

Open Professionals Education Network (OPEN) Resource Document Writing Multiple-Choice Questions: Item Analysis

Background

One of the unique advantages of multiple-choice questions is that you can get statistical information about how well your questions are working, also known as an item analysis. A typical item analysis contains the following information:

Item difficulty (p-value): The percentage of students that correctly answered the question, typically reported as a proportion, 0.0 to 1.00. P-values above 0.90 are very easy questions and should not be reused again for subsequent tests. P-values below 0.30 are very difficult questions and should be reviewed for confusing language, removed from subsequent tests, and/or highlighted for an area for re-instruction.

Item discrimination (R(IT)): The correlation between how well students did on the question and their total test score. The range is -1.00 to 1.00 and the higher the value, the more discriminating the item. A highly discriminating item indicates that the students who had high tests scores got the question correct whereas students who had low test scores got the question incorrect. Questions with discrimination values near or less than zero should be removed from the test.

Reliability coefficient (Alpha): A measure of how well the test questions are correlated with one another. The range is 0.0 to 1.0 and the higher the value the more reliable the overall test score. High reliability indicates that the questions are all measuring the same concepts, or general construct. Two ways to improve the reliability of the test are to 1) increase the number of questions in the test or 2) use questions that have high discrimination values in the test.

Distractor evaluation: A frequency table displaying the number and/or percent of students that selected a given distractor (answer choice). The quality of the distractors influence student performance on a test question. Although the correct answer must be truly correct, it is just as important that the distractors be incorrect. Distractors should appeal to low scorers who have not mastered the material whereas high scorers should infrequently select the distractors.

Check Your Understanding

ITEM	P	R(IT)	A	B	C	D	E
1.	0.72	0.34	667	187	38	30	10
2.	0.94	0.14	876	0	12	24	20
3.	0.35	0.42	330	98	74	183	247

Based on the item analysis, what would you do for each of the questions: keep it as is, eliminate it, or revise it? Note the correct answer is coded as "A" in this sample item analysis.

Feedback

Question 1: Keep it as is – difficulty, discrimination, and distractors all look good!

Question 2: Eliminate it – 94% got the question correct, too easy.

Question 3: Revise it - option E in particular, seems problematic.

Examples

Sample Item Analysis

N TOTAL = 932 MEAN TOTAL = 69.4 ALPHA = .84

ITEM	P	R(IT)	A	B	C	D	E
1.	0.72	0.34	667	187	37	30	10
2.	0.90	0.21	840	1	76	9	5
3.	0.60	0.39	561	233	46	88	4
4.	0.99	-0.06	923	3	3	3	0
5.	0.94	0.14	876	0	12	24	20
6.	0.77	-0.01	716	16	25	35	140
7.	0.47	0.31	432	107	68	165	157
8.	0.12	0.08	114	218	264	153	175
9.	0.08	0.04	75	64	120	67	606
10.	0.35	0.42	330	98	74	183	247

Correct answer = A

Overall, these ten questions worked reasonably well as the reliability of this test was 0.84. Some items could be eliminated or revised if used again and reinstruction may need to occur as well. Items 2, 4, and 5 are too easy as over 90% of students got these items correct. Items 8 and 9 are too difficult as less than 30% of students got these items correct. In item 8, students equally selected each of the distractors indicating guessing. In item 9, many students selected option E, indicating a problematic answer choice. Item 6 did not discriminate well indicating that students with a higher overall average test score did slightly worse on this item than the students with a lower overall average test score.

Summary & Additional Resources

Understanding how to interpret and use information based on student test scores is as important as knowing how to construct a well-designed test. Using feedback from your test to guide and improve instruction is an essential part of the process.

References:

DeVellis, R. F. (1991). *Scale development: Theory and applications*. Newbury Park: Sage Publications.

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items, 2nd ed.* Mahwah, NJ: Lawrence Erlbaum Associates.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Lawrence Erlbaum Associates.



This work is licensed under the Creative Commons Attribution 3.0 Unported License. To view a copy of this license, visit http://creativecommons.org/licenses/by/3.0/deed.en_US. Last updated December 5, 2014 by the Open Learning Initiative, Stanford University.