EXAMINING DISTRACTORS IN MULTIPLE-CHOICE QUESTIONS FOR CLASSROOM ASSESSMENTS: LEARNING FROM PSYCHOMETRICS

Yan Liu, Amery D. Wu, Minjeong Park, & Ernest Kwan
The University of British Columbia
Carleton University
yan.liu@ubc.ca

Psychometric research mainly involves the construction of tests/assessments and the development of techniques to ensure the quality of measurement. However, this field has been focusing more on large-scale tests/assessments. There has been less attention and efforts on classroom assessments. The present study is scoped in a broader goal of bridging psychometric research and classroom assessments. More specifically, we will introduce some psychometric techniques for instructors to examine the distractors in multiple-choice questions, including the most commonly used distractor analysis strategies, including a particular technique, differential options functioning (DOF) based on multinomial logistic regression. DOF is a versatile tool that can help to better understand students' misconceptions and learning gaps. A demonstration will be provided using an introductory course assessment.

INTRODUCTION

Multiple-choice questions (MCQ) or multiple-choice items have been widely used in educational testing, medical licensing examination, psychological tests, market research, etc. Nowadays, MCQs are still the most widely used assessment format for measuring students' knowledge and skills in K-12 and post-secondary education. The distractors/options are crucial, but they are often the neglected parts of MCQ. In educational assessments, good distractors can help to assess gaps in students' learning outcome. There is a variety of the analyses for evaluating the quality of distractors. In general, distractor analysis is used to provide insights into the effectiveness of the distractors and student' learning on the topics taught in the class.

The purpose of the present study is to demonstrate several useful methods of distractor analysis and focus on a particular technique, differential options functioning (DOF) proposed by Park and Wu (2017). This method aims to detect certain groups' possible misinterpretation and/or misunderstanding of the distractors that may lead to groups' response differences. This paper is organized in the following sequence: (1) a brief review on the commonly used methods for analyzing MCQ distractors, (2) a description of DOF method, (3) a demonstration of these methods, and (4) conclusion.

REVIEW ON DISTRACTOR ANALYSIS

The primary purposes of distractor analysis are to identify nonfunctioning and ill-functioning distractors as well as to improve the discrimination power of MCQs for distinguishing students with low ability from those with high ability. Three general approaches have been used or recommended for distractor analysis, classical test theory (CTT), item response theory (IRT) and cognitive diagnostic model (CDM). However, IRT and CDM are based on latent variable modeling techniques and require a large sample size and advanced psychometric training. Hence, IRT and CDM are not suitable for teachers and practitioners who would like to analyze classroom assessment data that are relatively small in size. Hence, we will only review the most commonly used strategies that suit class size data.

The commonly used CTT technique is to examine the percentage of students who choose each distractor. The percentage of students who choose each distractor can be used to detect low frequently selected distractors, which may be a candidate for "nonfunctioning distractor." Haladyna and Downing (1993) indicated that a distractor could be considered as a low frequency distractor if less than 5% of the students choose that distractor. The test developer can consider removing or modifying the distractor if the low frequency distractor is not written for a particular purpose (e.g., testing subtle misconception). An alternative way of examining the percentage of students is to use trace line plots (Wainer, 1989) to visualize the relationship between students' ability and the percentage of their selections of each option. Non-discriminating, nonfunctioning or ill-functioning

In M. A. Sorto, A. White, & L. Guyot (Eds.), *Looking back, looking forward. Proceedings of the Tenth International Conference on Teaching Statistics (ICOTS10, July, 2018), Kyoto, Japan.* Voorburg, The Netherlands: International Statistical Institute. iase-web.org

distractors can be easily identified from a trace plot. These two strategies are easy to implement by contingency table and graphics, so they are more popular than other methods.

DIFFERENTIAL OPTIONS FUNCTIONING

Differential options functioning (DOF) is based on multinomial logistic regression and does not assume any ordering among the options, nor does it require a sample size as large as IRT or CDM. This technique has been used for large scale assessment data for examining the MCQ distractors (Kato, Moen, & Thurlow, 2009; Park & Wu, 2017). In the context of achievement assessment, DOF can be used to help evaluate whether the effectiveness of the distractors remains to the same level between different groups of students who are equally capable (e.g., female vs. male) and whether the meaning of the distractors appear similar to the groups.

The procedures of DOF test include three nested sequential models, the model with the total test scores only (Model-1), the model with both the total test scores and the grouping variable (Model-2), and the model with an interaction of the total test scores and the grouping variable (Model-3). The models are specified as follows.

$$\text{Model 1}: \log \frac{P(Y=j \mid T)}{P(Y=k \mid T)} = a_j + b_1 T \quad (1)$$

$$\text{Model 2}: \log \frac{P(Y=j \mid T,G)}{P(Y=k \mid T,G)} = a_j + b_1 T + b_2 G \quad (2)$$

$$\text{Model 3}: \log \frac{P(Y=j \mid T,G)}{P(Y=k \mid T,G)} = a_j + b_1 T + b_2 G + b_3 (T * G) \quad (3)$$

where *j=1, 2, …, J* denotes the jth distractor of the item; *k* denotes the reference option (i.e., the correct answer); *T* is total test score; *G* is the grouping variable (e.g., female vs. male); *T***G* is the interaction of *G* and *T*.

Adopting this multinomial logistic regression analysis, the correct answer will be used as the reference and the probabilities of choosing the other options will be compared to the correct answer. The total score is treated as an approximate of student ability on the subject. Model-1 allows one to see the probability of endorsing a particular distractor (vs. the correct answer) after controlling for the ability. Model-2 tests whether there are group differences on selecting a particular distractor (vs. the correct answer) after controlling for the ability. Model-3 tests whether the direction and magnitude of the group differences vary depending on the level of student ability.

To assist our interpretation of the results, we will also use odds ratio, a type of effect size, and logistic characteristic curves in addition to the *p*-value of regression coefficients. As recommended by Cohen (1988), a small effect = 1.5, a medium effect = 3.5 and a large effect = 9. Usually an effect size of 2 or larger (0.5 or smaller) is considered as the threshold for a small effect in practice.

DEMONSTRATION OF DISTRACTOR ANALYSIS STRATEGIES

Sample & Measure

A total of 125 students responded to an introductory statistics assessment, Comprehensive Assessment of Outcomes in Statistics (CAOS) before taking a short statistics course. The students were graduate students in life sciences, 48% of female (*sex*; female=0, male=1), 72% of them had taken 1-2 introductory statistics course before the course and only 11% had not taken any statistics course before (*numCourses*;, 1 = none course, 2 = one course, 3 = two courses, 4 = three courses, 5 = four or more courses). *Sex* was used in the following demonstration as a grouping variable.

CAOS was developed by Robert delMas, Joan Garfield (University of Minnesota, USA), Ann Ooms (Kingston University, UK), & Beth Chance (California Polytechnic State University, USA) in 2006. This test consists of 40 items covering 11 topics, including data collection, data representation, normal distribution, probability, sampling variability, confidence intervals, significance tests, etc. All the items are in multiple-choice format, with two to five options for each item. We added another option, “I don’t know” as an alternative response.

Results of the Demonstration

Results Using CTT Approach. Table 1 and Figure 1 demonstrated how to examine the distractors via the percentages of students who selected each option based on the observed data. Table 1 shows that more students chose option A (54%) than did the correct answer, option C

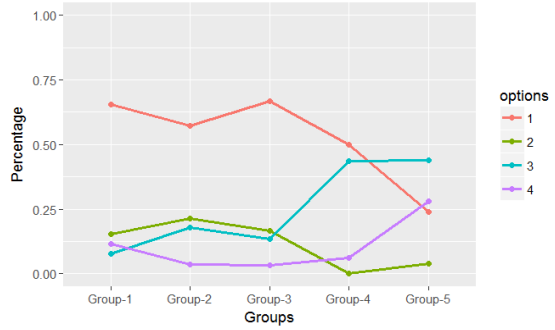
(23%), but only 10% of students indicated that they did not know the answer (option D). Distractor A apparently shows that a large number of students had misconception about sampling distribution of the mean or could not distinguish the concepts of standard error and standard deviation. Meanwhile, the trace plot in Figure 1 unveils useful information; high ability students tended to choose the correct answer, option C (blue line) more frequently, whereas low ability students tended to choose distractors A (red line) and B (green line) more frequently. Interestingly, almost no students with low ability indicated that they did not know the answer.

Table 1. Observed percentage of students who selected each option of item #32

	Option A	Option B	*Option C*	Option D
Item #32	67 (54%)	16 (13%)	*29 (23%)**	13 (10%)

Note. * denotes the correct answer; Option D is "I don't know".

Figure 1. Trace plot for item #32 with four options across five levels of ability groups



DOF Results. Table 2 and Figure 2 present the results of DOF analysis for item #32. Note that we report the results of Model-3 because the *p*-value and odds ratio together suggest a possible interaction effect between groups and ability. If we only use the *p*-value as the criterion, we would decide that there is no gender difference and no interaction effect in choosing option A and option D comparing to choosing the correct answer (option C). However, odds ratios suggest an interaction effects for all the distractors with a size of 5.77 for option A, 25.34 for option B, and 0.44 for option D, although the *p*-values for options A and C were larger than 0.05.

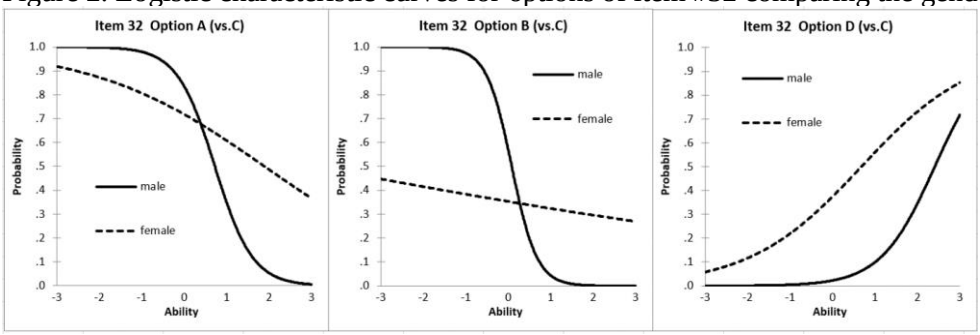
Table 2. *Wald* χ^2 test results for regression coefficients in Model-3

Item (Key)	Option	Predictor	*b*	*s. e.*	*Wald* χ^2	*p*	*Odds*
32 (C)	A	T	-2.25	0.77	-2.91	<0.001	0.11
		G	-0.69	0.69	-1.00	0.32	0.50
		*T*G*	*1.75*	*0.88*	*1.99*	*0.05*	*5.77*
	B	T	-3.36	0.95	-3.56	<0.001	0.03
		G	-0.88	0.86	-1.03	0.30	0.41
		*T*G*	*3.23*	*1.12*	*2.88*	*<0.001*	*25.34*
	D	T	1.57	1.76	0.89	0.37	4.79
		G	3.26	2.34	1.39	0.16	25.97
		*T*G*	*-0.81*	*1.89*	*-0.43*	*0.67*	*0.44*

To facilitate our interpretation of the meaning of interaction, the logistic curves are plotted in Figure 2. The interaction effects suggest that the lower ability students were more likely to choose the distractors A and B (vs. the correct answer C, see the graph on the left and in the middle). In contrast, higher ability students were more likely to choose option D "I don't know" (vs. the correct answer C, see the graph on the right). Moreover, there existed some differences in choosing the two distractors A and B between female and male groups with low ability; female students were less likely to choose the two distractors than did male students. Also, female students were more likely to choose option D "I don't know" than did male students, regardless of their

level of ability.

Figure 2. Logistic characteristic curves for options of item #32 comparing the gender groups



CONCLUSION

This study demonstrated several practical and useful psychometric tools that can help instructors to analyze the MCQ distractors and identify teaching or learning gaps. It is easy to see the misconceptions by using percentage table and trace plot. DOF analysis helps to know whether the distractors mean the same and work the same way for boys and girls.

Other variables of teaching and learning interest can be examined in a DOF analysis (in lieu of the sex variable in the current study) to see whether the distractors work the same way (e.g., previous knowledge and learning strategies).

REFERENCES

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale: Erlbaum.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice item? *Educational and Psychological Measurement, 53*, 999–1010.

Kato, K., Moen, R. E., & Thurlow, M. L. (2009). Differentials of a State reading assessment: item Functioning, distractor functioning, and omission frequency for disability categories. *Educational Measurement: Issues and Practice, 28(2)*, 28 – 40.

Lemann, N. (2000). *The Big Test*. New York, NY: Farrar, Strauss, and Giroux.

Park, M., & Wu, D. A. (2017). Investigating differential options functioning using multinomial logistic regression. *International Journal of Quantitative Research in Education, 4*, 94-119.

Wainer, H. (1989). The future of item analysis. *Journal of Educational Measurement, 26*, 191–208.

APPENDIX: A Multiple-choice Item Selected from CAOS Assessment

Item #32. It has been established that under normal environmental conditions, adult largemouth bass in Silver Lake have an average length of 12.3 inches with a standard deviation of 3 inches. People who have been fishing Silver Lake for some time claim that this year they are catching smaller than usual largemouth bass. A research group from the Department of Natural Resources took a random sample of 100 adult largemouth bass from Silver Lake and found the mean of this sample to be 11.2 inches. Which of the following is the most appropriate statistical conclusion?

A) The researchers cannot conclude that the fish are smaller than what is normal because 11.2 inches is less than one standard deviation from the established mean (12.3 inches) for this species.

B) The researchers can conclude that the fish are smaller than what is normal because the sample mean should be almost identical to the population mean with a large sample of 100 fish.

C) The researchers can conclude that the fish are smaller than what is normal because the difference between 12.3 inches and 11.2 inches is much larger than the expected sampling error.

D) I don't know.