Item Response Theory Analysis

IRT was developed in the 1940s and 1950s by multiple statisticians independently.

- Georg Rasch is credited with developing a statistical model to handle dichotomous decisions.
- Hua-Hua Chang (2018 FREMO conference, Oslo, Norway) has indicated that the decision to use the 50-50 point as the decision point arose from biological testing of drugs to determine their LD50--that is, the lethal dose at which 50% of animals expired.
- Frederic Lord at ETS developed in the 1950s and 1960s much of IRT as implemented in various tests including the SAT, GRE, GMAT, LSAT and the TOEFL.[1][2]

IRT attempts to model factors that explain the probability of a person with a certain ability getting an item of a certain difficulty correct. Most commonly IRT models the item difficulty (the point b at which there is a 50% chance of getting the item right), the item discrimination (the slope a of the item characteristic curve at point b), and the pseudo-chance probability of getting an item right when person ability is very low (the intercept c of the item characteristic curve when point b is at -3.00). Of course, there are many other sources of error (e.g., individual and contextual factors independent of the test and its administration) which are not accounted for in these models.

Nonetheless, person scores using IRT are dependent on how hard the items were that were answered correctly and how effectively they discriminate among people with lesser or greater knowledge. This is quite a difference to the classical model of simply summing all items that people got correct.

This open access paper provides explanation of the various IRT models and shows what happens to a test when the models are applied.

Brown, G. T. L., & Abdulnabi, H. (2017). Evaluating the quality of higher education instructor-constructed multiple-choice tests: Impact on student grades. *Frontiers in Education*, *2*(24). d ⓘ).3389/feduc.2017.00024 https://www.frontiersin.org/articles/10.3389/feduc.2017.00024/full

I have provided notes on figshare.com for learning how to do IRT making use of the data and R software.

Go to: https://doi.org/10.17608/k6.auckland.c.4198712.v1

There you will find 6 different presentations explaining and contrasting IRT with CTT. Also there is a data set you can use to practice analysis and a tutorial guide that will show you step-by-step how to analyse the test data using the free software R.

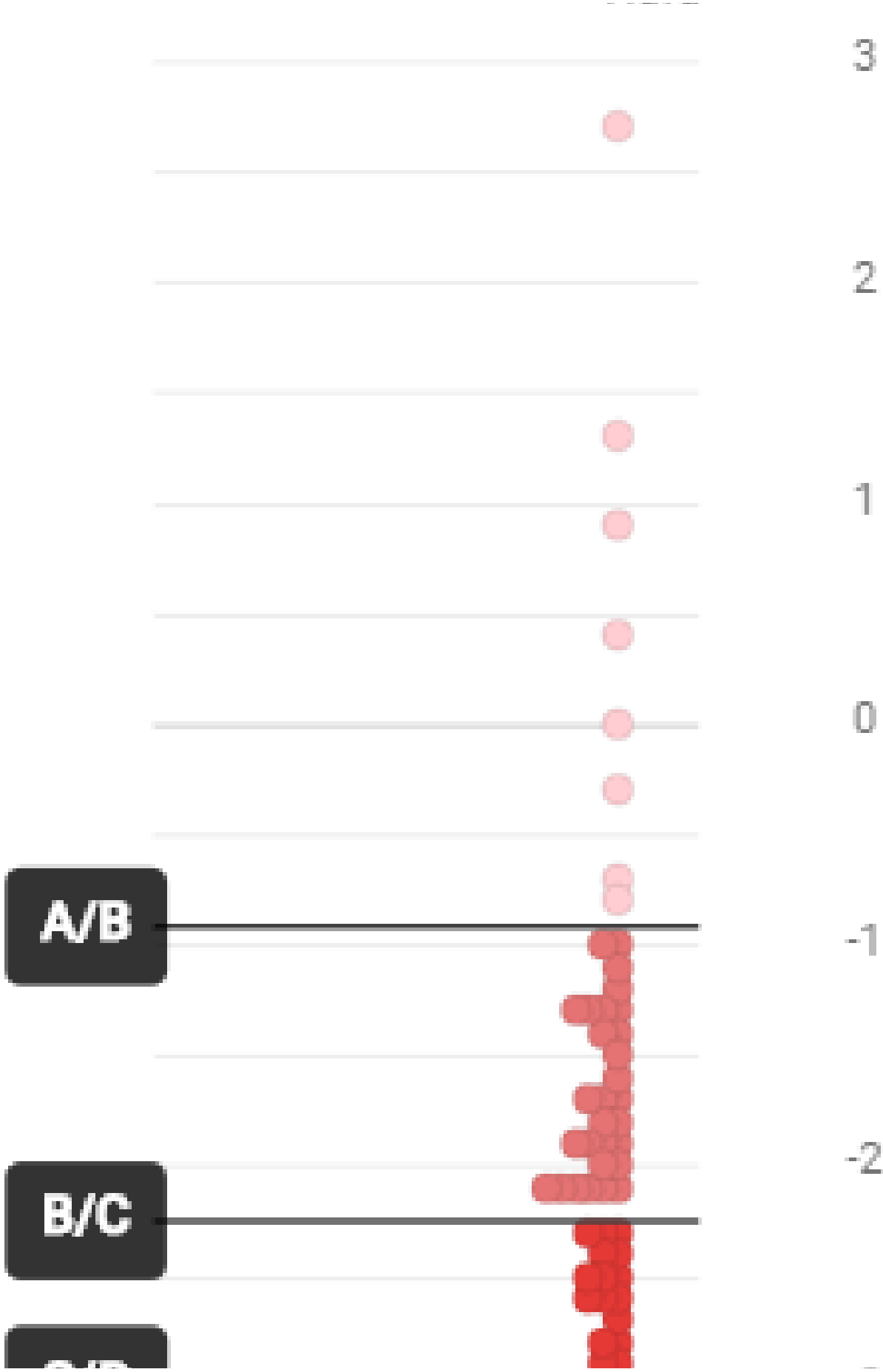
SmartStandardSet

Item Response Theory can be used to identify poorly performing items (as is the case in the article by Brown & Abdulnabi). Additionally, because it weights scores according to the item characteristics, rather than the sum of items correct, it can be used to adjust student scores and to inform grade boundary setting. A project at the University of Auckland (entitled *SmartStandardSet*) is developing a simple software to analyse multiple-choice tests and allow test administrators the opportunity to set grade boundaries based on the competencies implied by the test item content. The following slides illustrate this methodology.

54	99.3	1.84	-3.47	✓
25	98.5	2.3	-2.75	✓
62	97.6	2.3	-2.5	✓
60	97.3	0.83	-4.66	✓
42	96.9	1.29	-3.23	✓
39	96.4	2.22	-2.3	✓
31	96	2.37	-2.18	✓

IRT Analysis of Items

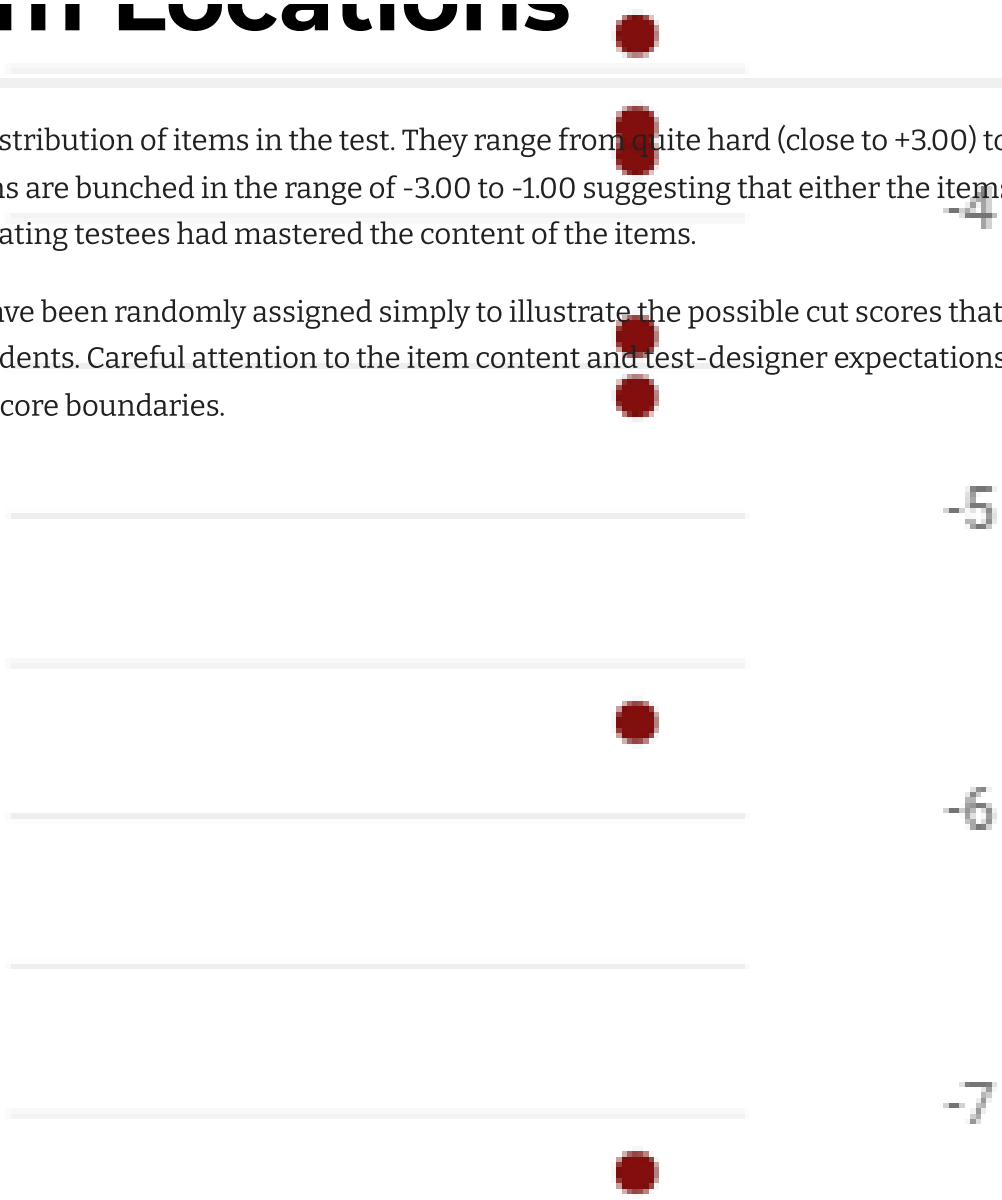
This screen shot shows the easiest items in a test. The % correct column shows the proportion of test takers who got the item right. The Discrimination column shows how steeply the ICC runs through the Item Weight (difficulty point b) point of the item. Negative weights indicate the item is easy and discrimination values close to zero show that an item does not separate well. Items that have negative discrimination would be shown with a red cross. Only when the % correct is exactly the same will items have identical slopes and discrimination.

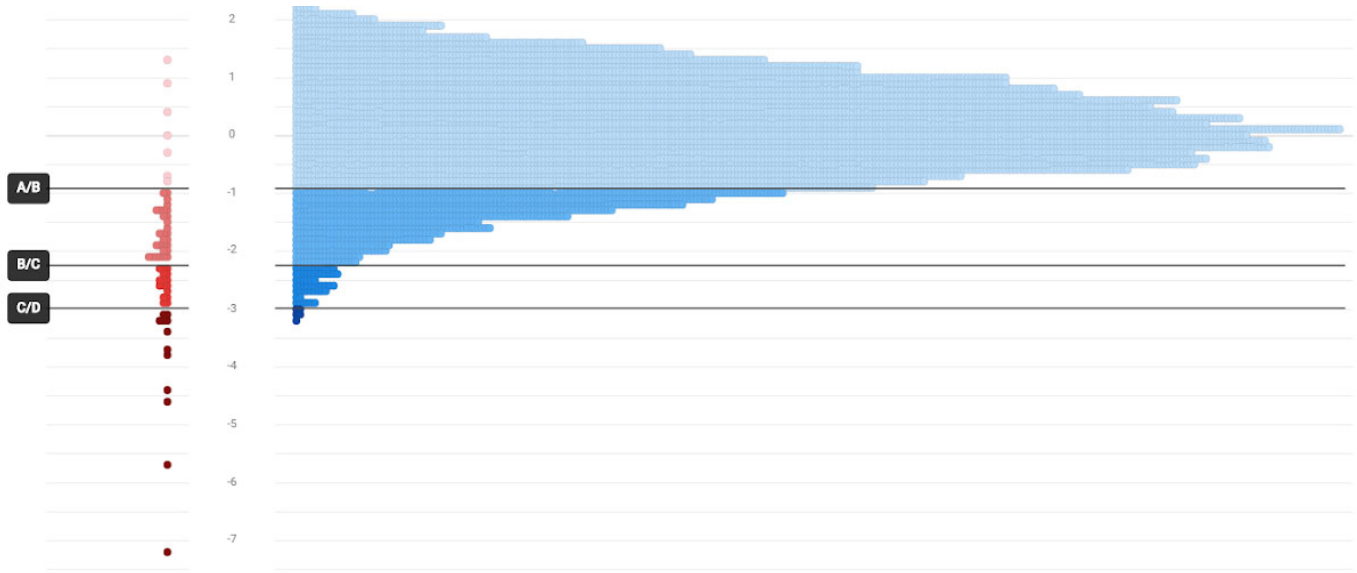

3
2
1
0
-1
-2
A/B
B/C

IRT Item Locations

This chart shows the distribution of items in the test. They range from quite hard (close to +3.00) to very easy (<-7.00). Most items are bunched in the range of -3.00 to -1.00 suggesting that either the items were too easy or the participating testees had mastered the content of the items.

The grade standards have been randomly assigned simply to illustrate the possible cut scores that could be used to separate students. Careful attention to the item content and test-designer expectations has to be given to set proper score boundaries.





Standards and Norms

This chart displays the performance of test takers relative to the items.

- It should be noted that almost no people were challenged the by few items below -4.00 suggesting that those items are wasted and their removal would make for a more efficient test.
- On the other hand, very few items were being used to evaluate the competence of the top third to half of all test takers. If decisions had to be made to separate those people, then more items would be needed there.
- On the other hand, if this were a mastery test (e.g., minimum competency to be safe as a learner driver) then it could be argued that any test take with a score >-1.00 clearly is satisfactorily competent.
- Only a person expert in the content of the test can determine where the pass-fail line or any grade standards should be drawn.