Latent Trait Models Highlight Deficits in Student Understanding

Robert Quinn, Petra Graham, Anne Karpin, Ayse Aysin Bilgin

Department of Statistics, Macquarie University, Australia

Aim

Final exams are typically set in order to assess course content knowledge and to provide evidence that students have achieved at least some minimum level of competence in the learning outcomes. Exam papers are typically archived on completion but they contain abundant information that can highlight topics where there is either adequacy or a shortfall in understanding.

Final exam papers from a first-year business statistics unit were randomly selected. Using item response theory models, the probability of a correct response to each of 56 question items was obtained as a function of item difficult and student ability. Item difficulties were extracted to enable the ranking of question items from least to most difficult. Results of modelling and the impact on future teaching will be presented.

Methods

A sample of 250 final exam papers was systematically selected from a very large cohort of 2017 first year introductory statistics students. The exam comprised 7 short answer inference questions that could be broken down into 56 binary items identifiable as either correct or incorrect according to the marking guide.

Item response theory (IRT) models including the constrained Rasch model (whereby all items are assumed to have the same discrimination), the unconstrained Rasch model and a latent trait model were tested. The best model was chosen as the one with the smallest AIC.

Item characteristic curves (ICC) are used to display item difficulty as a function of student ability and item information curves (IIC) indicate the discriminatory ability of each item.

The *LRT* package for R[1] was used for all analyses.

Results

Questions were categorised into broad concepts which are shown in Table 1. Table 1 indicates that students are struggling most with conclusion explanations and least with stating hypotheses and checking assumptions.

The exam questions focus on problems relating to tests of means for one-sample (Q1) and two-samples (Q2), tests of proportions (Q3), chi-squared goodness-of-fit tests (Q4), chi-squared tests of independence (Q5), paired t-tests (Q6) and regression (Q7).

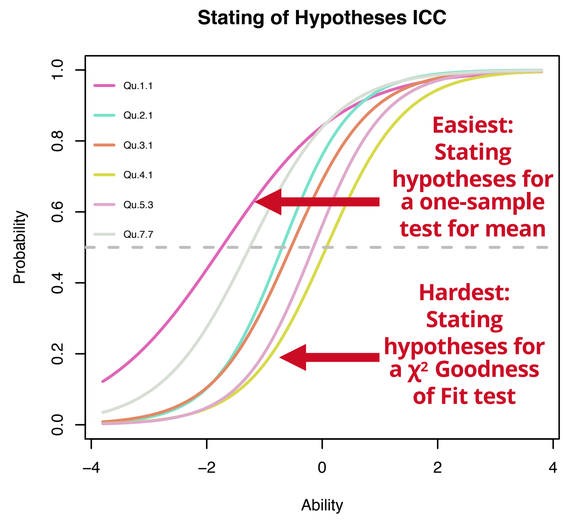
Question 7.6 was answered correctly by only **9%** of students, whilst **Question 4.6** was answered correctly by **89%** of students.

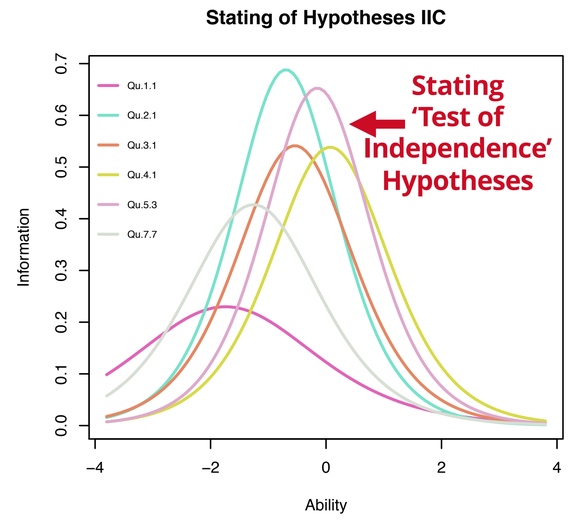
The **IRT Latent Trait Model** was selected as the best fitting model for this particular examination, due to holding the lowest AIC (12,864).

Table 1: *The category of each item assessed in the exam, coloured by the percentage of successful students.*

Question Types	Items
Stating of Hypotheses	1.1, 2.1, 3.1, 4.1, 5.3, 7.7
Test Assumptions	1.2, 2.2, 2.3, 3.2, 4.3, 5.4, 7.1, 7.2, 7.3
Calculation of Test Statistics / DF / Intervals	1.3, 1.4, 2.4, 2.5, 3.3, 4.4, 4.5, 5.5, 5.6, 6.1, 6.2, 7.8, 7.9
P-Value / Decision	1.5, 1.6, 2.6, 2.7, 3.4, 3.5, 4.6, 5.7, 7.10
Conclusion	1.7, 1.8, 2.8, 3.6, 4.7, 4.8, 5.8, 6.3, 6.4, 7.6, 7.11, 7.12
Other*	4.2, 5.1, 5.2, 7.4, 7.5, 7.13, 7.14

Key: **≤ 25% Correct** **25-50% Correct** **50-75% Correct** **≥75% Correct**

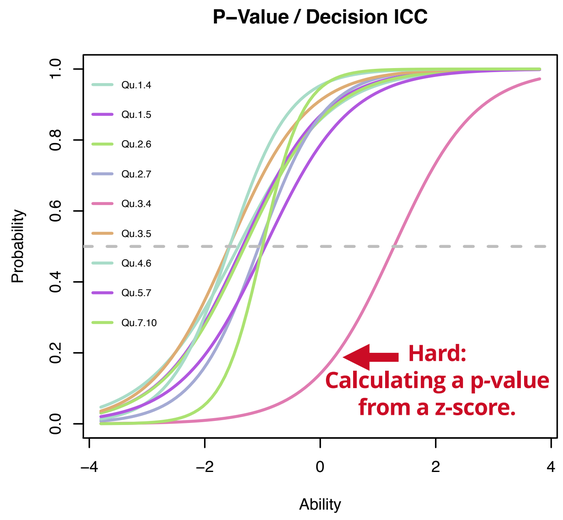


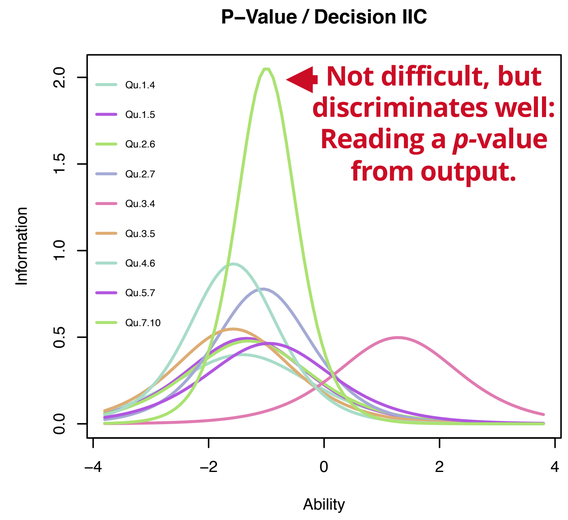


Stating of Hypotheses

Most of these questions are 'easy' (The ICC shows the ability required to have a 50% change of getting the answer correct is 0 or less for all questions.)

The IIC shows relatively high peaks indicating good discrimination for all questions, except 1.1, which appears to be too easy and does not discriminate particularly well.

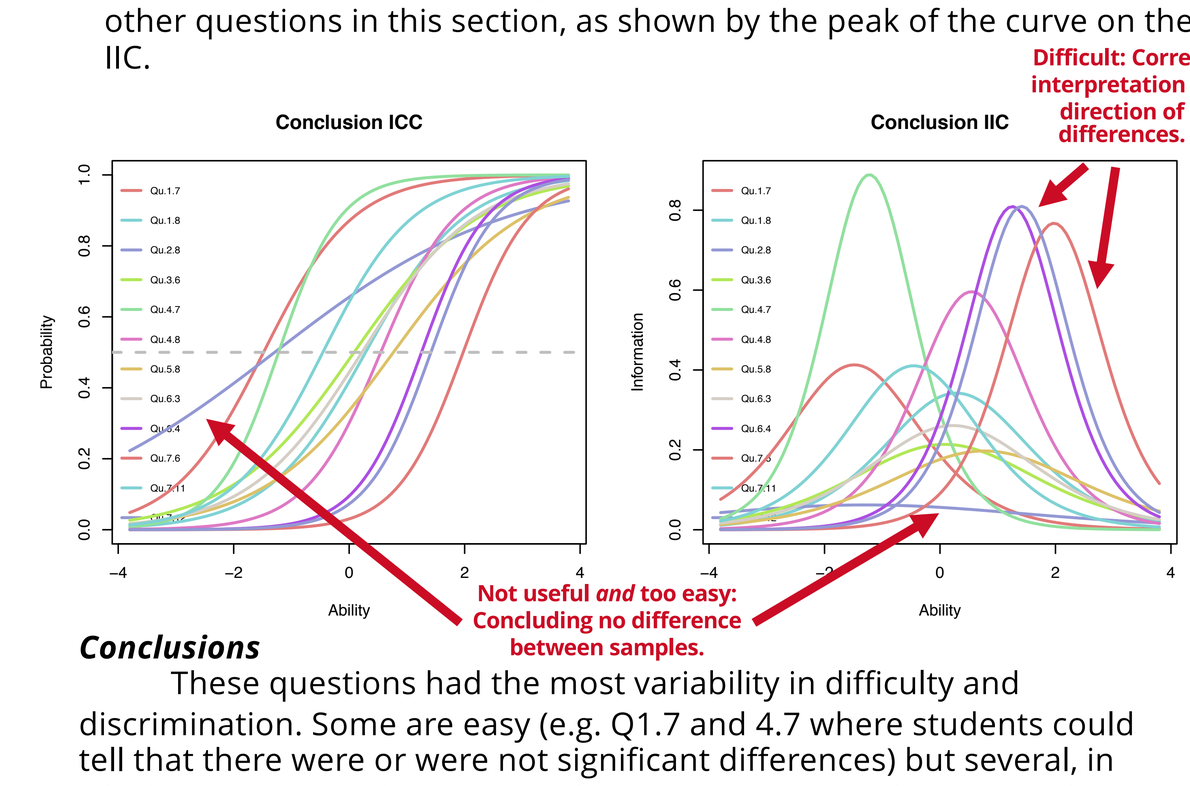




P-value/Decision

These are mostly easy, except for Q3.4 which appears to be extremely difficult; weaker students will not get this right.

Q7.10 is not particularly difficult, but discriminates better than the other questions in this section, as shown by the peak of the curve on the IIC.



Conclusions

These questions had the most variability in difficulty and discrimination. Some are easy (e.g. Q1.7 and 4.7 where students could tell that there were or were not significant differences) but several, in which more comprehensive conclusions were required, were much more difficult (e.g. Q6.4, 7.6, 7.12).

Conclusions and Future Actions

In response to our findings we

- Introduced mandatory practice quizzes to encourage formative learning on the key topics that students struggled with.
- Plan to change the way we teach the interpretation and writing of conclusions, making these parts more interactive and providing useful summaries of techniques.
- Plan to be more careful with respect to assessing material that contains assumed knowledge not necessarily carefully covered in class.

MACQUARIE University
SYDNEY·AUSTRALIA

[1] D. Rizopoulos, "ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses", *Journal of Statistical Software*, vol. 17, no. 5, 2006.