Item Analysis of Classroom Tests
(Summarized from Oosterhof, A. Classroom Applications of Educational Measurement)

In the first part of the semester we learned about item analysis for scales and questionnaires (e.g., item-total correlation, change in Cronbach's alpha if item is deleted, etc.). Such information was used to identify poorly performing items which could either be revised or eliminated from a larger scale of items. Item analysis for classroom tests is similar, but sometimes performed differently since teachers rarely have sophisticated software to perform such analyses. The goal, however, is the same: to identify poorly performing items so they can be revised for future use. Note, however, sometimes items perform poorly not because of the item, but because the content was not taught well. This suggests the problem is not with the item but with instruction. A complete item analysis for classroom tests has several components, each of which will be described below.

1. Test Analysis

When examining a test, one may first note how students performed on the whole test. If most students performed very well on a test, that could mean

- the test content was taught well, students learned the content, and scored high on the test – this is an ideal situation; or
- the test was too easy and could not adequately assess student content understanding.

If student overall test performance was poor, this could mean

- instruction was poor, and this is reflected in poor performance by students on the test;
- material was more difficult than expected so student performance was lower than hoped;
- test content did not align well with instruction or objectives, so the test lacks content validity;
- the test was simply too difficult; or
- a combination of several issues listed above.

2. Item Analysis: Item Difficulty

Item difficulty refers to the percentage or proportion of students who correctly responded to an item. If, for example, 43% of the students correctly responded to item X, then X has an item difficulty of 43% or .43. Usually item difficulty will be presented in proportion, not percentage, format.

For items scored dichotomously – scored as 1 = correct and 0 = incorrect – the item difficulty is the sum of students who answered the item correctly divided by the number of students who completed the test. For example, if 15 students answered an item, but only 8 students answered the item correctly, the item difficulty would be 8/15 = .53 or 53%.

For items that are graded using a rubric or have a range of scores (e.g., short-answer item #1 is worth 3 points, essay item #2 is worth 5 points), these item scores are converted to proportion correct and then the mean of those proportions is used to determine item difficulty. For example, short-answer item #1 is worth 3 points total. Say four students complete the item with scores of 3.0, 2.5, 2.0, and 3.0. Those scores are converted to proportion correct:

- 3.0 / 3 = 1.00 (100%)
- 2.5 / 3 = .83 (83%)
- 2.0 / 3 = .67 (67%)
- 3.0 / 3 = 1.00 (100%)

Item difficult is the mean of these proportions = (1.00 + .83 + .67 + 1.00) / 4 = 3.5 / 4 = .875. In case this formula is not clear, the devisor, 4, is the number of students, the number of scores, for item #1.

3. Item Analysis: Item Discrimination

Item discrimination, another component of item analysis, refers to the ability of the item to discriminate between more and less knowledgeable students. Item discrimination is calculated by finding the difference in item difficulty between

two sets of students, those with above average performance and those of below average performance. Average performance is defined as the total test score for each student, i.e., sum of all test items or the overall percentage correct on the test.

Ideally, the two groups are defined such that the top performing 1/4 (25%) or 1/3 (33%) of the class on the test represents the above average group and the bottom 1/4 (25%) or 1/3 (33%) the below average groups. Sometimes the top 1/2 (50%) is compared to the bottom 1/2 (50%) if the number of students is small. In most cases for classroom tests, the number of students will be small, usually less than 30, so best to use the top 1/3 and bottom 1/3, or the top half and bottom half, for calculating item discrimination.

Item discrimination for a given item would be:

item discrimination =	item difficulty for top performing group	−	item difficulty for low performing group

If the top 1/3 of students all correctly answered the chosen item, their item difficulty would be 1.00, and if 40% of the bottom 1/3 correctly answered the item, their item difficulty would be .40. The item discrimination would be:

item discrimination = 1.00 - .40 = .60

The larger the discrimination index, the better performing the item in terms of distinguishing between knowledgeable students and less knowledgeable students for the content domain and skill level sampled by the item. For classroom, teacher-constructed tests, discrimination indexes of .20 and greater are good when using the upper 1/4 vs. lower 1/4, and one should expect lower levels of discrimination when using larger groups as the basis of the upper and lower groups, such as upper 1/3 vs. lower 1/3, or upper 1/2 vs. lower 1/2. Some measurement specialists recommend only using the top 10 and bottom 10 performers, even if a hundred (or more) students took the test. As these recommendations reveal, there is no uniform agreement on how to define the top and bottom group, so for your own practice work with whatever seems best for you. Key is to have a large sample for both groups, i.e., 10 or more students in each group.

The discrimination index ranges from -1.00 to 1.00. Negative values indicate that less knowledgeable students are answering the item correctly more often than more knowledgeable students. This is a signal that something is wrong with the item and it is deficient. A common problem with such items is ambiguity, i.e., the item has more than one correct response or no correct response.

Another method for calculating the discrimination index is to calculate the correlation between the item score and the total test score across the sample of students. This correlation is called the item-total correlation and is the same item-total correlation learned for item analysis with scales and questionnaires. This is especially the method used when one wishes to calculate discrimination for an item that is scored with multiple points or partial credit, like essay or short-answer items.

Normally, one will use all students, not just the top and bottom groups, to calculate the item-total correlation. As with the discrimination index, the item-total correlation ranges from -1.00 to 1.00. The better the item discriminates, the larger will be the positive item-total correlation. Like negative discrimination indices, negative correlations indicate that the item is not properly discriminating and signals a problematic item.

To illustrate both methods for calculating discrimination, listed below are the total test scores and the item performances for two groups of students. Note that all information is sorted by Total Test Score, so it is easy to identify top and bottom performers.

Table 1: Example Discriminations: Note that one should sort the table or spreadsheet by Total Test Score to more easily identify top and bottom test performers

Student	Group Identification	Item 1 Multiple Choice (1 = correct, 0 = incorrect)	Item 20 Essay (Partial Score out of possible total 5)	Item 20 Essay (Proportion Correct out of 5 Points)	Total Test Score (% Correct)
Bill	Top 1/3	1	4.5	4.5/5 = .90	97
Brandon	Top 1/3	0	4	4.0/5 = .80	93
Bertha	Top 1/3	1	4.5	4.5/5 = .90	89
Brianna	Top 1/3	1	3.5	3.5/5 = .70	87
Brittany	Middle 1/3	1	4	4.0/5 = .80	85
Bailey	Middle 1/3	1	3.5	3.5/5 = .70	80
Brenda	Middle 1/3	0	4	4.0/5 = .80	79
Bonnie	Middle 1/3	1	2.5	2.5/5 = .50	76
Bryan	Bottom 1/3	0	3	3.0/5 = .60	75
Bart	Bottom 1/3	0	3.5	3.5/5 = .70	74
Barney	Bottom 1/3	1	2.5	2.5/5 = .50	73
Bernie	Bottom 1/3	0	2	2.0/5 = .40	66

The Item 1 difficulty the top group is 3/4 = .75, and 1/4 = .25 for the bottom group. The item discrimination for Item 1 is:

Item 1 discrimination = .75 - .25 = .50

The item-total correlation for all 12 students uses the 1,0 scoring for Item 1 and the Total Test Score is Pearson r = .36, which, while lower, corresponds with the above discrimination index. Both the discrimination index of .50 and the correlation of .36 tells us students with higher scores also performed better on item 1, which is what we hope to find with a good performing item.

For the Item 20, the essay item, discrimination can be found by calculating the difference in mean proportion correct between the two groups. For the Top 1/3 group, the mean item difficulty is (.90+.80+.90+.70)/4 = .825. For the bottom 1/3 group the mean item difficulty is (.60+.70+.50+.40)/4 = .55.

Item 20 discrimination = .825 - .55 = .275

The item-total correlation for Item 20 is found by correlating the Item 20 score, either raw score or proportion correct, with the Total Test Score for all students. Since Pearson r is invariant to linear transformations, it does not matter whether raw score or proportion score is used for Pearson r. For Item 20, the item-total correlation is .853. Both item discrimination and item-total correlation are positive which indicates the essay item functions as it should by discriminating between more and less knowledgeable students.

Finally, there is a direct relationship between item difficulty and item discrimination. The more (or less) difficult the item, the less it discriminates. When item difficulty approaches the half-way point (.50), item discrimination will most likely be maximized.

4. Item Analysis: Distractor Analysis for Multiple-choice Items

Distractor analysis enables one to determine the pattern of responses across all options of a multiple-choice item. This is useful because it allows one to learn which are the more successful distractors. If an item is performing appropriately, then the upper performing group should choose the correct option more frequently than any other option, and, as

noted in the discrimination discussion, should choose the correct option more often than the lower performing group. Additionally, the lower performing group should choose distractors more often than the upper performing group.

Table 2: Distractor Analysis of Item 1 (Numbers are percentages)

Item X	A	B*	C	Omit
Upper 1/4	25	75	0	0
Lower 1/4	25	25	50	0
All Students	25	58	15	0

Note. The Omit category represents students who did not answer the item.
*Correct response.

Note that Item 1, illustrated in Table 2, behaves accordingly. That is, the upper group chose B more often than the lower group, and the lower group chose distractors (options A and C) more often than the upper group.

Should the pattern illustrated in Table 2 not occur, then the item will probably need some revision. For example, if lower group chooses the correct response more often than the upper group, then the item is most likely ambiguous. Or, if the upper group chooses one of the distractors more often than the lower group (but a larger percentage of the upper groups still chooses the correct response), then that distractor needs revision. The last type of problem that may be observed is the case in which more students (both upper and lower) choose a distractor rather than the correct option. When this occurs, students view the distractor as more correct than the correct option, and the item (or instruction) should be carefully reviewed.

Distractor analysis is beneficial in learning for which skills and content students are having the greatest and least success. Oftentimes items will contain distractors that represent common mistakes, and when students select such distractors with great frequency, this is a clear indication that further instruction is necessary. Thus, proper interpretation of distractors may lead to alterations to instruction.

Note: The three components of item analysis (difficulty, discrimination, and distractor analysis) should be viewed cautiously when one has a small sample of students. Ideally large groups, say 50 or greater, are needed for reliable analysis. But when small numbers are present, an analysis should be viewed as preliminary, and one should collect more data from additional classes over time.

5. Detecting Ambiguities That are Causing Students Difficulty

The item difficulty index indicates which items are causing students difficulty. However, a more detailed examination of item analysis may reveal additional problems.

5a. Interpreting an Item Analysis

If an item is causing some difficulty, then one must next determine whether the difficulty results from the item's ability to discriminate between more and less knowledgeable students. The next step is the examine the distractor analysis. Usually one need not examine, in detail, the distractor analysis if the item has suitable difficulty and discrimination. In sum, to interpret an item analysis, one must first examine the item difficulty, then item discrimination, and finally distractor analysis.

When an item has low discrimination and moderate difficulty, then the item is most likely ambiguous and should be revised, or perhaps instruction was less than adequate and should be corrected. When examining an item, one should consider all aspects: stem, correct option, and distractors. Moreover, one should also ensure that the item reflects the desired capability and performance objective.

5b. Using Student Input to Interpret an Item Analysis

One should first perform an item analysis of all items on the test. Once items are identified as questionable, based upon the item analysis, the instructor should next survey students and find why they selected various distractors, or why the item posed difficulty for students. Often this discussion will reveal either student misunderstandings, or problems with instruction. In addition, this communication may reveal problems inherent with the item.

6. Item Analysis for Item Formats Other than Multiple-Choice

It is possible to perform an item analysis for (a) Completion and Short-Answer Items, (b) Essay Items, and (c) Alternate-Choice Items. The item analysis will typically consist of difficulty and discrimination. Distractor analysis is unique to multiple-choice items.

6a. Completion and Short-Answer Items

Since these items can usually be scored as either 1 (correct) or 0 (incorrect), calculation of difficulty and discrimination is the same as discussed earlier. If one assigns partial points for short-answer items, then one may calculate both difficulty and discrimination using the procedures described above and below for essay items.

6b. Essay Items

With essay items one may incorporate partial grades (e.g., 4 out of 5 possible points), so item difficulty requires that item scores be converted to proportion correct, then the mean proportion correct for that item is taken across students to determine item difficulty. Using data from Table 1, the mean proportion correct for item 20 is shown below.

Table 1 again

Student	Group Identification	Item 1 Multiple Choice (1 = correct, 0 = incorrect)	Item 20 Essay (Partial Score out of possible total 5)	Item 20 Essay (Proportion Correct out of 5 Points)	Total Test Score (% Correct)
Bill	Top 1/3	1	4.5	4.5/5 = .90	97
Brandon	Top 1/3	0	4	4.0/5 = .80	93
Bertha	Top 1/3	1	4.5	4.5/5 = .90	89
Brianna	Top 1/3	1	3.5	3.5/5 = .70	87
Brittany	Middle 1/3	1	4	4.0/5 = .80	85
Bailey	Middle 1/3	1	3.5	3.5/5 = .70	80
Brenda	Middle 1/3	0	4	4.0/5 = .80	79
Bonnie	Middle 1/3	1	2.5	2.5/5 = .50	76
Bryan	Bottom 1/3	0	3	3.0/5 = .60	75
Bart	Bottom 1/3	0	3.5	3.5/5 = .70	74
Barney	Bottom 1/3	1	2.5	2.5/5 = .50	73
Bernie	Bottom 1/3	0	2	2.0/5 = .40	66
Means		.583	3.458	.691	81.16

Mean Proportion Correct Item 20 = (0.9 + 0.8 + 0.9 + 0.7 + 0.8 + 0.7 + 0.8 + 0.5 + 0.6 + 0.7 + 0.5 + 0.4)/12 = .691

This difficulty can also be calculated from the ratio of mean points earned divided by points possible. The maximum points possible for item 20 is 5, and the mean score for item 20 is

Mean Grade Item 20 = (4.5 + 4 + 4.5 + 3.5 + 4 + 3.5 + 4 + 2.5 + 3 + 3.5 + 2.5 + 2)/12 = 3.458.

Difficulty for item 20 is the mean grade divided by the total points possible: 3.458 / 5 = .691.

For a quicker and less formal analysis, a simplified approach can be used for calculating difficulty for partial grade items like essays. To do this, first determine what is a minimally acceptable response level, in terms of points awarded, and calculate the percentage of students who scored above (or below) this level. For example, an essay item may be worth a total of 10 points, yet one may decide that a minimum of 6 points is needed to be consider acceptable performance on the essay. One then calculates the proportion of student who scored 6 or more points and this proportion represents the item difficulty. Item discrimination can also be calculated from this proportion. After determining the top and bottom performing groups for the whole test, one may calculate, for each group, the percentage of students awarded 6 or more points to the essay item. Thus, for example, the top performing group may have 85% receive 6 or more points for the essay, and the bottom group may have 53% receive 6 or more, so the item discrimination would be .85 - .53 = .32.

A better approach, however, is the use the item-total correlation. That is, calculate the correlation between the points received for the essay item by each student with the total score received by each student on the test. The higher the positive correlation, the better the item discriminates.

5c. Alternate-Choice Items

Since these items can usually be scored as either 1 (correct) or 0 (incorrect), calculation of difficulty is the same as discussed for multiple-choice items, and the procedure for calculating discrimination is identical to that discussed above for multiple-choice formats.

6. Self-Test

Use Table 4 to answer items 1 through 14.

Table 4: An item analysis for two items.

Item 1		A	B	*C	D	E	Omit
	Upper 1/3	4%	6%	74%	5%	11%	0
	Lower 1/3	21%	12%	38%	25%	4%	0
	All Students	12%	10%	54%	15%	9%	0

54%, or 15 of 28 students, taking the test correctly answered item; 36 percentage points separate the upper and lower groups on the correct answer (option C).

Item 2		A	B	C	D	E	Omit
	Upper 1/3	28%	42%	15%	0%	15%	0
	Lower 1/3	13%	51%	10%	0%	26%	0
	All Students	20%	46%	13%	0%	21%	0

46%, or 13 or 28 students, taking the test correctly answered item; -9 percentage points separate the upper and lower groups on the correct answer (option B)

1. What is the difficulty of item 1?
2. What is the difficulty of item 2?
3. What is the discrimination of item 1?
4. What is the discrimination of item 2?
5. Which options, distractors, within item 1 are performing appropriately?

6. Which options within item 2 are performing appropriately?

Items 7 through 14 are interpretations of the information in Table 4. Indicate (yes or no) whether each interpretation is correct.

7. Item 1 is easier than item 2.
8. Item 1 is more ambiguous than item 2.
9. The analysis suggests the discrimination of item 1 would be improved if option E were revised.
10. The analysis suggests that the discrimination of item 2 would be improved if option E were revised.
11. If reviewing the wording of item 1 indicates that option A is a good distractor, the teacher should evaluate the adequacy of instruction relevant to this item.
12. If reviewing the wording of item 2 indicates that option A is a good distractor, the teacher should evaluate the adequacy of instruction relevant to this item.
13. Overall, item 1 appears to be well written.
14. Overall, item 2 appears to be well written.

Items 15 through 19 suggest some benefits of using student input to help interpret an item analysis. Indicate (yes or no) whether each of these statements represents a benefit.

15. Test scores can be more readily corrected for student guessing.
16. Ambiguity within a test item can be identified.
17. Ambiguity within instruction given students can be identified.
18. Misconceptions learned by students can be addressed

To answer the following items, use Table 5 below. Before you can answer the following items, you must (a) convert each response to correct or incorrect (except for the essay item 6), (b) find the test total score for each student, then (c) sort the table in order of total score to perform the item analysis.

Table 5: Test Results for 12 Students

Student	Item 1: T/F	Item 2: MC B is correct	Item 3: MC A is correct	Item 4: MC C is correct	Item 5: MC A is correct	Item 6 Essay Max score = 5	Total Score
Student 1	true	B	A	C	B	4	
Student 2	true	A	B	C	A	4	
Student 3	false	B	C	A	C	3	
Student 4	false	B	A	B	A	4	
Student 5	false	A	B	C	B	4	
Student 6	true	A	A	A	A	5	
Student 7	true	C	C	B	C	2	
Student 8	true	B	A	C	A	4	
Student 9	true	B	A	C	A	4	
Student 10	true	C	B	A	B	3	
Student 11	true	B	A	C	A	5	
Student 12	false	C	C	B	C	2	
Mean							

19. Calculate item difficulty for each of the six test items.
20. Calculate item discrimination for each of the six test items. Use a 50/50 split, top 6 vs bottom 6.
21. Perform a distractor analysis for Item 4.

Self-Test Answers

1. What is the difficulty of item 1?
   .54 (or 54%)

2. What is the difficulty of item 2?
   .46 (or 46%)

3. What is the discrimination of item 1?
   .74 - .38 = .36

4. What is the discrimination of item 2?
   .42 - .51 = -.09 (negative, this indicates something is wrong)

5. Which options, distractors, within item 1 are performing appropriately?
   A B C D, distractor E is a problem since more upper students selected it than lower students

6. Which options within item 2 are performing appropriately?
   Only distractor E; the other response options are inverted, more upper students selected distractors (A C D) than lower students, and more lower students answered it (selected option B) correctly than upper students.

7. Item 1 is easier than item 2.
   Yes, difficulty was .54 vs .46, more answered item 1 correctly so easier.

8. Item 1 is more ambiguous than item 2.
   No, distractors and correct response performed as they should expect for option E in item 1.

9. The analysis suggests the discrimination of item 1 would be improved if option E were revised.
   Yes, option E is pulling more upper students than lower students and this negatively affects discrimination.

10. The analysis suggests that the discrimination of item 2 would be improved if option E were revised.
    No, option E is performing correctly for item 2, the other options, or the item stem, must be revised or instruction revised.

11. If reviewing the wording of item 1 indicates that option A is a good distractor, the teacher should evaluate the adequacy of instruction relevant to this item.
    No, distractor A of Item 1 is performing as it should therefore instruction appears to be sound.

12. If reviewing the wording of item 2 indicates that option A is a good distractor, the teacher should evaluate the adequacy of instruction relevant to this item.
    Distractor A of item 2 is not performing well, too many upper students select it relative to lower students. Either the item stem, distractor A, or instruction must be revised.

13. Overall, item 1 appears to be well written.
    Yes

14. Overall, item 2 appears to be well written.
    No

15. Test scores can be more readily corrected for student guessing.
    Yes

16. Ambiguity within a test item can be identified.

Yes

17. Ambiguity within instruction given students can be identified.

Yes

18. Misconceptions learned by students can be addressed

Yes

19. Calculate item difficulty for each of the six test items.

Below in Table 5 are student answers converted to correct = 1 and incorrect = 0 numbers for each item except for item 6 which does not need this conversion. This conversion makes calculation of difficulty and discrimination easier.

Table 5: Test Results for 12 Students

Student	Item 1: T/F	Item 2: MC B is correct	Item 3: MC A is correct	Item 4: MC C is correct	Item 5: MC A is correct	Item 6 Essay Max score = 5	Total Score
Student 11	1	1	1	1	1	5	10
Student 8	1	1	1	1	1	4	9
Student 9	1	1	1	1	1	4	9
Student 6	1	0	1	0	1	5	8
Student 1	1	1	1	0	0	4	7
Student 2	1	0	0	1	1	4	7
Top 50% students above, bottom 50% below, sorted by total test grade.							
Student 4	0	1	1	0	1	4	7
Student 5	0	0	0	1	0	4	5
Student 3	0	1	0	0	0	3	4
Student 10	1	0	0	0	0	3	4
Student 12	0	0	1	0	1	2	4
Student 7	1	0	0	0	0	2	3
Mean	8/12 = .66	6/12 = .50	7/12 = .58	5/12 = .41	7/12 = .58	3.66	
						3.66/5 =.73	

Item discrimination

Item 1 = .66
Item 2 = .50
Item 3 = .58
Item 4 = .41
Item 5 = .58
Item 6 = .73

20. Calculate item discrimination for each of the six test items. Use a 50/50 split, top 6 vs bottom 6.

See table next page.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6
Top 50% Item Difficulty	1.00	0.66	0.83	0.66	0.83	4.33/5 = .86
Bottom 50% Item Difficulty	0.33	0.33	0.33	0.33	0.16	3.00/5 = .60
Item Discrimination	1.00-.33 = .67	.66-.33 = .33	.83-.33 = .50	.66-.33 = .33	.83-.16 = .67	.86-.60 = .26

21. Perform a distractor analysis for Item 4.

To make this easier, all other items were deleted, and then students sorted by total score just like above.

Student	Item 4: MC C is correct	Total Score
Student 11	C	10
Student 8	C	9
Student 9	C	9
Student 6	A	8
Student 1	C	7
Student 2	C	7
Student 4	B	7
Student 5	C	5
Student 3	A	4
Student 10	A	4
Student 12	B	4
Student 7	B	3

Next, count the times each option is selected by upper and lower scoring students and report counts in a table, like below.

Counts reported, not percentages, in table

Item 4		A	B	*C	Omit
	Upper 1/2	1	0	5	0
	Lower 1/2	2	3	1	0
	All Students	3	3	6	0

All options appear to be performing as expected. More of the lower scoring students selected distractors A and B than upper scoring students, and upper scoring students selected the correct response, C, 5 times more often than lower performing students.