Multiple-Choice Construction Checklist

© 2001 Robert Runté, Ph.D. Faculty of Education, University of Lethbridge Runte@uleth.ca

Whole Item

1. The item represents a significant educational objective. ❑
   "If I could only ask one question, is this the question I would ask?"
   "Is this what I want students to retain from the course ten years from now?"
2. The item addresses a single problem. ❑
3. The item does not rely on any other item for its correct answer. ❑
4. The item is free of cultural, gender, religious, or other biases. ❑
5. The item does not require more than its fair share of the student's total time for the test. ❑
6. The vocabulary is appropriate to the student's grade level. ❑
   That is, the syntax and vocabulary are not challenging, unless the specific intent of the question is to test syntax and vocabulary.
7. The item does not measure personal opinion. ❑
   UNACCEPTABLE: "The greatest inventor of the 20th Century was..."
   ACCEPTABLE: "The most prolific inventor of the 20th Century was..."
8. The entire item is contained on the same page. ❑
   Where a passage or data for an item takes the entire page, it is acceptable to place related questions on the *facing* page. (Avoid placing data on a back of the page forcing students to flip back & forth.)
9. Key words from a reading passage or data set are **not** repeated in the stem or alternatives ❑
10. The entire item, including incorrect alternatives, has been proofread/spell checked. ❑

Stem

1. The stem is self-contained and introduces what is expected of the student. ❑
   The examinees should be able to read the stem and answer the question before reading the alternatives.
2. The stem is clearly worded and free of ambiguity. ❑
3. The stem is free of irrelevant or unnecessary detail. ❑
4. The stem asks a question that has a definite answer. ❑
5. All conditions and qualifications necessary to make the intended response definitely the best of the available ❑
   alternatives have been specified.
   Wrong: "The BEST sound is obtained from a flute when..."
   The criteria for "best" must be specified:
   Better: "The LOUDEST sound is obtained from a flute when..."
   "The CLEAREST sound is obtained from a flute when..."
6. In BEST-answer items (as above example), the key qualifier in the stem has been emphasized by the use of capital letters, boldface, and/or underlining. ❑
7. The stem does not provide grammatical clues to any alternative; that is, the stem is grammatically correct in its relationship to each alternative. ❑
8. The use of absolutes (e.g., "always", "never") has been avoided. ❑
9. The use of negatives has been avoided. Where negatives must be used, the negative wording has been emphasized by the use of capital letters, bolding, underlining, etc. ❑
   WRONG: "Which of the following countries does not have a Marxist government?"
   Slightly BETTER: "Which of the following countries does <u>NOT</u> have a Marxist government?"
   BETTER: "All of the following countries have Marxist governments **EXCEPT ONE**. Choose the **EXCEPTION**:"
   BEST: "An example of a capitalist nation is:"
10. "Of the following" used only in best-answer questions ❑

Multiple-Choice Construction Checklist (Continued):

Alternatives

1. The alternatives are all appropriate to the question asked or implied by the stem. ❑
2. The distracters are all plausible. ❑
3. The use of trickery has been avoided. ❑
4. All alternatives are consistent with higher levels of learning; that is, the item would still be correct (though easier) in a more senior course. ❑
5. Distracters do not present false information. ❑

Poor: The event that triggered WWII was
a) the bombing of Pearl Harbor in 1938
b) Germany's invasion of Poland
c) England's invasion of the Rhineland
d) the American assassination of Churchill

Better: The event that triggered Canadian entrance into WWII was the
a) Japanese bombing of Pearl Harbor
b) German invasion of Poland
c) Italian invasion of Ethiopia
d) Serbian assassination of Arch Duke Ferdinand

6. The alternatives are stated as briefly and simply as possible. ❑
7. Wording common to all alternatives has been placed in the stem. ❑

WRONG: A perception check is a
(a) verbal statement that reflects our interpretation of what others say
(b) verbal statement that reflects how others interpret what we say

RIGHT: A perception check is a verbal statement that reflects
(a) our interpretation of what others say
(b) how others interpret what we say

8. The correct alternative is not stated in textbook or stereotypical language. ❑
9. One alternative is not significantly longer than the others. ❑
10. Key words in the stem have not been repeated in any of the alternatives. ❑
11. The alternatives are grammatically consistent with the stem, and parallel in grammatical structure, type of content, length, and complexity. ❑
12. Overlapping alternatives have been avoided. ❑

23. Who was dancing in the story?
(a) Jane was dancing.
(b) Mary was dancing.
(c) Jill was dancing.
*(d) All the ladies were dancing.
(d) subsumes (a), (b) and (c), but (a), (b) and (c) nevertheless remain correct answers.

13. "All of the above" has not been used. ❑
Student may see that alternative (a) is correct, choose it, and go on to the next question without noticing the "all of the above" option. (In effect, students get the item wrong, even though they chose a correct answer to the question posed) Furthermore, if examinees can identify one alternative as incorrect, they can automatically eliminate "all of the above", so question no longer involves guessing out of four alternatives, but has been narrowed down to one out of two
14. Combinations of alternatives (e.g., "both A and B") have been avoided. ❑
15. "None of the above" has not been used as an alternative. ❑
Students may legitimately eliminate "correct" answer through beyond the course knowledge
Calculation items are an exception, and may use "none of the above".
16. Keyed answer is inarguably correct. ❑
17. Alternatives on mathematics and science tests show correct number of significant digits (except where item is assessing student's knowledge of significant digits). ❑
18. Absolutes such as "always" and "never" have been avoided. ❑
19. The number of alternatives is consistent, at least within groups of items. ❑
Until recently, four alternatives were recommended. Current studies suggest, however, that attempts to use four alternatives often results in the adoption of implausible fourth alternatives, and that this may lead to an overall decline in the standards applied to other alternatives. Three alternatives are therefore recommended.

Multiple-Choice Construction Checklist (Continued):

Key

1. The keyed answer is the only correct, or clearly the best, answer. ❑
2. The placement of the keyed answer has been varied; i.e., same number of 'a's, 'b's, 'c's, etc. ❑
3. Alternatives have been arranged in ascending or descending order. ❑
   Numerical alternatives are arranged in either ascending or descending order.
   Non-numerical alternatives are arranged in logical order (e.g., alphabetical, chronological, positive to negative), or where there is no obvious continuum, by length of the alternative in either ascending or descending order.
4. Every item is independent of every other item. ❑
   Choosing the correct answer to any item does not depend on students having chosen correct response to a previous item.
   Information provided in one question does not provide clues to another.
5. Each question is worth the same number of marks. ❑
   Instead of assigning items different weights, important topics may be assigned additional questions.

Context-Dependent Multiple-Choice Items

1. The answer to one item is independent of the answer to all previous and following items on the same passage or data. ❑
2. The number of questions asked on a single passage or data set reflects the length and complexity of the passage. ❑
3. The passage or data presented is new to all students; or material with which all students have had an equal opportunity to familiarize themselves. ❑
4. Visual material is clearly reproduced. ❑
5. Any questions related to the passage are on the same page, or at minimum, on the facing page. ❑

Compiling a Multiple-Choice Test

Use a table of specifications (blueprint) to ensure representative sample of course content.

Sample Table Of Specifications

Bloom's Taxonomy / Subject Content	Knowledge & Comprehension	Application	Analysis, Synthesis & Evaluation	Totals
Art History	40%	5%	5%	50%
Art Appreciation	5%	5%	40%	50%
Totals	45%	10%	45%	100%

- usually a table of specifications would have more detailed content; limit of two categories here to keep illustration simple
- often combine categories on Bloom's taxonomy to simplify presentation, provide some flexibility, etc.
- totals tell one at a glance what percentage of course emphasis given to each topic and what percentage lower and higher level mental processes

Sample Rationale:

- Both topics equally important so = 50% each
- art history requires memorization of names and dates, so relatively more weight on knowledge and comprehension than on analysis, synthesis, or evaluation
- art appreciation requires student to analyze art object and synthesize personal response to produce critique (evaluation), so this unit will have relatively more weight on higher mental activities
- table demonstrates overall balance between lower and higher mental activities

Matching Question Checklist:

1) All items are homogeneous ❑
2) Columns have been given appropriate descriptive titles ❑
3) The "premises" are on the left, "options" are on the right ❑
4) The shorter words are in the second (i.e., right hand, "options") column ❑
5) Columns are arranged alphabetically or numerically ❑
6) There are more options than premises; or each option may be used more than once ❑
7) Clear directions have been provided ❑
8) The whole question is contained on the same page ❑

Completion Items Checklist:

1) What is to go in the blank has been clearly specified ❑
2) Generally, one blank per item ❑
3) The blank is at the end of the item (where possible) ❑
4) All answer blanks have been put in single column to facilitate grading. ❑
5) Statements have not been taken directly from the text or class notes ❑
   -promotes memorization without understanding
   -statements out of context are often ambiguous
6) Unintended hints have been avoided: ❑
   -'an' and 'a's before blanks; plurals, etc; length of blanks ❑

True-False Checklist:

1. Statements are unequivocally true or false. ❑
2. Each item is as significant as possible. ❑
3. Only one central idea is included in each statement. ❑
4. "Specific determiners"(i.e., unintended clues to the answer) have been avoided. ❑
5. Ambiguous terms like "frequently" and "in most cases", etc. have been avoided. ❑
6. Negative, and particularly double negative statements, have been avoided. ❑
7. The statement is short and uses simple language. ❑
8. There is no pattern in the order of responses. ❑
9. Approximately half the statements are "false" ❑
10. Trick questions have been avoided. ❑
11. Source materials have been paraphrased before being included on the test. ❑
    Otherwise encourages memorization without understanding.
    Statements may be ambiguous out of context
    Can lead to copyright problems
12. Directions to students are clear. ❑
13. Scoring method is clear to students (e.g., is guessing penalized?). ❑
14. Students circle "true" or false", rather than write "T" or "F". ❑
15. Students have been instructed to correct false statements: ❑

Sample true/false format:

	Items	Corrections
T F	1. Insects have eight legs.	
Sample student answers:		
T F	1. ~~Insects~~ have eight legs.	spiders
T F	1. Insects have ~~eight~~ legs.	six

Item Analysis in the University Classroom

© 2001 Robert Runté, Ph.D. Faculty of Education, University of Lethbridge Runte@uleth.ca

Although one will not want to take the time to conduct an item analysis for every quiz, there are a number of occasions when item analysis can be useful to the classroom teacher:

1. Where the examination accounts for a significant proportion of the students' mark.
   Item analysis will allow you to identify questions which were mis-keyed, which were much too difficult, which have two defensible answers, or which are now revealed to have been 'trick' questions. These questions can then be eliminated from the test before it is returned to students, or their grades calculated.

2. Where an examination covers material on which you require detailed diagnostic information.
   Item analysis will reveal not only those areas in which students are weak, but also the nature of the errors students are making.

3. Where items from an examination are to be used again.
   Writing good test items is extremely time consuming. Using item analysis statistics, even items that performed poorly can be revised and reused. Studies have shown that it is up to five times faster to revise items than to replace them with completely new ones (and freshly written replacement items would likely contain their own flaws until revised on the basis of item analysis statistics.)

4. Where one wishes to improve one's test writing skills.
   Every instructor should undertake at least the occasional item analysis to refine their question-writing skills. In our experience, instructors are constantly surprised at how much they can learn, and how quickly they can improve, based on the statistical feedback from just a few items. Item analysis should therefore be considered part of an instructor's continuing professional development.

5. Where one wishes to evaluate the effectiveness of program changes.
   Questions for which one has developed a strong statistical profile over a number of years provide normed standards against which the performance of the current class may be compared. Thus, if all other factors have been held constant, the effectiveness of a new textbook or teaching strategy may be evaluated by observing whether student performance on these items is above or below previously established norms.

All University of Lethbridge instructors have direct access to computer scoring equipment (e.g., scantron) and item analysis software, making item analysis a simple matter of having students use the correct computer-scanable answer sheets and feeding the results into the machinery.

It is nevertheless useful to know how to do an item analysis by hand, both because there may be occasions when you want to conduct an analysis after the test has already been written (without scan sheets) and because realizing one can do this by hand helps to demystify the whole process.

Eight Easy Steps For Classroom Item Analysis

[Adapted from Sax, *Principles of Educational and Psychological Measurement and Evaluation, 3rd Edition,* (Belmont, Ca: Wadsworth, 1989).]

Here, then, there are the eight steps for classroom item analysis, using only simple arithmetic.

1. Score each answer sheet and write the total score in the corner.

2. Sort the answer sheets into rank order from top to bottom.

3. Divide the pile in half. (If you have an odd number of students, set the "middle" student's paper aside until step 7, so that the top and bottom piles will have the exact same number of students. Where you have given the test to 100 or more students, divide the pile into thirds. Put the middle third aside, and treat the top and bottom thirds as the top and bottom groups in the instructions below.)

4. Using only the top half of the class, record the number of students who responded to each alternative. (That is, how many students chose "1-a", how many "1-b", and so on.) Repeat the procedure for the bottom half of the class, but keep the records for the two groups separate.

5. For each question, subtract the number of students in the lower group who got the item correct from the number of students in the upper group who got it correct. Repeat for all the test items.

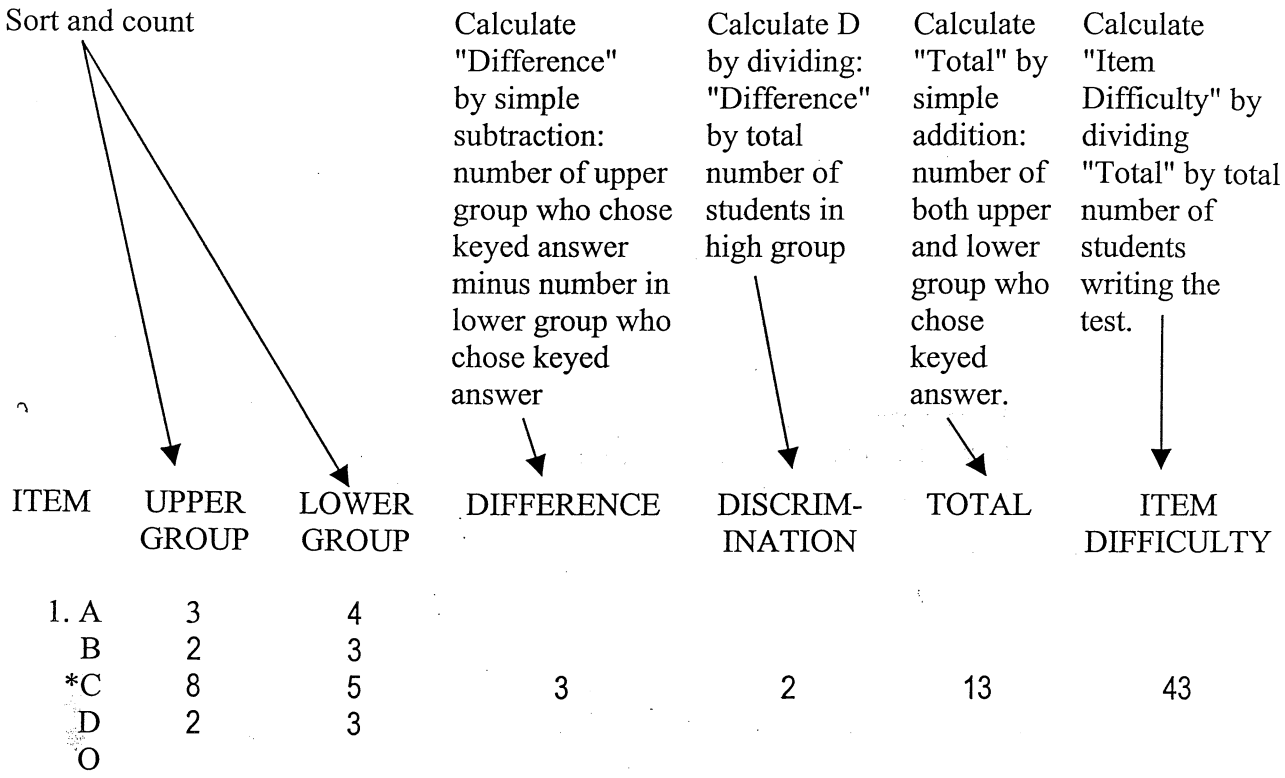
6. For each question, divide the difference between top and bottom groups obtained in step 5 by the number of students in the upper group.

7. For each question, total the number of students (both upper and lower group) who got it right. (For large samples of over 100 where you have been dealing with only the top and bottom thirds, include the middle third in *this* calculation. Similarly, if you had an odd number of students and had to put the middle student's paper aside for steps 3-6, count that paper back in at this point.)

8. For each question, divide the total from step 7 by the total number of students in the class or sample.

You should end up with something similar to the sample item analysis on the next page.

Sample Item Analysis for a Test, 30 students



ITEM	UPPER GROUP	LOWER GROUP	DIFFERENCE	DISCRIM-INATION	TOTAL	ITEM DIFFICULTY
1. A	3	4				
B	2	3				
*C	8	5	3	2	13	43
D	2	3				
O						

*=Keyed Answer O= OMIT (students who did not respond to the item)
Note, it is only necessary to calculate "difference", "discrimination", "total" and "item difficulty" for the keyed answer.

Interpreting Item Analysis Statistics To Revise Items

(Adapted from Sax, *Principles of Educational and Psychological Measurement and Evaluation, 3rd Edition,* [Belmont, Ca: Wadsworth, 1989.])

There a number of simple guidelines for revising items based on an analysis of their statistical performance.

1. <u>The ideal pattern for a multiple-choice item is one in which a majority of the high group selected the keyed answer, and the majority of low group were evenly distributed between the three distracters.</u>

   Because we selected the high and low groups based on their performance on this particular test, they are by definition the strongest or weakest students on these particular tasks. It is therefore possible for us to determine whether individual questions are working by seeing how the high and low groups performed on each item. We are not evaluating students here, but whether the question is carrying its fair share of the assessment process.

2. The higher the discrimination index (calculated in step 6) the better.
Remember, the object of a norm referenced test is to discriminate between levels of ability, so a good question is one which distinguishes well between the top and bottom groups. A negative discrimination index means one of two things:

(A) The skills tested by this question are unrelated to the skills tested by the majority of the other questions on the test. For example, a question on how to read a chart on a History examination may test quite different skills then the majority of the questions which focused on the students recall of historical events. We might not expect the top group in history to be the same group as for chart reading, and a negative discrimination on that item would therefore be understandable. Although ideally an examination should only test a single topic or set of skills, there may be occasions (as in this example) where the curriculum calls for quite diverse skills within a single subject.

(B) On the other hand, a negative discrimination may suggest that there is something wrong with the item itself. For example, there may be some ambiguity in the wording that is causing students in the high group to second guess the intent of the question, and so choose a wrong alternative, while the weaker students are missing this subtle/unintended qualification, and therefore choosing the correct answer. Such ambiguity should be edited out of the question.

3. Distracters should be more attractive to students in the low group than the high group.
If high group students consistently choose a wrong alternative, they may be reading something other than what was intended into this wrong answer.

4. Each distracter should attract at least 5% of the total group, or be replaced with something more plausible.

5. No distracter should be more attractive than the keyed answer.
If one is, reexamine the question to make sure the question was not mis-keyed. (If it was mis-keyed, rescore the tests before returning them to students.) If the alternative is clearly incorrect, then revise the item to make this distracter less plausible.

6. If an alternative is equally attractive to the high group as the keyed answer, there may be two defensible answers.
The item should be carefully reviewed, and if a case can be made for the popular alternative, student marks should be adjusted accordingly before the test is returned. If the alternative is clearly wrong, the question should be revised, either by weakening the popular alternative, or by clarifying the stem.

7. If student responses are equally distributed among all the alternatives (that is, including the keyed answer) it strongly suggests that students are guessing.

   This is an acceptable pattern for the low group, but if the high group show this pattern, the question is either so ambiguously worded that students were unable to identify what was being asked, or this material was not adequately covered in the course. An example of the latter might be where several teachers have agreed to share a common end of year examination, but this particular skill or fact was not taught in the class being analyzed. In that case, the question should be dropped for this class, and marks adjusted accordingly. Normally, however, random distribution among all the alternatives indicates a lack of clarity in the question.

8. Four choice items with a difficulty level of less than .35 should be avoided.
   Questions that are too difficult may frustrate and discourage students. Furthermore, such items do not discriminate well, are likely to violate rules 1-7, and are unreliable because they encourage guessing. A difficulty level of .25 on a four choice item, for example, suggests completely random guessing, and a correct score measures the student's "luckiness", rather than knowledge of the subject. Some difficult questions in the .35 to .45 range are acceptable to provide the best students with some challenge, provided that the items have acceptable discrimination indices (as calculated in step 6).

9. Four choice items with a difficulty level of higher than .85 should be avoided, unless addressing a curricular objective of particular import, or as part of a mastery test.
   Questions that are too easy are unlikely to discriminate well. Given that one always has more learning objectives to assess than the limited number of items which can be included on the test, asking something that nearly everyone will get correct is essentially a waste of a question. A few items in the .80 to .85 range may be placed at the beginning of the test, or strategically scattered across objectives, to encourage students and reduce test anxiety.

10. Where a number of students omitted a question, this may indicate problems with either the question or the test.
    If the question occurs near the end of the test, it may mean that these students failed to turn the last page to find the question (that is, a test formatting problem) or that the test is too long and some students were unable to get to this item. This need not reflect poorly on the item itself, which may be reused unchanged, but on a shorter or better laid out test. If the item occurs earlier in the test, it may again indicate a formatting problem (students missed it because it was buried in the midst of graphics, etc.) or that the question was baffling. By checking to see if the omits came from both top and bottom groups and by examining the difficulty level, it should be possible to determine if students were baffled because the item was too difficult, or because the question was poorly worded.

Item Banking

Once items have been analyzed and revised, they can be banked.

Item banking not only allows you to organize your items in a systematic fashion for easy retrieval and reuse, but actually allows you to build better tests. By tracking the statistical performance of banked items over several tests, monitoring the effect of each revision or refinement, one is able to gain valuable insights into which techniques work best for the particular course level, subject matter and client group, and thereby design increasingly sophisticated evaluation instruments.

For example, as you build up a bank of questions over the years, you can ensure that the average difficulty level of the questions for each cell on your table of specifications is roughly equivalent. This avoids the problem where, for example, an instructor gives ten easy questions on topic A and ten hard ones on topic B. Just looking at the test average the teacher may think the examination a fair one, not realizing that s/he is inadvertently discriminating against those students who would have scored much higher had the Topic B questions been easy and the Topic A questions difficult. The same principle applies to the various cognitive levels of Bloom's taxonomy.

Typical Output From Scan Sheets

```
MOST DIFFICULT QUESTIONS
========================

 Question:   28   Score:    0/19   Difficulty:  100.0%
 Question:   23   Score:    2/19   Difficulty:   89.5%
 Question:   21   Score:    4/19   Difficulty:   78.9%
 Question:   13   Score:    5/19   Difficulty:   73.7%
```

The "most difficult questions" printout is useful for quickly identifying mis-keys (#28 is almost certainly and #23 is likely) or questions that were too difficult because the concept was not taught this term (ran out of time, changed texts, etc.) or questions that are too poorly worded for students to correctly interpret.

```
ITEM ANALYSIS
                              =============

 Question    6)  Correct:     14     73.7%          Incorrect:      5     26.3%
                 (14 chose A)                       (2 chose B)
                                                    (3 chose C)

 Question    7)  Correct:      7     36.8%          Incorrect:     12     63.2%
                 (7 chose A)                        (12 chose D)
```

The "item analysis" table quickly shows you which items may have alternatives that are not working (no one chose "D" in question 6) or working *too* well (too many students chose "D "on Question 7 – when more students choose a wrong answer over the correct one, one should check whether the question is a mis-key, that the preferred "wrong" answer is not in fact defensible, or that students have misread or misunderstood a poorly written question).

Typical Output From Scan Sheets (continued)

TOTAL DISTRIBUTION TABLE

WARNING: Some of the data in this table represent percentage figures.
As fractional values have been truncated, percentages may not sum to 100.

				PERCENTAGE TOTALS						ABSOLUTE TOTALS					
QUESTION	LOW	KEY	MEAN	BL	A	B	C	D	E	BL	A	B	C	D	E
1		B	0.79	0	21	78	0	0	0	0	4	15	0	0	0
2		A	1.00	0	100	0	0	0	0	0	19	0	0	0	0
3		A	0.63	0	63	36	0	0	0	0	12	7	0	0	0
4		A	0.95	0	94	5	0	0	0	0	18	1	0	0	0
5		B	0.32	0	68	31	0	0	0	0	13	6	0	0	0
6		A	0.74	0	73	10	15	0	0	0	14	2	3	0	0
7		A	0.37	0	36	0	0	63	0	0	7	0	0	12	0
8		D	0.95	0	0	0	5	94	0	0	0	0	1	18	0
9		E	0.79	0	5	0	15	0	78	0	1	0	3	0	15

The "total distribution table" serves the same purpose as the item analysis table and is usually easier and faster to read and interpret. (Note, in this example, questions 1-5 are True/False, so can only have "A" or "B" as the answer.)

CHOICE DISTRIBUTION TABLE

Difficulty: The percent of the groups tested (top and bottom quartiles) that answered the item correctly.
Discrimination: A score which measures how well an item discriminates between the more knowledgeable (top quartile) and the less knowledgeable (bottom quartile). The index ranges from -1.0 to 1.0, where a value of -1.0 would indicate an item that was answered correctly by all of the bottom quartile and incorrectly by all of the top quartile. Scores of .40 and above indicate a very discriminating item.

				LOW 25% OF CLASS						MID 50% OF CLASS						HIGH 25% OF CLASS							
QUESTION	LOW	KEY	MEAN	BL	A	B	C	D	E	BL	A	B	C	D	E	BL	A	B	C	D	E	Diff.	Disc.
1		B	0.79	0	50	50	0	0	0	0	18	81	0	0	0	0	0	100	0	0	0	0.75	0.50
2		A	1.00	0	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	1.00	0.00
3		A	0.63	0	25	75	0	0	0	0	72	27	0	0	0	0	75	25	0	0	0	0.50	0.50
4		A	0.95	0	75	25	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0.88	0.25
5		B	0.32	0	100	0	0	0	0	0	72	27	0	0	0	0	25	75	0	0	0	0.38	0.75
6		A	0.74	0	75	0	25	0	0	0	63	18	18	0	0	0	100	0	0	0	0	0.88	0.25
7		A	0.37	0	25	0	0	75	0	0	27	0	0	72	0	0	75	0	0	25	0	0.50	0.50
8		D	0.95	0	0	0	0	100	0	0	0	0	9	90	0	0	0	0	0	100	0	1.00	0.00
9		E	0.79	0	0	0	25	0	75	0	9	0	18	0	72	0	0	0	0	0	100	0.88	0.25
10		D	0.47	0	0	25	25	25	25	0	9	9	36	45	0	0	0	25	0	75	0	0.50	0.50
11		B	0.74	0	25	75	0	0	0	0	18	72	0	0	9	0	25	75	0	0	0	0.75	0.00
12		A	0.95	0	75	25	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	0.88	0.25
13		A	0.26	0	25	0	50	25	0	0	9	9	18	63	0	0	75	0	0	25	0	0.50	0.50
14		A	1.00	0	100	0	0	0	0	0	100	0	0	0	0	0	100	0	0	0	0	1.00	0.00

The choice distribution table is the most useful table for item analysis because it includes item discrimination, and compares the performance of top and bottom groups). Look for positive discriminations of 0.01 and above. Questions 2, 8 and 14 do not discriminate between students and should be revised for or eliminated from future tests. (Note, in this example, questions 1-5 are True/False, so can only have "A" or "B" as the answer.)