Test Item Analysis & Decision Making

Offered by the Measurement and Evaluation Center, University of Texas - Austin

Analyzing Multiple-Choice Item Responses

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.

Using statistical information to review your multiple-choice test can provide useful information. Three of these statistics are:

1. **Item difficulty, P:** the percentage of students that correctly answered the item.
   - Also referred to as the p-value.
   - The range is from 0% to 100%, or more typically written as a proportion as 0.0 to 1.00.
   - The higher the value, the *easier* the item.
   - P-values above 0.90 are very easy items and should not be reused again for subsequent tests. If almost all of the students can get the item correct, it is a concept probably not worth testing.
   - P-values below 0.20 are very difficult items and should be reviewed for possible confusing language, removed from subsequent tests, and/or highlighted for an area for re-instruction. If almost all of the students get the item wrong there is either a problem with the item or students did not get the concept.
   - Optimum difficulty level is 0.50 for maximum discrimination between high and low achievers.

To maximize item discrimination, desirable difficulty levels are slightly higher than midway between chance (1.00 divided by the number of choices) and perfect scores (1.00) for the item. Ideal difficulty levels for multiple-choice items in terms of discrimination potential are:

Format	**Ideal Difficulty**
Five-response multiple-choice	.60
Four-response multiple-choice	.62
Three-response multiple-choice	.66
True-false (two-response multiple-choice)	.75

2. **Item discrimination, R(IT):** the point-biserial relationship between how well students did on the item and their total test score.
   - Also referred to as the Point-Biserial correlation (PBS)
   - The range is from 0.0 to 1.00.
   - The higher the value, the more discriminating the item. A highly discriminating item indicates that the students who had high tests scores got the item correct whereas students who had low test scores got the item incorrect.

- Items with discrimination values near or less than zero should be removed from the test. This indicates that students who overall did poorly on the test did *better* on that item than students who overall did well. The item may be confusing for your better scoring students in some way.

A guideline for classroom test discrimination values is shown below:

0.40 or higher	very good items
0.30 to 0.39	good items
0.20 to 0.29	fairly good items
0.19 or less	poor items

3. Reliability coefficient (ALPHA): a measure of the amount of measurement error associated with a test score.

- The range is from 0.0 to 1.0.
- The higher the value, the more reliable the overall test score.
- Typically, the internal consistency reliability is measured. This indicates how well the items are correlated with one another.
- High reliability indicates that the items are all measuring the same thing, or general construct (e.g. knowledge of how to calculate integrals for a Calculus course).
- Two ways to improve the reliability of the test are to 1) increase the number of questions in the test or 2) use items that have high discrimination values in the test

Reliability	Interpretation
.90 and above	Excellent reliability; at the level of the best standardized tests
.80 - .90	Very good for a classroom test
.70 - .80	Good for a classroom test; in the range of most. There are probably a few items which could be improved.
.60 - .70	Somewhat low. This test needs to be supplemented by other measures (e.g., more tests) to determine grades. There are probably some items which could be improved.
.50 - .60	Suggests need for revision of test, unless it is quite short (ten or fewer items). The test definitely needs to be supplemented by other measures (e.g., more tests) for grading.
.50 or below	Questionable reliability. This test should not contribute heavily to the course grade, and it needs revision.

Distractor Evaluation

Another useful item review technique to use is **distractor evaluation**.

The distractor should be considered an important part of the item. Nearly 50 years of research shows that there is a relationship between the distractors students choose and total test score. The quality of the distractors influences student performance on a test item. 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. Reviewing the options can reveal potential errors of judgment and inadequate performance of distractors. These poor distractors can be revised, replaced, or removed.

One way to study responses to distractors is with a frequency table. This table tells you the number and/or percent of students that selected a given distractor. Distractors that are selected by a few or no students should be removed or replaced. These kinds of distractors are likely to be so implausible to students that hardly anyone selects them.

Caution when Interpreting Item Analysis Results

W. A. Mehrens and I. J. Lehmann provide the following set of cautions in using item analysis results (Mehrens, W. A., & Lehmann, I. J. (1973). Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart and Winston, 333-334):

- Item analysis data are not synonymous with item validity. An external criterion is required to accurately judge the validity of test items. By using the internal criterion of total test score, item analyses reflect internal consistency of items rather than validity.

- The discrimination index is not always a measure of item quality. There is a variety of reasons an item may have low discriminating power:
    a) extremely difficult or easy items will have low ability to discriminate but such items are often needed to adequately sample course content and objectives;
    b) an item may show low discrimination if the test measures many different content areas and cognitive skills. For example, if the majority of the test measures "knowledge of facts," then an item assessing "ability to apply principles" may have a low correlation with total test score, yet both types of items are needed to measure attainment of course objectives.

- Item analysis data are tentative. Such data are influenced by the type and number of students being tested, instructional procedures employed, and chance errors. If repeated use of items is possible, statistics should be recorded for each administration of each item.

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.

Lord, F.M. (1952). The relationship of the reliability of multiple-choice test to the distribution of item difficulties. Psychometrika, 18, 181-194.

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

Activity: Item Analysis

Below is a sample item analysis performed by MEC that shows the summary table of item statistics for all items for a multiple-choice classroom exam. Review the item difficulty (P), discrimination (R(IT)), and distractors (options B-E).

Item Analysis (sample of 10 items) – correct answer is "A"

Summary Table of Test Item Statistics

<test name>

N TOTAL = 932 MEAN TOTAL = 69.4 S.D. TOTAL = 10.2 ALPHA = .84

ITEM	P	R(IT)	NC	MC	MI	OMIT	A	B	C	D	E
1.	0.72	0.34	667	71.56	67.66	1	**667**	187	37	30	10
2.	0.90	0.21	840	70.11	69.02	1	**840**	1	76	9	5
3.	0.60	0.39	561	72.66	65.47	0	**561**	233	46	88	4
4.	0.99	-0.06	923	69.34	69.90	0	**923**	3	3	3	0
5.	0.94	0.14	876	69.76	68.23	0	**876**	0	12	24	20
6.	0.77	-0.01	716	69.34	69.57	0	**716**	16	25	35	140
7.	0.47	0.31	432	72.76	66.16	3	**432**	107	68	165	157
8.	0.12	0.08	114	71.61	68.39	8	**114**	218	264	153	175
9.	0.08	0.04	75	70.78	69.03	0	**75**	64	120	67	606
10.	0.35	0.42	330	75.24	63.54	0	**330**	98	74	183	247
.											
.											
40.											

Which item(s) would you remove altogether from the test? Why?

Which distractor(s) would you revise? Why?

Which items are working well?

Activity: Item Breakdown

Below is a sample item analysis performed by MEC that shows the breakdown for one item on a multiple-choice classroom exam. Review the splits, mean for each alternative, and pattern of responses to the distractors (options B-E) within each split.

Item Analysis (sample of 10 items) – correct answer is "A"

Table 1
Item Breakdown

	ITEM NO.	3			KEYED RESPONSE = A		
SPLIT	OMIT	A	B	C	D	E	SUM
1	0	199	20	1	13	0	233
2	0	157	55	5	16	0	233
3	0	119	66	14	31	3	233
4	0	86	92	26	28	1	233
SUM	0	561	233	46	88	4	932
MEAN	0.0	72.8	64.7	61.1	65.7	63.3	
	P TOT = 1.00		P = .60		R(IT) = .39		

What does the pattern of responses for the correct and incorrect alternatives across the various splits tell you about the item?

Which distractor(s) would you revise? Why?

Activity: Frequency Distribution

Below is a sample item analysis performed by MEC that shows the frequency distribution of total scores for a multiple-choice classroom exam. Review the frequency, percentile ranks, standard scores, and percentages for each raw score.

Frequency Table of Raw Scores

<test name>

N TOTAL = 932 MEAN TOTAL = 69.4 S.D. TOTAL = 10.2 ALPHA = .84

RAW SCORE	FREQ	PCTL RANK	PERCENT CORRECT	STAND. 50-10 SCORE	PCT
99	1	100	99.00	79.05	.1
97	1	100	97.00	77.09	.1
95	1	100	95.00	75.12	.1
94	1	100	94.00	74.14	.1
96	2	99	93.00	73.16	.2
92	3	99	92.00	72.18	.3
91	8	99	91.00	71.19	.9
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
70	40	54	70.00	50.57	4.3
69	35	50	69.00	49.59	3.8
68	38	46	68.00	48.61	4.1
67	36	42	67.00	47.63	3.9
66	43	37	66.00	46.65	4.6
65	42	33	65.00	45.66	4.5
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
42	2	1	42.00	23.08	.2
41	1	1	41.00	22.10	.1
40	1	1	40.00	21.12	.1
39	1	0	39.00	20.13	.1
37	1	0	37.00	18.17	.1
36	3	0	36.00	17.19	.3

How can you use the frequency counts (and related percentages) to determine how the class did as a whole?

How would you use the standard scores to compare 1) the same students across different tests or 2) the overall scores between tests?