What Those Numbers Mean

A Guide to Item Analysis

Ahmad Fuad bin Abdul Rahim

Medical Education Department

School of Medical Sciences

Universiti Sains Malaysia

First Publication January 2010

ISBN: 978-983-43381-8-3

Edited by

Dr. Ahmad Fuad bin Abdul Rahim
Medical Education Department
School of Medical Sciences
Universiti Sains Malaysia Health Campus
Kubang Kerian, Kelantan Darul Naim

Published by:

KKMED Publications
Medical Education Department
School of Medical Sciences
Universiti Sains Malaysia Health Campus
Kubang Kerian, Kelantan Darul Naim
Downloadable at
http://www.medic.usm.my/dme/index.php?option=com_content&task=view&id=54&Itemid=86

Published in Malaysia

Contents

List of Boxes

List of Tables and Figures

■ Introduction

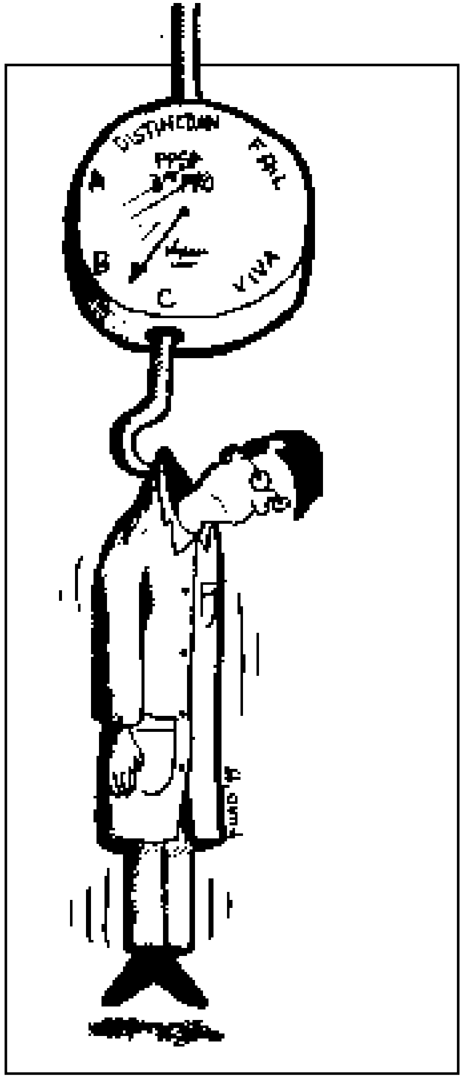
I have often been asked about “what those numbers mean”. Those numbers are the item analysis indices printed out by the computer in the Academic Section, PPSP, after every examination.

Now, I am as numerophobic as the next person. But at least, I tell myself, this is not statistics. So, to the best of my ability, I will try to make some sense out of these numbers, in the hope that this understanding will encourage us all to use these indices in improving our examinations, in turn improving the quality of education our students experience. For what greater task is there more worthy of our best, if not the education of our young generations?

This guide started off way back in 1998 as an in-house document used in our faculty development workshops. Many thanks to Professors Rogayah Jaafar, Yasmin Anum, Zalina Ismail and Hamiadji Tanuseputro for going through the initial drafts and for their encouragement and suggestions. My thanks also to Dr. Muhamad Saiful Bahri Yusoff for his encouragement and help to get this published.

Ahmad Fuad bin Abdul Rahim
Medical Education Department
December 2009

■ Chapter 1: Function of Item Analysis



An examination is, in some ways, like a weighing scale (See Box 1). Both are measurement tools. We use a weighing scale to measure, obviously, the weight of things. Like the extra adipose tissue we have on our bodies. We use an examination to measure, hopefully, a student's knowledge, skills or attitudes.

We often 'service' or calibrate our weighing scales, to make sure they are measuring accurately. We make sure, for instance, the spring is not too loose, making us heavier than we actually are. (The spring can be too tight, but we don't mind that, do we?) Similarly we 'calibrate' or improve our examination by looking at the data obtained from it. There is a lot that can be learned from examination data that could help us in improving our examination questions, especially from objective question items. This is called item analysis.

Box 1

How is an examination like a weighing scale?

As mentioned, both are measurement tools. In both cases, we use the tools to make decisions, like deciding, if one is underweight or overweight, pass or fail. Both tools can be accurate or inaccurate.

But there are differences too. For one thing, a weighing scale measures weight; something definite. You either weight 80 kilograms, 90 kilograms or some other weight, give or take a few grams.

An examination, on the other hand, measures something abstract. Knowledge. Skills. Attitude. Can we measure all the knowledge that a student has, and give it a value? Can we say student A has 87 skillograms of clinical skills, or student B measures 55 on the attitudometer? I don't think so; not yet, anyway. So, at best it is an approximation of what the student has.

What is the implication for us teachers? Given its approximate nature, compared to the decisions we sometimes have to make with it (Shall this student repeat another year? Can this student pass as a safe doctor?) the onus is on us to make it as accurate, valid and reliable as we can.

Item analysis has its roots in norm-referenced testing (Linn and Gronlund 1995 p.315, Ebel and Frisbie, 1991 p.221 and 225) (See Box 2). This fact is important to know because the qualities we are looking for in the question differs depending on whether we are running a norm-referenced examination or a criterion – referenced one. However, both types of examinations can still benefit from item analysis in that it can help to identify "... faulty items and can provide information about student misconceptions and topics that need additional work". (Linn and Gronlund 1995 p.315).

Although in doing item analysis we look at data concerning the overall performance of students in one examination, the information we get refers to each question, or item in the examination.

There are four questions that can be answered by item analysis:

1) Did the item function as intended?

2) Were the test items of appropriate difficulty?

3) Were the test items free of irrelevant clues and other defects?

4) Was each of the distracters effective (in multiple-choice items)?

All questions except for number two are relevant for future test construction in both a norm- and criterion-referenced situation. Question two, however, is more suited for a norm-referenced situation. We will see why later.

Apart from the benefits mentioned, there are other benefits of item analysis:

- It provides a guide for the effective discussion with students after the examination.
- It can guide remedial work to be done on students.
- It can guide general improvement in instruction.
- It increases the educator's skill in test construction

Box 2

Norm-referenced examinations and Criterion-referenced examinations

The value of the examination results comes when it is compared to a standard. In norm-referenced examinations, the results a student obtains is compared to the results of his peers in the same examination. The main principles is that "…one's peers are used to set the standards for the assessment of comparative ability and relative attainment". It usually results in a fixed percentage of students who fail or excel, and is useful for comparison and ranking of students.

In criterion-referenced examinations, students are compared to set criteria. The basic principle is that "….the student's absolute performance is assessed which is not relative to his or her peers but to some pre-set criterion determined by the faculty".

Norm-referenced examination has definite limitations, at least where medical education is concerned. For a discussion on this issue, please refer to Turnbull, J.M., 1989, from which the material in this box is taken from.

■ Chapter 2: The Components of Item Analysis

The Difficulty Index The difficulty index for an item is commonly defined as the percentage of students who gets the item right (Linn and Gronlund 1995 p.320). As such it is sometimes called the facility index (Dixon 1994). Some institutions define it as the percentage of students who gets the item wrong. The former definition is more common but the most important thing is to make sure how the index is defined (See box 3). Pay attention, folks: the percentage refers to the percentage of students from the total of the lower and upper groups only. The assumption is that the middle group students follow the same pattern (Linn and Gronlund 1995 p.320). For multiple true-false items some authors define the difficulty index for the whole question as the mean marks for that question (Fleming 1984). Looking at the definition, we have to realize that the index is not determined by the content alone but also by the students who attempted them. In fact, this is true for all the indices in item analysis (Ebel and Frisbie 1991 p.228).

As mentioned previously the difficulty index is useful in norm-referenced examinations. To obtain the maximum distinction between the high and low-achieving students, it is preferred that items be of middle difficulty so that the score distribution spreads out (Ebel and Frisbie 1991 p.231, Linn and Gronlund 1995 p 315). Items of low or high difficulty will not be selected for further use.

In criterion-referenced examinations a question's difficulty is related to the criteria that it is measuring. If the thing that the student is required to be able to do (the criteria) is difficult, then the question measuring it will be difficult. If the criteria happens to be easy, then the question will be easy (Ebel and Frisbie 1991 p.223). The difficulty index, in this criterion-referenced situation, is used to identify too-difficult questions (Ebel and Frisbie 1991 p.237). Possible causes of too-difficult items include a wrong answer key, more than one answer for an item, a question on rare or trivial areas, the problem not stated clearly or the item positioned at the end of the test so that many students cannot attempt it (Cox 1976). Even then, if none of these causes is thought to be responsible for making the item too difficult, and the questions well made, and agreed to test an important area, the question should be accepted as it is. It should be pointed out, too, that items with low indices of difficulty are not automatically good. They should be scrutinized to see for possible causes such as weak choices or requiring only a low level of understanding. In other words, judgement is still required on the part of the examiner. In accordance with this, Dixon (1994) advocates the use of the discrimination index for weeding out both too-easy and too-difficult items and gives a target range of 20-80%.

Box 3

The Difficulty Index

As defined, the difficulty index is the percentage. Using the former definition, the smallest value is a zero. That means none of the students who has attempted that particular question, or that particular alternative in a multiple true-false MCQ, got it right. Tough cookie. The highest value, no prizes for guessing this one right, is a hundred. That means everybody got it right.

In some places, instead of percentage, the index is given in decimals. So we would see values ranging from zero to one. The principle remains: low values, difficult question. High values, easy question.

The Discrimination Index The idea behind discrimination is, if you use a measuring tool, you should be able to discriminate between the things that you measure. Remember our weighing scale? You would expect that it would be able to indicate the weight of a person accurately; an overweight person should have a higher value. If everyone using the weighing scale gets a reading of 50 kilograms, you would throw it in the trash and buy a new one.

Similarly, for examinations you would expect it to discriminate the students. The discrimination index of an item is defined as "...the degree to which it discriminates between students of high and low achievement". (Linn and Gronlund 1995 p.321). In other words the question should be able to pick out the 'good' students from the 'bad' ones; more 'good' students will be able to answer the item, less 'bad' students will be able to answer it.

If you look at it, this seems to be a norm-referenced concept. We are trying our best to find questions which can help us to rank out, or discriminate between high-achieving and low-achieving students. Actually, in criterion-referenced examinations, the discrimination index is still useful. "The items in a criterion-referenced test should also discriminate between students as long as some students have not learned the content measured by those items" (Ebel and Frisbie 1991 p.224). The difference between these two is that in norm-referenced examinations, items that fail to discriminate are regarded as poor items. In criterion-referenced examinations, it is not necessarily so. More of this later.

The discrimination index is calculated in several ways. One way is by calculating the difference in the number of students in the upper levels who got the item right as compared to the number of students in the lower levels who got it right, after the students have been ranked (Linn and Gronlund 1995 p. 325, Ebel and Frisbie 1991 p.225) (See Box 4).

Box 4

Calculating the Discrimination Index

Because the 'good versus bad students' method is used here in PPSP, I will try to explain the procedure in more detail.

1. Students are ranked according to their overall results. Let us say that there are a hundred students who took the MCQ 1 paper in the recent examination. We look at the results and rank the students from number one (who got the highest score) to the last (who got the least).
2. We calculate how many students there are (actually, the computer does all the work) in 27 percent of the student number, that is 27 students out of a hundred in this case. Let us call this number A.
   Why 27 percent? It seems that this is the optimum compromise between the two requirements: that the two groups have as many students as possible and that they are as far apart as possible (Ebel and Frisbie 1991 p.227)
3. We identify the 'good students group' by taking the top 27 students, and the 'poor students group' by taking the last 27.
4. For a particular question or a particular branch in a five response MCQ, the discrimination index in calculated as (pay attention here, folks!) the difference between the number of students in the 'good students group' who got that question correct, X, (let's say 25 students) and the number of students in the 'poor students group' who got it correct, Y, (let's say 16) divided by A (look at step 2 if you forgot what A is).
   That is (25-16)/27, or 9/27, or 0.33.

For the multiple true-false type of question other methods of calculating the index have been described. This includes using the Pearson's product-moment correlation coefficient for the whole question and using the point-biserial correlation coefficient or the phi coefficient for the individual items (Fleming 1984). In using correlation coefficients, we expect good questions to have a positive correlation with students' results. Those who did well overall, should do well with the question. Poorly designed questions will have low or even negative correlation (See Box 5).

The discrimination index is affected by sampling error (Ebel and Frisbie 1991 p.231) and therefore the higher the student number the more reliable it is. However this does not mean that it does not have any value where small student numbers are concerned. They are still useful for overall test improvement.

The index of discrimination is used in selecting and revising items (Ebel and Frisbie 1991 p232). For item selection the higher the index the better. Table 1 is a useful guide.

For item revision we are looking for poorly discriminating items in the norm-referenced context (Ebel and Frisbie 1991 pp.233). Possible causes of low discrimination include a too-easy or too-difficult item (Dixon 1994) (See Box 6). In the criterion-referenced situation many good items have a low index of discrimination, even zero (Ebel and Frisbie 1991 pp.237). This is because scores in the criterion-referenced examinations tend to be negatively skewed and have low variability. In other words, judgement is still required when looking at the indices and doing item revision. However, no item, in both criterion- and norm-referenced examinations, is useful if the index is negative and such an item is subject to scrutiny.

Table 1 Guideline for Using the Discrimination Index in Item Analysis (Ebel and Frisbie 1991 p.232)

Index of discrimination	Item evaluation
0.40 and up	Very good items
0.30 to 0.39	Reasonably good but possibly subject to improvement
0.20 to 0.29	Marginal items, subject to improvement
0.19 or less	Poor items, to be rejected or improved by revision

Lastly it is good if we can remember that item discriminating power does not indicate item validity. The index uses internal criterion; it measures whatever the test is measuring regardless of it's validity (Linn and Gronlund 1995 pp.325).

Box 5

Discrimination Index

As described, the discrimination index is a measure, of how the 'good' students are doing versus the 'poor' students on a particular question. Knowing this, we expect the value of the discrimination index to range between 1 (all 'good' students correct versus no 'poor' students correct by the former method, or the maximum value for a positive correlation by the latter method) to –1 (figure it out yourself). The acceptable values are given in Table 1.

Box 6

Discrimination and Difficulty Index: The Relationship

Here were see the relation between the two indices: very difficult or very easy questions are not able to discriminate between students (Ebel and Fresbie 1991 pp.223-224). This fact is important in norm-referenced examinations, less so in criterion-referenced ones.

Other indices Dixon (1994) wrote about other indices that can be useful in the evaluation of items. One is the *Correctness index*, defined as the percentage of those candidates recording either a true or false response for a particular branch in a multiple true-false response MCQ who gave the correct response. The target range is 20-80%; a low index may mean that students are attempting the item but are getting it wrong. *The Branch Popularity index* is determined as the percentage of all candidates sitting the examination who record either a true or false response for this branch; this can be used to identify an item that is avoided by students for a particular reason. The target range is 80-100%. The popularity may be related to the stem, which can be determined by the *Stem Popularity index*, defined as the percentage of all candidates sitting the examination who record either a true or false response for at least one branch of this stem. The target range is 90-100%. All of these indices can be determined if we are provided with the number of students who got the item correct, the number of students who get the item incorrect and the number of students who did not attempt the item.

> *"Examinations are formidable even to the best prepared, for the greatest fool may ask more than the wisest man can answer."*
>
> *Charles Colton*
> *1780-1832*

■ Chapter 3: Going Through the Item Analysis: Item Revision

The examination is over. The results are out. We have the computer printout of the item analyses in our hands. All those numbers. Where do we begin? How do we begin? How can item analysis help us?

As mentioned in the first section, we use item analysis to help us in:

1. identifying good questions to be included in future examinations

2. identifying poorly performing questions for improvement

3. identifying areas learned poorly by students for future improvement in instruction and feedback to students

The way we use item analysis also depends on the way we want to interpret the results of that particular examination. Do we want to pick out the good students from the bad ones (e.g. for selection, grouping students) or do we want to see if they can perform up to a certain predetermined standard (e.g. in professional courses, as in PPSP)?

This brings us back to the issue of Norm-referenced (NR) or Criterion-referenced (CR) examination (See Box 2)

What has this got to do with item analysis, you ask? Depending on whether the examination is NR or CR, we use the item analysis to pick out MCQs of different qualities. It means that a 'good' MCQ for NR is not necessarily 'good' for CR.

Put simply, in NR examinations we want questions of middle difficulty (around 0.5) because we want the normal curve to be, well, as normal as possible. We also want the questions to have as high a discrimination index as possible.

It's quite different in CR examinations. We still expect questions to discriminate, but we look at the discrimination indices in a different light. In CR examinations, questions with a high discrimination index means the question have discriminated students who have learned the content from those who have not. That is useful, and good. But those with a low discrimination index does not necessarily mean that the item is not useful, or not good. It may mean that ALL of the students tested, or the majority of them, have learned the content, which is also what we want. About the only condition in NR examinations when the discrimination index should make us say "uh-oh, let's look at this one" is when they are negative. It means that the 'good' students are getting the item wrong, and the 'bad' ones right. That is cause for concern in both NR and CR examinations.

Before we go on the 'how' of item analysis, a word of caution. It is very possible to blow up the importance of the item analysis out of proportion. As you have seen, judgement is very much required in attaching meaning, or weight, to the indices. They are not sacred, incontestable values.

■ Chapter 4: Item Analysis: A Flowchart to Help You

After the preliminary discussion, I offer here a flowchart that you can use to look at the analysis. This is only a suggested way of looking at things, you may have your own way.

Mind, too, that although the flowchart suggests that you start with the discrimination index, it does not necessarily mean that the discrimination index is more important than the difficulty index. We look at it first because low discrimination can be caused by a question being too difficult or too easy (see Box 6) , so it is logical that we look at the cause after we find a reason to do so.

The flowchart is also designed with the type X MCQs (Multiple True-false) in mind. For the type A MCQs (Single Best Answer) the item analysis is slightly different. More of that in the next chapter.

Item analysis is also very much a team effort, preferably departmental. We need content experts to consider the possible explanations for the item analysis, and perhaps input from those doing the teaching to take full advantage of the information.

Have fun!

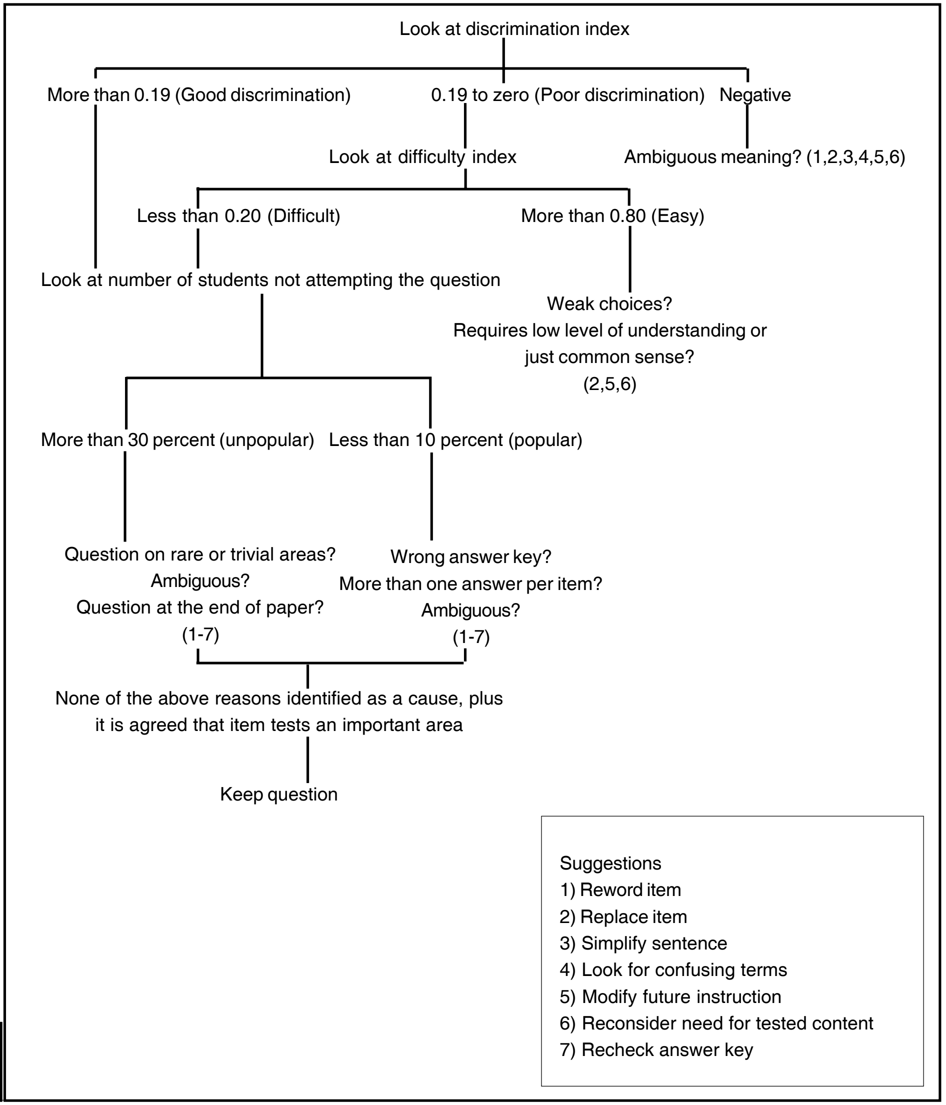

Look at discrimination index
More than 0.19 (Good discrimination)
0.19 to zero (Poor discrimination)
Negative
Look at difficulty index
Ambiguous meaning? (1,2,3,4,5,6)
Less than 0.20 (Difficult)
More than 0.80 (Easy)
Look at number of students not attempting the question
Weak choices?
Requires low level of understanding or
just common sense?
(2,5,6)
More than 30 percent (unpopular)
Less than 10 percent (popular)
Question on rare or trivial areas?
Ambiguous?
Question at the end of paper?
(1-7)
Wrong answer key?
More than one answer per item?
Ambiguous?
(1-7)
None of the above reasons identified as a cause, plus
it is agreed that item tests an important area
Keep question
Suggestions
1) Reword item
2) Replace item
3) Simplify sentence
4) Look for confusing terms
5) Modify future instruction
6) Reconsider need for tested content
7) Recheck answer key

Figure 1 Flow chart for using item analysis in item revision

■ Chapter 5: Item Analysis for Type A MCQs

For type A MCQs, the approach is a bit different although the main concepts explained in the previous chapters remains the same. To make things easier, we will refer our discussion to a set of hypothetical item analysis data of a hypothetical type A MCQ (refer Figure 2).

	Option	A	B*	C	D	E
Group						
Upper		5	40	4	6	4
Lower		8	11	17	20	3

Difficulty Index: 0.43 Discrimination Index: 0.49

Figure 2 A Hypothetical Type A MCQ With Its Hypothetical Indices

Difficulty index Because for type A MCQs there is only one correct answer, the difficulty index refers to the performance of the students on that correct option. In figure 2 the correct option, or the key, is option B. 51 students, 40 from the upper group and 11 from the lower, chose the correct option out of a total of 118 students (59 in the upper group and 59 from the lower). So the Difficulty Index for the question is 0.43 (51 divided by 118).

Discrimination Index For the same reason the Discrimination Index for the question is considered to be the discrimination on the correct option. In figure 1 40 students from the upper group chose the correct option while 11 students chose it from the lower group. The Discrimination Index is thus 40-11=29 divided by 59 (the number of students in a group) giving us 0.49.

Aim of Item Analysis Referring back to our earlier discussion, there are four questions that can be answered by item analysis, which are

1. Did the item function as intended?
2. Were the test items of appropriate difficulty?
3. Were the test items free of irrelevant clues and other defects?
4. Was each of the distracters effective (in multiple-choice items)?

Questions 1, 2 and 3 applies both to types X and A MCQs. Question 4, however, applies to type A MCQs and is another different feature in the item analysis of type A MCQs.

In type A MCQs, the function of the options other than the correct one is to present to the students seemingly plausible answers to the question, alternatives that might seem attractive due to the students misunderstanding or lack of knowledge. That is why they are called distracters. In this aspect item analysis can tell us whether they have been effective in doing their function, that is attracting the attention of the students, expectedly those in the lower group, to choose them as the correct answer. Some authors recommend that attention be given to the pattern of responses rather than the difficulty and discrimination indices (Case, Susan M. and Swanson, David B., 1998, p.107).

Examples To make the point clear there is nothing like giving examples. All the examples are modified from an excellent handbook and resource for both the construction of MCQs and item analysis, Constructing Written Test Questions for the Basic and Clinical Sciences, Second Edition (1998) by Susan M. Case and David B. Swanson, National Board of Medical Examiners, Philadelphia, USA. This handbook is downloadable free from the internet at http://www.nbme.org/.

For all the examples, it is assumed that there are a hundred students in total. Therefore there are 27 students in each the upper and the lower group. Lets look at the first one.

Option	A	B*	C	D	E
Group					
Upper	1	1	20	3	2
Lower	5	4	9	7	2

Difficulty Index: 0.19 Discrimination Index: -0.1

Looking at the discrimination index will set our alarm bells ringing, as we see this question has a discrimination of -0.1. It is also very difficult, looking at the difficulty index. When we look at the pattern of responses, this might well be due to a miskeyed item. The correct answer looks like C, but again, here, as in any item analysis situation, a content expert has to look at the question to make sure. If the key is indeed C, then the discrimination index becomes 0.41 (can you work it out?) and the difficulty index becomes 0.53; that's very nice and the question does not need any rewriting.

Second example:

Option	A	B	C*	D	E
Group					
Upper	0	1	20	3	3
Lower	0	1	10	8	8

Difficulty Index: 0.56 Discrimination Index: 0.37

This question has good indices and suitable for reuse. Options A and B, however, will benefit from rewriting because few students, from both groups, selected it as the answer. Perhaps it is obviously wrong and therefore is not doing its job of distracting the students well.

On to the third example:

Option	A	B	C*	D	E
Group					
Upper	10	3	9	2	3
Lower	5	6	2	7	7

Difficulty Index: 0.20 Discrimination Index: 0.26

Nine students from the upper group and two students from the lower group selected the correct answer. This is a very difficult item plus having a 'bad' response. Observe that many of the upper group students are misled by option A; the item may be poorly worded. It is worth checking again: is it a fair option? Is it clearly worded? Is it equally correct?

Last one:

Option	A	B	C*	D	E
Group					
Upper	5	5	9	6	2
Lower	5	6	2	7	7

Difficulty Index: 0.20 Discrimination Index: 0.26

This item has an identical breakdown for option C as the previous example; it has the same difficulty and discrimination index. However, looking at the pattern of responses, this item may be acceptable because those who don't know the answer in the upper group are spread out evenly among the distracters. Scrutiny is still needed for options A, B and D to ensure they are clearly worded, correct and unambiguous.

Reference list
(These make good reading too)

1. Case, Susan M. and Swanson, David B., Constructing Written Test Questions For the Basic and Clinical Sciences, 2nd Edition, 1998, National Board of Medical Examiners, Philadelphia

2. Cox, K.R., Quality Control in the Part I F.R.A.C.S. Examination, 1976, The Australian and New Zealand Journal of Surgery, (3), August, 46, pp. 269-277

3. Dixon, R.A., Evaluating and Improving Multiple-Choice Papers: True-false Questions in Public Health Medicine, Medical Education, 1994, 28,pp. 400-408

4. Ebel, R.L. and Fresbie, D.A., Essentials of Educational Measurement, 5th Edition, 1991, Prentice-Hall, New Jersey

5. Fleming, P.R., The Administration of a Multiple-choice Question Bank, 1984, Medical Education, 18, pp. 372-376

6. Linn, R.L. and Gronlund, N.E., Measurement and Assessment in Teaching, 7th Edition, 1995, Prentice-Hall, New Jersey

7. Tumbull, J.M., What is. Normative versus Criterion-referenced Assessment, Medical Teacher, Vol. 11, No. 2, 1989, pp. 145-150.

Invitation

How do you find this monograph? Easy to understand? Needs some modifications? Something to be added? I would like your comments please.

You can e-mail me at fuad@kb.usm.my.