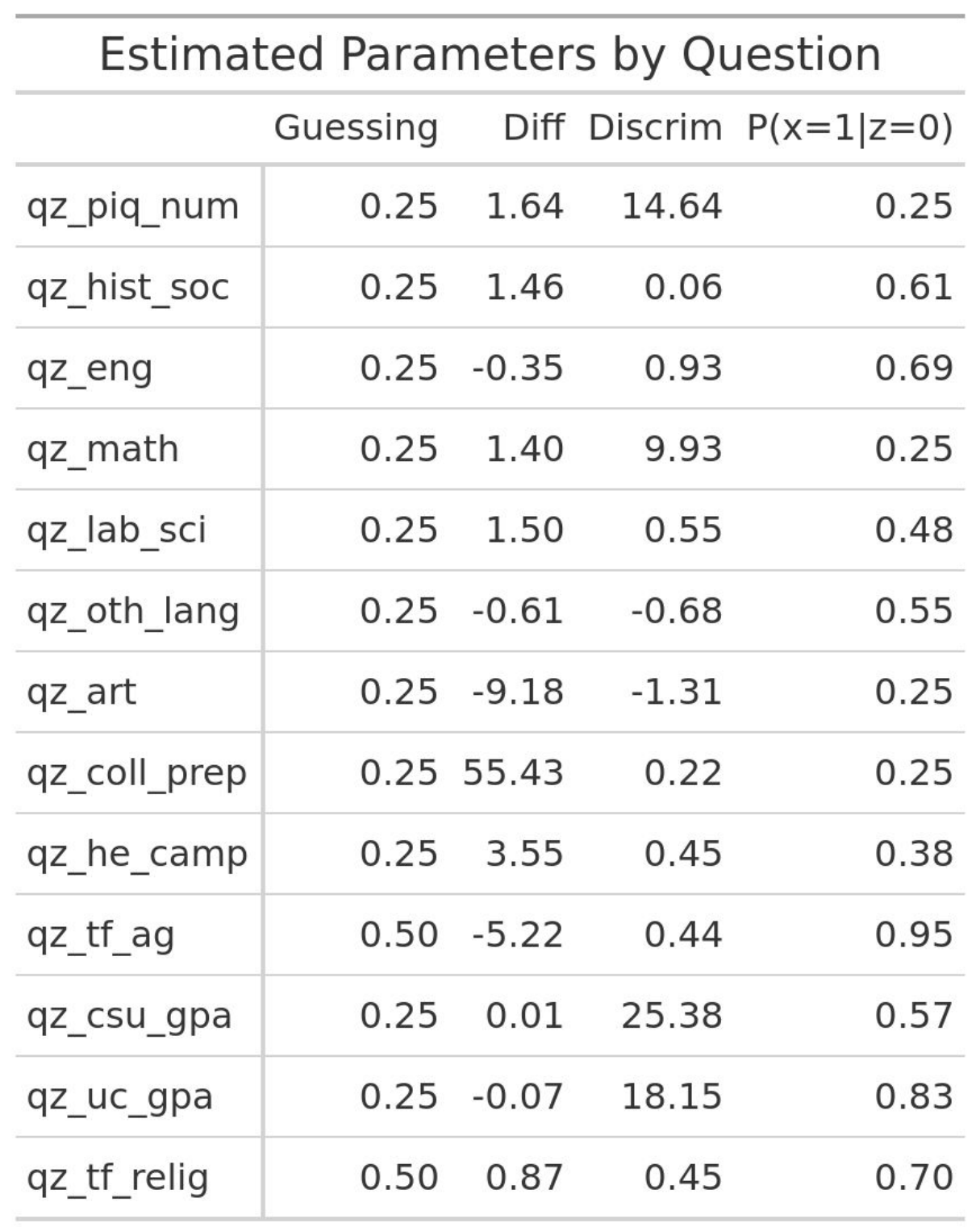
All analysis is done to decide how best to balance two goals:

1. **The test should reliably discriminate between different levels of knowledge**
2. **The test should use as few questions as possible**

The first goal is accomplished by selecting questions that have high discrimination parameters, but also selecting for questions with appropriate difficulty. For example, it does not help to measure a 3rd grader's math ability for by making them solve integrals even though these questions may discriminate for math knowledge very well.[3] One solution to this problem would be to ask all the questions we have, which would give a very accurate measurement. Test takers don't have time for this, which is why the second goal must be considered.

RESULTS

Comparing the 2-factor model with the 1-factor model yielded a p-value of 0.993, indicating that a second factor does not explain a significant amount of variance in the responses and, therefore, the **unidimensionality assumption is met** and the standard IRT model can be used to estimate the question parameters (Table) and the test information function (Figure 1).



Estimated Parameters by Question

	Guessing	Diff	Discrim	P(x=1|z=0)
qz_piq_num	0.25	1.64	14.64	0.25
qz_hist_soc	0.25	1.46	0.06	0.61
qz_eng	0.25	-0.35	0.93	0.69
qz_math	0.25	1.40	9.93	0.25
qz_lab_sci	0.25	1.50	0.55	0.48
qz_oth_lang	0.25	-0.61	-0.68	0.55
qz_art	0.25	-9.18	-1.31	0.25
qz_coll_prep	0.25	55.43	0.22	0.25
qz_he_camp	0.25	3.55	0.45	0.38
qz_tf_ag	0.50	-5.22	0.44	0.95
qz_csu_gpa	0.25	0.01	25.38	0.57
qz_uc_gpa	0.25	-0.07	18.15	0.83
qz_tf_relig	0.50	0.87	0.45	0.70

Table 1

The test information function shows that the test is **very effective at measuring knowledge for people near or above the average knowledge** level, with approximately 97% of measurable information being for individuals within 1 standard deviation of the mean or higher. This indicates that the test is ineffective at measuring the difference between below average and far below average individuals. It also has specific areas where questions discriminate very well. The reason for this can be seen in more detail in Figure 3, where the individual information curves for `qz_uc_gpa`, `qz_csu_gpa`, `qz_math`, and `qz_piq_num` are very effective in specific region. Figure 2 shows the rest of the questions, for which some have good information and discrimination (i.e. the question regarding the English requirement) and some are far worse (i.e. the question about college preparatory electives). Overall, this is reflected in the estimated parameter tables.

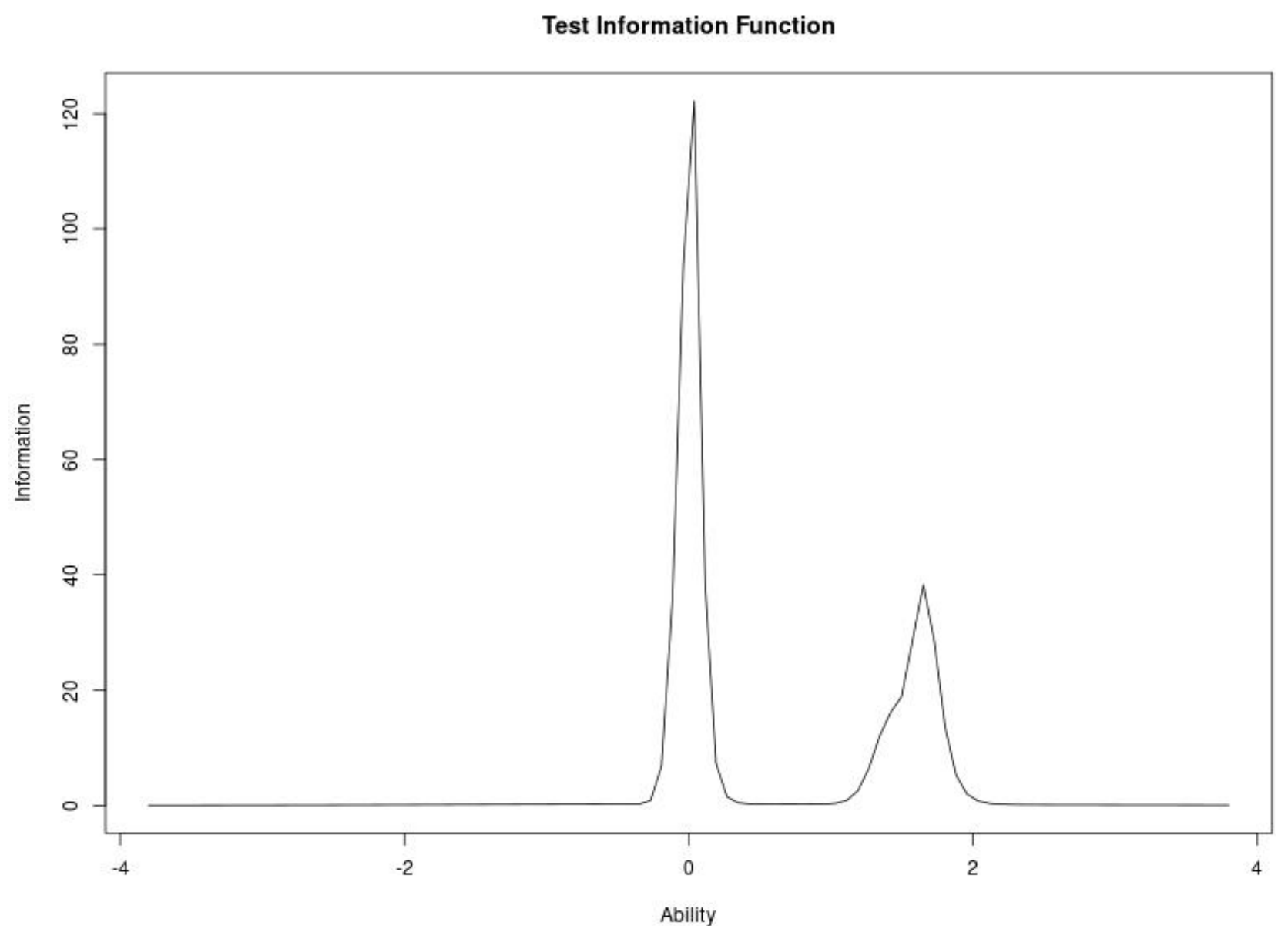


Figure 1, Test Information

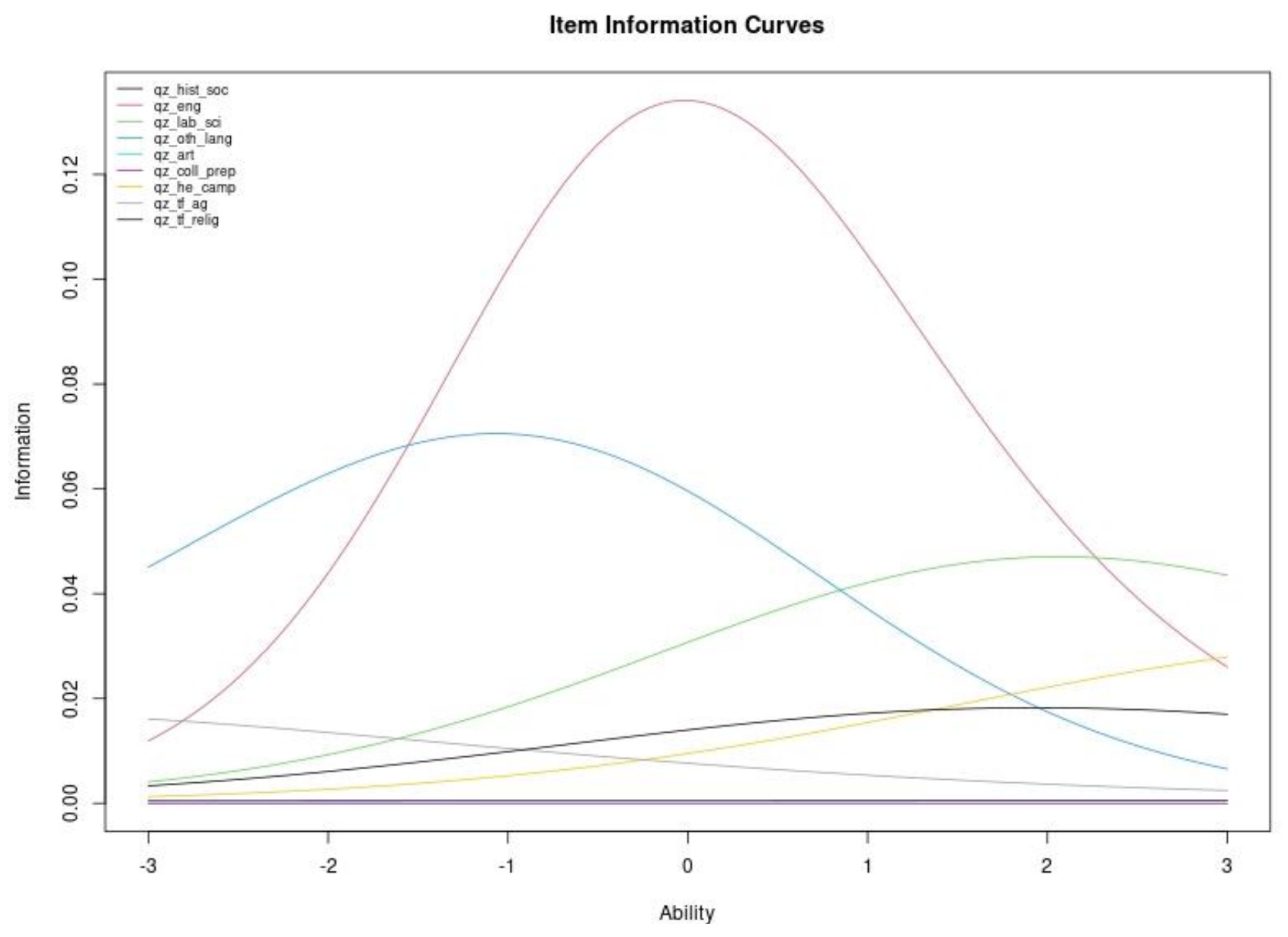


Figure 2, IIC

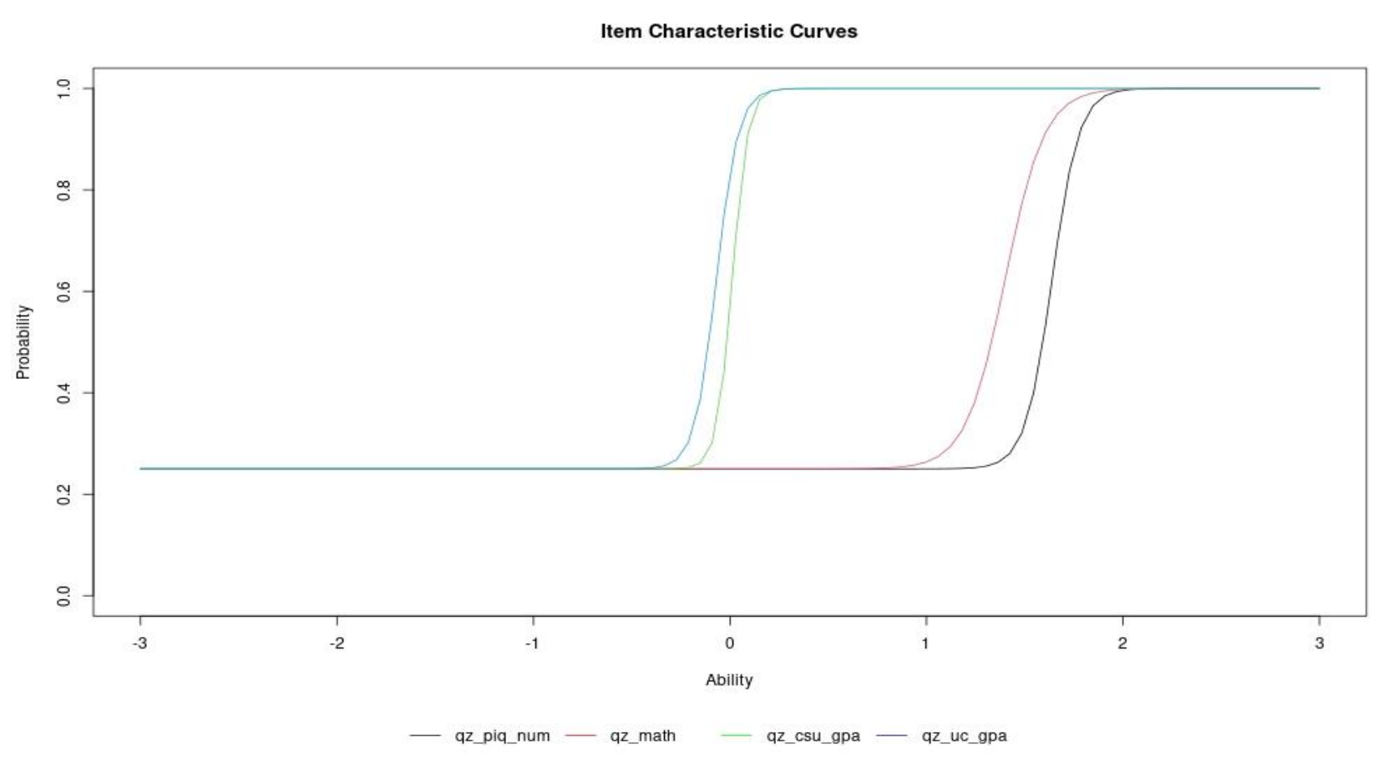


Figure 3, ICC

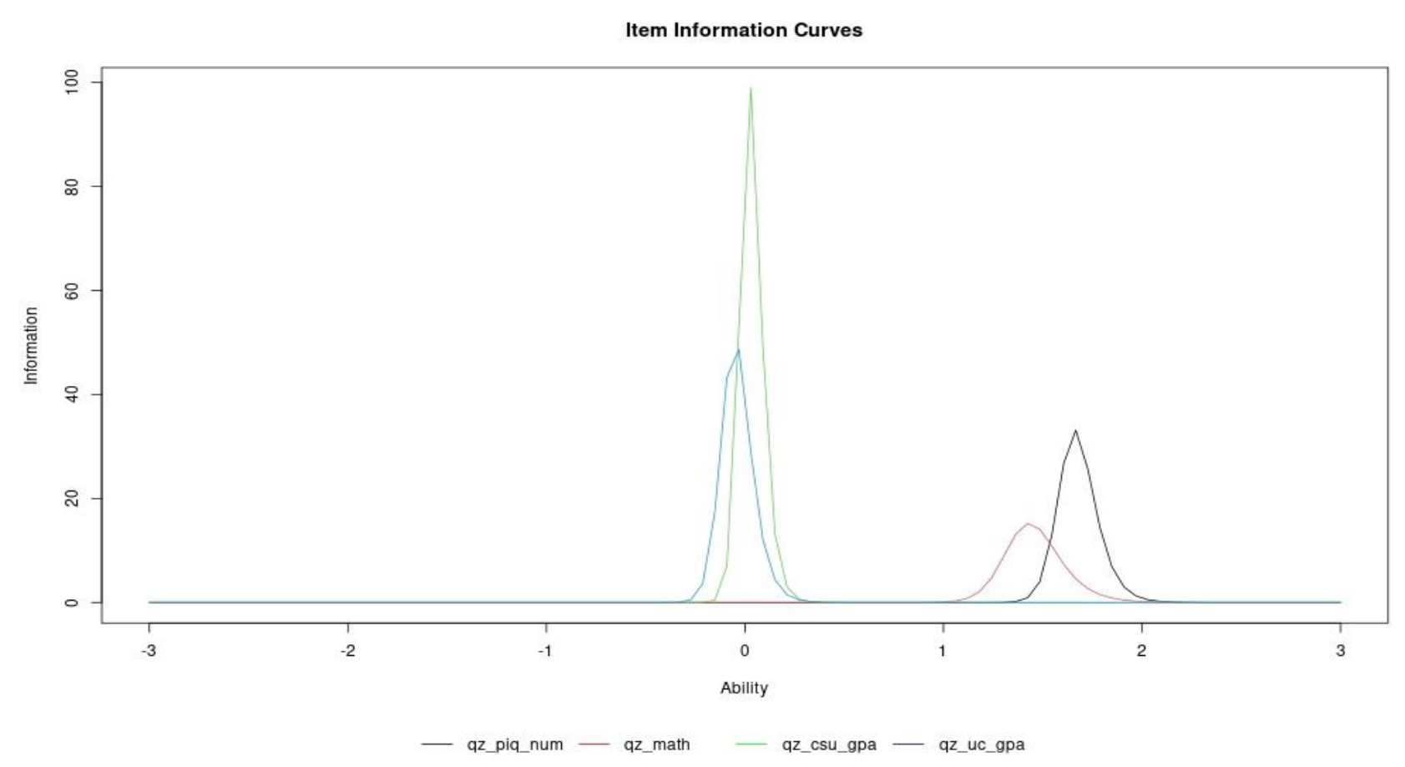


Figure 4, IIC

THE QUIZ

Students answered 53.6% of questions correctly.
Can you do better?

RECOMMENDATIONS

Modifications to Questions

Many questions need to be removed or significantly changed. Recommendations for this include but are not limited to:

- `qz_tf_ag` is far too easy with the average student answering it correctly 95% of the time. This isn't necessarily bad, but it also discriminates very poorly between below and above-average preparedness. While it could be removed from the quiz, changing it to a 4-option multiple choice would reduce the success of guessing and improve discrimination.
- `qz_coll_prep` should also be removed or modified since it is drastically harder than any other question in the quiz, and of the information it provides, only 2.5% is for students within 2 standard deviations of the mean preparation.
- `qz_art` has exactly 25% correct answers and its lack of discrimination indicates that this is because most students are just guessing. This question should be removed.
- `qz_he_camp` is a poorly written question. It could be argued that "private institutions" have more campuses than community colleges with a broad enough definition of a higher education institution.

Changing Question Assignment Procedure

It is also recommended that the procedure for assigning questions should be changed. One option is a stratified sampling procedure. This method divides the question bank into easy, medium, and hard buckets based on the estimated difficulty parameter and each students should be randomly assigned 3 questions, one from each bucket. This would be make scores more reflective of knowledge and would also prevent circumstances where students are asked two very similar questions in the same survey (i.e. `qz_csu_gpa` and `qz_uc_gpa`).

One issue with this procedure is that it would discriminate poorly in-and-around the mean since the harder questions only discriminate at more than 1.5 standard deviations above the mean knowledge. If the goal of testing is to differentiate between average students and the difference between a slightly and far below average student is not worth measuring, the best design would be to sample randomly from the questions with difficulties near 0.

Computerized Adaptive Testing

The best solution would be adopting a Computerized Adaptive Testing (CAT) method. This method, which has recently gained popularity for standardized testing, has the simple premise of asking harder questions after a respondent is correct and easier questions after they are incorrect.[3] For example, a CAT version of this quiz would start by asking respondents `qz_csu_gpa` to discriminate between above and below average. If they get this question correct, `qz_math` could be asked to narrow the estimate for z_i being in either [0,1) or [1,∞). Unfortunately, implementing a CAT system is difficult because the test administration software needs to be much more complex than what EAOP currently uses and all questions need to have an accurately measured difficulty score for results to have any meaning.

References

[1] Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. Journal of Statistical Software, 17(5), 1–25. https://doi.org/10.18637/jss.v017.i05

[2] Phil Chalmers (2021). mirt: A Multidimensional Item Response Theory Package for the R Environment. Version 1.44.0. https://doi.org/10.18637/jss.v048.i06

[3] Bock, R. D., & Gibbons, R. D. (2021). Item response theory (First edition.). John Wiley & Sons, Incorporated.