An Analysis of Multiple Choice Item Quality

Melissa Van Bussel, Kara Fitze, Wesley S. Burr and Aaron Slepkov

Background Information

Post-secondary institutions frequently use Multiple-Choice (MC) items on classroom tests and examinations, with such assessments often constituting a large portion of students' final grades. The MC format allows for faster and easier marking for professors, especially with growing enrollment rates and class sizes. Given that there is such a wide use of MC testing in postsecondary settings, it is valuable to consider the quality of MC items (questions) on such tests. In this work we examine the difficulty indices of tests, their overall discriminatory power, and the effectiveness of each MC item's distractors.

Data Recovery

- Fifty-two different MC tests and examinations from a variety of departments at Trent University were collected
- The sample was composed of midterm tests and final examinations at the Undergraduate level, that took place between 2012 and 2016, and ranged in length from 17 to 100 questions
- There was no prior certainty on the length of each item, although it appeared that the majority of MC items were 4 or 5 options long

Definitions

Multiple Choice Vocabulary

Item: An individual multiple choice question
Keyed option: The correct answer for a given question
Distractors: Remaining options that are not the keyed option

Characteristics of Multiple Choice Items

Difficulty Index (*p*): The proportion of examinees who select the keyed option for an individual item. Restricted to values ranging from 0 (no examinee selects the keyed option) to 1 (all examinees select the keyed option).

Distractor Efficiency: A measure of the test's ability to produce *functional* distractors. Functional distractors can be defined subjectively, however, the most widely accepted definition is that at least five percent of examinees select an option for it to be considered functional, and the distractors must be negatively correlated with overall test scores.

Discriminatory Power: The degree to which more knowledgeable students select the keyed option over less knowledgeable students. This is measured by an item's discrimination coefficient, denoted r_{pb}. The discrimination coefficient reports a point-biserial correlation between examinees' overall test scores and the scores they obtain on a particular item under consideration.

Item-included Correlation: The student's overall test score correlated with their performance (0 or 1) on the item under consideration.

Item-excluded Correlation: A student's test score without the given item, correlated with the score they obtained on the item.

Methods

- We created a robust (error-checking) R function to take a single test file as its input, perform an analysis of various test parameters, and return a data frame of the parameter values.
- This data frame includes measures of the mean difficulty index, values of the mean item-included correlation, and mean item-excluded correlation, mean distractor efficiency and Cronbach's α.
- Functional distractors were defined as non-keyed options that were selected by at least 5% of the examinees

p_mean	p_sd	p_max	p_min	scores_mean	scores_sd	r_mean	r_sd	r_prime_mean	r_prime_sd
0.671	0.197	0.972	0.225	0.671	0.129	0.294	0.147	0.208	0.146
0.468	0.244	0.959	0.162	0.468	0.097	0.222	0.135	0.098	0.140
0.640	0.235	0.966	0.150	0.640	0.122	0.294	0.133	0.236	0.137
0.638	0.190	0.954	0.215	0.638	0.144	0.332	0.105	0.295	0.107
0.610	0.226	0.984	0.053	0.610	0.147	0.331	0.129	0.224	0.124

Table 1: Sample of the data returned from the function call. Here **p_mean** represents the mean item difficulty**, p_min** and **p_max** are the minimum and maximum difficulty indices per test respectively, **r_mean** stands for the mean item-included discrimination coefficient*, and* **r_prime_mean** is the average item-excluded discrimination coefficient.

Results

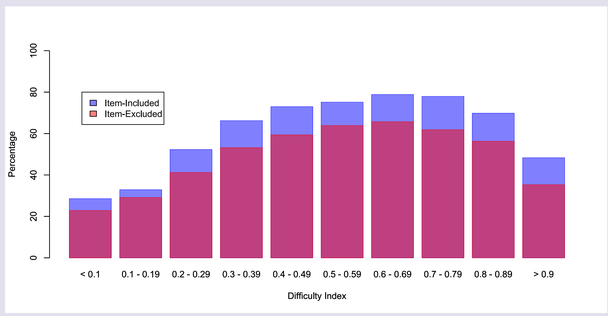

Figure 1: Percentage of Items with a Discrimination Coefficient of +0.20 or Greater as a Function of Item Difficulty

For items with a difficulty index smaller than 0.30, less than half of the items met or exceeded the general accepted value for the item-included discrimination coefficient (+0.20). Similarly, for items with a difficulty index less than 0.20, fewer than 50% of the items met or exceeded this 0.20 value (item-excluded coefficient). There is a pattern that the item-excluded correlation coefficient has a lower percentage of discrimination for given item difficulty values in comparison to the item-included correlation.

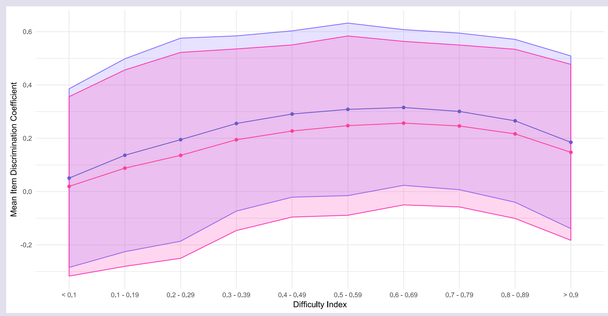

Figure 2: Mean Item Discrimination Coefficient as a Function of Item Difficulty

This plot gives a visual interpretation of the relationship between the difficulty index and the discrimination coefficient of every MC item in the data. It suggests that there may be linear and quadratic components of relation between both variables. The mean discrimination coefficient appears to be higher when the difficulty indices are between a range of abut 0.3 and 0.89, and then significantly lower for the really easy and really difficult items. The confidence bands clearly demonstrate that item-included discrimination coefficients are inflated.

It has long been known (Cureton, 1966) that the point-biserial correlation between an item and the test of which it is a part is spuriously inflated. One suggestion for a corrected average point-biserial estimate can be formed as Equation (1), with n the total number of items, S_t the standard deviation of scores on the total test, p the proportion passing item i, $q = 1 - p$, and y the *Z*-score for p.

$$\text{Equation (1): } r_{pb}{}' = \sqrt{\frac{n}{n-1}}\frac{r_{pb}\,S_t - \frac{pq}{y}}{\sqrt{S_t^2 - \sum p_i q_i}},$$

This work has been so-far been focused largely on demonstration of the spurious inflation of the correlation using practical real-world test examples, as despite the long literature indicating the issue, point-biserial (item-included) correlations are used to this day in the education and testing literature to demonstrate item quality (and to suggest overall test quality). Future work includes expansion of the testing database to include a wider range of tests, so as to provide a concrete, practical, Canadian undergraduate reference point for expected discrimination and level of spurious inflation.

Research Questions

- How does item difficulty affect the discriminatory power?
- What are the empirical differences between the results for item-included correlations and item-excluded correlations?

References and Acknowledgements

DiBattista, D., & Kurzawa, L. (2011). Examination of the Quality of Multiple-Choice Items on Classroom Tests. *Canadian Journal for the Scholarship of Teaching and Learning*, (2):2.

Curreton, E.E. (1966). Corrected Item-Test Correlations. *Psychometrica*, (31):1.

Melissa Van Bussel and Kara Fitze have been supported through the NSERC Undergraduate Student Research Awards program.