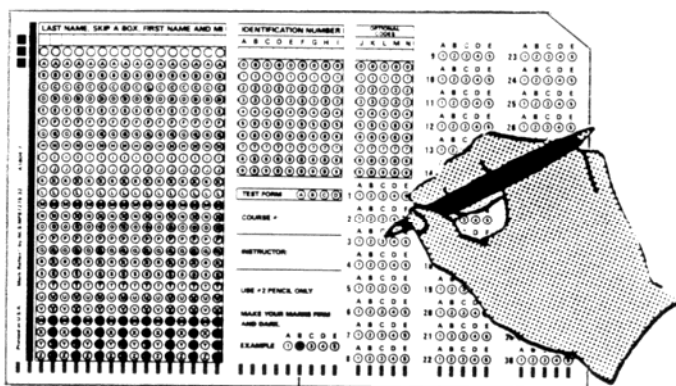


# Preparing and Evaluating Essay Test Questions

Technical Bulletin #36



Evaluation and Examination Service  
The University of Iowa  
(319) 335-0356

## PREPARING AND EVALUATING ESSAY TEST QUESTIONS

The purpose of this bulletin is to provide guidelines regarding the preparation of essay test items. This bulletin covers the relative characteristics of essay and objective test items and suggestions on when to use essay items. Guidelines for scoring and for analyzing the quality of essay test items are also included.

### COMPARISON OF ESSAY AND OBJECTIVE TEST ITEMS

There are several similarities between essay and objective test items. First, either type can be used to measure almost any important educational achievement.

Second, both essay and objective test items can be used to encourage students to develop important cognitive outcomes, understanding of principles, organization and integration of ideas, and application of knowledge. Neither test item readily encourages rote learning. However, the user of either type of test item may intentionally or unknowingly lead to tests that require mere memorization of facts, details, or lists. What a test measures is more a function of what the item writer writes than it is a function of the item format.

Third, the use of either item type involves the exercise of subjective judgment. Decisions about what type of items to use, how to word items, and how many items to use are common test development procedures. The subjectivity generally associated with essay test items stems from the judgments required by the scoring procedures.

Finally, in both cases the value of the obtained test from an essay or objective tests depends on item objectivity and reliability. Test scores that are independently verifiable by another scorer are said to be objective. A lack of objectivity is rarely a problem with objective tests, but it is frequently a serious limitation of essay tests. It is possible to obtain highly reliable, accurate scores, usually with considerable effort, from both types of tests.

Despite these similarities there are good reasons for deciding to use one item type rather than the other in particular testing circumstances. These are some of the differences often taken into account in making the choice between essay and objective test items:

1. An essay question requires students to plan their own answers and to express them in their own words. An objective item requires examinees to choose among several designated alternatives.
2. Students spend most of their time thinking and writing when taking an essay test. They spend most of their time thinking and reading when taking an objective test.
3. With objective test items the student's task, and the basis on which the examiner will judge the degree to which it has been accomplished, are stated more clearly than they are in essay items.

4. An essay test consists of relatively few, more general questions that call for rather extended answers. An objective test ordinarily consists of many rather specific questions requiring only brief answers.
5. An essay examination is relatively easy to prepare but rather tedious and difficult to score accurately. A good objective examination is relatively tedious and difficult to prepare but comparatively easy to score.
6. An essay test affords students a great deal of freedom to express their individuality in the answers they give and freedom for the examiner to be guided by his or her own preferences in scoring the answer. An objective exam affords freedom for the test constructor to express personal knowledge and value but limits the freedom for students to show how much or how little they know or can do.
7. The quality of an objective test is determined largely by the skill of the test constructor. The quality of an essay test is determined largely by the skill of the test scorer.
8. An objective test permits, and occasionally encourages, guessing. An essay test permits, and occasionally encourages, bluffing.
9. The distribution of numerical scores obtained from an essay examination can be controlled to a considerable degree by the grader; the distribution of scores from an objective examination is determined almost entirely by the test itself.

Whether to use essay or objective test items can be decided by weighing the factors cited above. However, the personal preferences, skills, and experience of the test constructor, independent of these other factors, often carry great weight in the decision.

#### WHEN TO USE ESSAY TESTS

The suggestions below are based largely on factors identified in the previous section. Essay tests are recommended for measuring educational achievement when:

1. The group to be tested is small, and the test items will not be reused with another group.  
An essay test is efficient to prepare and score when the number of examinees is small. As that number increases, however, efficiency favors the objective form. On the second point, the security of a small number of items is more difficult to maintain than for a larger number of items.
2. The instructor wants to encourage the development of student's skill in written expression.
3. The instructor is more interested in exploring student attitudes than in measuring achievements.

Whether instructors should be more interested in attitudes than achievements, and whether they should expect honest expressions of attitudes in a test situation seems open to question.

4. The instructor is more confident of his or her proficiency as a critical reader than as an imaginative writer of good objective test items.

The skills required for these two tasks are quite different. We should not expect to find them in equal abundance among all test constructors. For this reason, the experiences and preferences of an instructor often dictate which type of test will be used.

5. Time available for test preparation is shorter than time available for test scoring.

Of course instructors who have been able to build a large bank or pool of objective test items over time need be less concerned about test preparation time than those who start from scratch. For some courses, time available is a function of the number of teaching assistants on hand for essay test scoring.

#### GUIDELINES FOR WRITING ESSAY ITEMS

1. Ask questions or establish tasks that will require the student to demonstrate command of essential knowledge.

This means that students should not be asked merely to reproduce material heard in a lecture or read in a textbook. To "demonstrate command" requires that the question be somewhat novel or new. The substance of the question should be essential knowledge rather than trivia that might be a good boardgame question.

2. Ask questions that are determinate, in the sense that experts (colleagues in the field) could agree that one answer is better than another.

Questions that contain phrases such as "What do you think..." or "What is your opinion about..." are indeterminate. They can be used as a medium for assessing skill in written expression, but because they have no clearly right or wrong answer, they are useless for measuring other aspects of achievement.

3. Define the examinee's task as completely and specifically as possible without interfering with the measurement process itself.

It is possible to word an essay item so precisely that there is one and only one very brief answer to it. The imposition of such rigid bounds on the response is more limiting than it is helpful. Examinees do need guidance, however, to judge how extensive their response must be to be considered complete and accurate.

4. Generally give preference to specific questions that can be answered briefly.

The more questions used, the better the test constructor can sample the domain of knowledge covered by the test. And the more responses available for scoring, the more accurate the total test scores are likely to be. In addition, brief responses can be scored more quickly and more accurately than long, extended responses, even when there are fewer of the latter type.

5. Use enough items to sample the relevant content domain adequately, but not so many that students do not have sufficient time to plan, develop, and review their responses.

Some instructors use essay tests rather than one of the objective types because they want to encourage and provide practice in written expression. However, when time pressures become great, the essay test is one of the most unrealistic and negative writing experiences to which students can be exposed. Often there is no time for editing, for rereading, or for checking spelling. Planning time is shortchanged so that writing time will not be. There are few, if any, real writing tasks that require such conditions. And there are few writing experiences that discourage the use of good writing habits as much as essay testing does.

6. Avoid giving examinees a choice among optional questions unless special circumstances make such options necessary.

The use of optional items destroys the strict comparability between student scores because not all students actually take the same test. Student A may have answered items 1-3 and Student B may have answered 3-5. In these circumstances the variability of scores is likely to be quite small because students were able to respond to items they knew more about and ignore items with which they were unfamiliar. This reduced variability contributes to reduced test score reliability. That is, we are less able to identify individual differences in achievement when the test scores form a very homogeneous distribution. In sum, optional items restrict score comparability between students and contribute to low score reliability due to reduced test score variability.

7. Test the question by writing an ideal answer to it.

An ideal response is needed eventually to score the responses. If it is prepared early, it permits a check on the wording of the item, the level of completeness required for an ideal response, and the amount of time required to furnish a suitable response. It even allows the item writer to determine if there is any "correct" response to the question.

8. Specify the time allotment for each item and/or specify the maximum number of points to be awarded for the "best" answer to the question.

Both pieces of information provide guidance to the examinee about the depth of response expected by the item writer. They also represent legitimate pieces of information a student can use to decide which of several items should be omitted when time begins to run out. Often the number of points attached to the item reflects the number of essential parts to the ideal response. Of course if a definite number of essential parts can be determined, that number should be indicated as part of the question.

9. Divide a question into separate components when there are obvious multiple questions or pieces to the intended responses.

The use of parts helps examinees organizationally and, hence, makes the process more efficient. It also makes the grading process easier because it encourages organization in the responses. Finally, if multiple questions are not identified, some examinees may inadvertently omit some parts, especially when time constraints are great.

#### SOME EXAMPLES

**1A. What is feudalism?**

The bonds on a reasonable response to this question are not apparent. To say that "feudalism is a form of government used by the nobility in the middle ages" gives a correct answer to the question, but is it enough?

**1B. Under feudalism, describe how a typical baron would be considered a nobleman, knight, vassal, and lord all at the same time.**

The examinee's task is more straightforward in the second item; there should be less doubt about what is required for a complete answer.

**1C. What political, economic, and military relationship existed between a king, a lord, and the serfs of a fiefdom in a typical middle ages kingdom?**

This item asks a different question from #1B but illustrates how general item #1A is. The ideal response to items 1B and 1C might be legitimate responses to item 1A as well.

**2A. How do you think badminton compares with tennis?**

This item illustrates an indeterminate question because it solicits the respondent's opinion. Clearly this item does not require the examinee to demonstrate command of important knowledge. A better alternative would be to ask about the similarities and differences between the two games in terms of specified criteria.

**2B. Describe the similarities and differences between badminton and tennis in terms of equipment, basic rules, and strategy for winning.**

**3A. What does an automobile transmission transmit?**

An item like this usually reflects a lack of depth on the part of the item writer in deciding what to test. "Power" is a simple correct answer to this question. A better question could be written if the set of related propositions to be included in the correct answer are delineated by the item writer in advance of writing the item. Here are some possibilities.

- a. The transmission sends power from the engine to the driveshaft.
- b. Engine power consists of speed and torque (twisting power).
- c. The transmission can control the ratio of speed to torque.

**3B. Describe how the operation of an automobile transmission functions differently in these two situations:**

- (1) The car is beginning to move away from a stop sign.**
- (2) The car is traveling down a hill at about 55 mph.**

## GRADING ESSAYS

The main purpose of this section is to describe common methods for scoring essay test responses. The guidelines provided are designed to enhance rater reliability, or in other words, to optimize agreement among different readers about the score to be assigned to any given essay response.

Reliability is the characteristic of a set of measurements that relates to the accuracy of the scores. For essay tests it is necessary to distinguish between three types of reliability: scorer, examinee, and examination. Each of these types of reliability emphasizes or accounts for certain types of errors. For example, scorer reliability estimates indicate the extent to which the ratings of one grader would be replicated by another. When scorers are in agreement about the numbers to be assigned to each of several essay responses, the scorer reliability is high. Readers of the same response might disagree about the score to be assigned for any of several reasons:

1. The interpretation of the essay question differs.
2. Their perception of an "ideal response" differs.
3. How they weigh such factors as grammar, spelling, and organization differs.
4. They have different biases about the examinee based on past test performance.
5. They have different standards of quality for each of the available rating categories.

Examinee reliability has to do with the consistency of a test taker's performance on different occasions. If an examinee's test performance is typical--not influenced unduly by poor health, mental or physical fatigue, or temporary memory fluctuations--he or she should be able to repeat that performance almost exactly. To the extent that the performance of a group of examinees is consistent on the same tasks on different occasions, examinee reliability is said to be high.

Finally, examination reliability has to do with the accuracy of test scores as indicators of scores examinees would receive on the entire population of test items or test tasks. Would examinees obtain the same scores, or would their scores be in the same rank order, if a similar but different set of test questions were to be used? If the test items are a large and representative sample of all of the possible items that could be used, the examination score reliability likely will be satisfactory. Many other factors affect test score reliability other than the sampling of test content.

The suggestions presented in this bulletin are intended to help the user attain a reasonably high level of reader or scorer reliability. In each case the goal is to gain control over factors that can contribute to errors in scoring essay examination responses.

## SUGGESTIONS FOR GRADING ESSAYS

### 1. Chose either the analytical or holistic (global-quality) method.

In analytical scoring, essential parts of an ideal response are identified and the assigned score is based upon the number of parts included in the response. Each part is evaluated individually. This scoring method requires that the instructor develop an ideal response and create a scoring key or guide. The scoring key provides an absolute standard for determining the total points awarded for a response. Student responses are compared to the scoring standard and not to the responses of their classmates.

In holistic scoring the reader forms an impression of the overall quality of a response and then transforms that impression into a score or grade. The score represents the quality of a response in relation to a relative standard such as other students in the class (rank order) or an absolute standard like a set of sample papers with predetermined scores.

Holistic scoring tends to be simpler and quicker than analytical scoring, and in some situations, may yield more reliable scores. A disadvantage to holistic scoring is that it does not provide a clear justification of the assigned scores, nor does it give any indication of how a student's score may have fallen short of the standard.

### 2. Score the responses question-by-question rather than student-by-student.

Answers to one question on all students' papers should be read before going on to the next question. This procedure is required for holistic scoring, but it is advantageous for analytical scoring as well. Students' scores on one question should not be allowed to influence how subsequent responses are scored. Question-by-question scoring helps to reduce extraneous influences. Papers should be shuffled following the scoring of each question, so that order in the stack will not result in a scoring bias based on previous questions.

### 3. Disassociate the identity of students from their responses during the grading process.

No matter how fair an instructor may hope to be, if he or she knows the name of the student whose paper is being graded, it will be extremely difficult to ignore other information, related or unrelated, in scoring that paper. Such influences can be reduced by using number codes, rather than names, as identifiers on essay papers. Responses to different questions by the same student can be disassociated by requiring the use of separate sheets of paper for each essay question.

4. Regardless of the scoring method used, determine in advance what aspects of the response will or will not be judged in scoring.

The starting point for using the analytical approach to scoring is identifying the crucial elements of an ideal response. Should these elements include handwriting quality, spelling ability, and proper use of grammar and punctuation? Ordinarily the scorer might point out problems related to any of these areas, but the essay score should not be influenced positively or negatively by them. Except in courses on writing, language, and linguistics, the mechanics of written expression are not the focus of instruction and, consequently, they should not be a focus of grading. Determine whether or not points will be deducted for extraneous information in an answer.

5. Obtain independent scoring on at least a sample of responses from multiple readers.

The only way to estimate rater reliability and to assess the relative level of objectivity in scoring is to have more than one rater read some of the papers. In large courses that employ teaching assistants, multiple ratings are fairly easy to obtain. Where assistants are not available, instructors who use essays might assist one another by reading a sample of each other's essay responses so that reliability can be estimated.

#### EXAMPLE OF ANALYTICAL SCORING

Question: What are the principal reasons why research in the social sciences has not progressed as far as that in the biological and physical sciences?

Ideal Response: Since social scientists are themselves part of what they are attempting to study, they cannot achieve the objectivity that is possible in the more precise sciences. Further, the conclusions they reach frequently run counter to deeply held prejudices of people, and hence are unacceptable. Feeling that many of the social interactions of people are not susceptible to scientific study, the public has been less willing to subsidize social research than medicine, for example. Lastly, the scientific study of nature has a much longer history than the scientific study of human behavior. This longer history has provided a much larger body of data and theory from which to move forward.

After writing a response, the instructor extracts the essential parts and assigns quantitative weights to each part. The grading guide for this sample essay item shows a brief statement of each major point and the corresponding number of points:

1.	Scientists part of their subject	+1
2.	Prejudice	+1
3.	Lack of financial support	+1
4.	Short history	+1
5.	Small body of facts and theory	+1
6.	Additional correct information (each)	+1
7.	Incorrect statements (each)	-1

Students should be told in advance when the scoring scheme includes reductions for incorrect statements. Not only is such a warning fair but it is also likely to reduce bluffing by students who are unprepared.

The considerations in weighting the scores from several essay items to form a composite test score have been outlined in EES Technical Bulletin #5, "Assigning Course Grades." Methods of estimating scorer reliability are described in textbooks on educational and psychological measurement.

#### EXAMPLE OF HOLISTIC SCORING

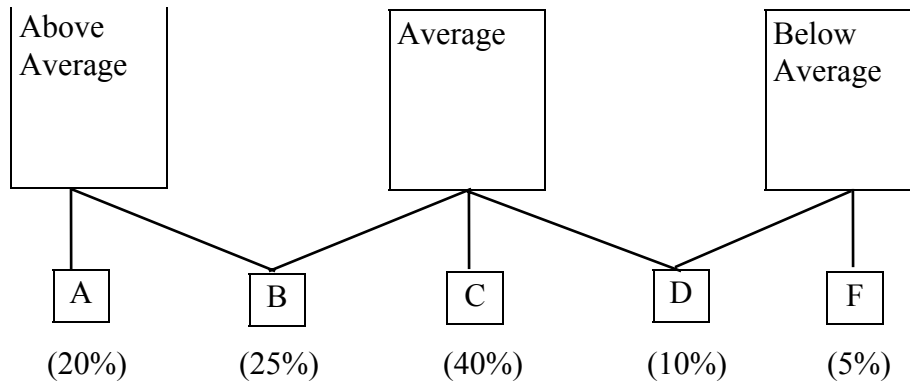
If holistic scoring were applied to the previous example, an instructor might read each response quickly for a general impression and place the papers into three stacks:

Above Average
------------------

Average
---------

Below Average
------------------

After sorting all the responses the papers in the above average stack would be shuffled, re-read, and divided into sub-stacks, depending upon the number of grade categories determined by the instructor. This same procedure is followed for the average and below average categories. The final sort might look something like the following:



(Adapted from Ebel, R.L. and Frisbie, D.A. (1991) p. 195)

The percentages represent the approximate number of papers the instructor expects to have in each grade category.

Holistic scoring can also incorporate a more structured scoring key. For example, students might be asked to list three impacts that the North American Free Trade Agreement has had on border towns in Texas and Mexico and to give an example for each impact. The scoring key might look like this:

- 6 = 3 impacts and 3 appropriate examples
- 5 = 3 impacts and 2 examples
- 4 = 3 impacts and 1 example
- 4 = 2 impacts and 2 examples
- 3 = 2 impacts and 1 example
- 2 = 2 impacts and no examples
- 2 = 1 impact and 1 example
- 1 = 1 impact and no examples
- 0 = no impact or irrelevant response

The difference between this example of holistic scoring and analytical scoring is that the impacts and examples are not predetermined by the grader. In this case every student could conceivably come up with a unique set of responses.

## THE USE OF ITEM ANALYSIS

The purposes of this section are to explain item analysis procedures and the concepts of item difficulty and discrimination, and to suggest ways of using item analysis to evaluate the quality of essay questions. The methods and concepts described are applicable to objective test items also, but the focus in this section is on their use for essay questions.

### ITEM ANALYSIS

Item analysis is a general term that encompasses a variety of methods for summarizing and analyzing the responses of students to test items. Certain patterns of responses can indicate desirable and undesirable features of the item or of the scoring procedure employed. Often these methods suggest why an item has not functioned effectively and how it might be improved. Use of item analysis may also help an instructor improve his or her test writing skill by identifying flaws in items previously written. In addition, a systematic review of student answers may reveal needed changes in instructional emphasis, especially if nearly all students do poorly on a question.

### ITEM DIFFICULTY

This term may be interpreted as "how hard is this item?". That is, how does the performance of a group of examinees compare with the highest possible level of performance? For example, if a maximum of eight points could be earned on an essay question and all examinees received eight points, we would call the item an "easy" one. If all examinees obtained a score of zero on the item, it would be a "difficult" item. Because of differences in the ability level of class members, the complexity of the concept(s) covered by items, the purpose of the test, and other factors, no one difficulty level can be suggested as optimal for all test items.

### A Method of Quantification

We have described difficulty as the extent to which a group of examinees approaches the highest possible level of performance on an item. That is, how high their average item score (points earned on the question) is. The performance level of a group on a question may be expressed as the difference between their average item score and the lowest possible item score (in most cases, zero). Similarly, the highest level of performance can be expressed as the difference between the highest and lowest possible item scores. Our difficulty index (P) will be this ratio:

$$P = \frac{\bar{X} - X_{\min}}{X_{\max} - X_{\min}} \quad (1)$$

where 100 times P is the index of item difficulty in percentage units,  $\bar{X}$  is the average item score for all students,  $X_{\min}$  is the smallest item score possible and  $X_{\max}$  is the highest item score possible. When P is multiplied by 100, it ranges from 0% (for a very difficult item) to 100% (for a very easy one).

A formula for P which gives identical values, but is simpler to use is

$$P = \frac{\sim fX - nX_{\min}}{n(X_{\max} - X_{\min})} \quad (2)$$

where  $\sim fX$  is the total number of points earned by all students on the question, n is the number of students, and  $X_{\max}$  and  $X_{\min}$  have the same meaning as above.

### Calculating P - An Example

Suppose an essay item has been completed by 12 students and that the scoring provided for possible scores of 0, 1, 2, or 3. The following results might have occurred:

Item Score (X)	No. of Students Earning Each Score (f)	Total Points Earned by f Students (fX)
3	3	9
2	2	4
1	2	2
0	5	0
	<u>n = 12</u>	<u><math>\sim fX = 15</math></u>

For these results, all possible item scores were listed in the first column (denoted by X). Then the number of students earning each possible score (denoted by f) was recorded in the second column next to the corresponding item score. The third column (fX) contains the total points earned by the students at each score level (i.e., the item score multiplied by the number of students who earned that score). The third column was then summed to obtain the total number of points earned by all students on that item ( $\sim fX$ ). For this item,  $X_{\max} = 3$  and  $X_{\min} = 0$ , so

$$P = \frac{15 - (12)(0)}{12(3 - 0)} = \frac{15}{36} = 0.42$$

or about 42%. This indicates that the item is of moderate ( $P=50\%$ ) difficulty, as it should be if it is to discriminate properly between levels of achievement.

It is possible to estimate the difficulty index for an item based on only a sample of the students. In computing the item discrimination index in the next section, the item scores of groups of high and low scoring students are used. These same two groups may be used to estimate the difficulty index for the entire class. The assumption is made that the average of the high and low scoring students is a reasonable estimate of total group performance. The difficulty indices for the high and low groups are averaged to obtain an estimate of the item difficulty for the entire class.

### ITEM DISCRIMINATION

Discrimination may be thought of as the degree to which students with varying levels of achievement perform differently on a question. That is, do high achieving students usually produce better responses than lower achieving students? If so, we say the item is functioning properly in differentiating between high- and low-achieving students. We expect students who earn a high score on the test as a whole to do better on each item than those whose test score is lower. (In fact, since a test score is usually just the sum of the item scores, there will be some built-in positive relationship for most items.)

The index of discrimination is the difference in item difficulties between groups of students with high and low test scores. Positive discrimination for an item results when the high scoring group obtains a higher average score on the item than does the low scoring group. A test composed of items with high positive discrimination indices will more likely yield reliable scores (hence reliable grades) than a test whose items have low and negative discrimination indices.

### A Method of Quantification

Since we have described discrimination as the difference in difficulty between high and low scoring students, it seems natural to select two extreme groups, compute index  $P$  for each group and find the difference between  $P$ -values. Under most conditions, the best method of grouping students for this computation is to take the highest and lowest 27% of the examinee group. Many essay examinations, however, are given in small classes (say less than 40 students) for which 27% groups would contain less than 10 students. A practical method of grouping so as to retain enough students in the two groups (to avoid distortion due to a single student's response) would be to use the highest and lowest 10 test scores for classes with 20 to 40 students. For larger classes, groups containing 25% or 27% of the class may be used. The discrimination index is then

$$D = P_U - P_L \quad (3)$$

where  $P_U$  and  $P_L$  are the difficulty indexes described above for the upper (U) and lower (L) groups.

### Calculating D - An Example

Suppose an essay item with possible scores 1 to 4 inclusive had been completed by 30 students. We would select the 10 papers with the highest test scores and the 10 with the lowest scores for our upper (U) and Lower (L) groups, then for each question, tabulate the scores as before:

Item Score (X)	Upper Group		Lower Group	
	Number Earning Score ( $f_U$ )	Total Points $f_U X$	Number Earning Score ( $f_L$ )	Total Points $f_L X$
4	5	20	1	4
3	3	9	1	3
2	1	2	4	8
1	1	1	4	4
	$n_U = 10$	$\sim f_U X = 32$	$n_L = 10$	$\sim f_L X = 19$

For this example  $X_{\max} = 4$  and  $X_{\min} = 1$  so that  $X_{\max} - X_{\min} = 3$ ,

$$P_U = \frac{32 - 10(1)}{10(3 - 1)} = \frac{22}{20} = 0.73$$

and

$$P_L = \frac{19 - 10(1)}{10(3 - 1)} = \frac{9}{20} = 0.30$$

D is then  $P_U - P_L = .43$ . Since this value is positive, it suggests that the question has desirable properties, but how large should D be? A handy (but crude) guide for groups of 30 or fewer is to divide the total number of students taking the test by 100 and use this value as a standard for good discrimination. In our example, there were 30 students in the class so a value of .30 would be used. Our discrimination index was .43, larger than .30, so the question had very acceptable discrimination. For classes larger than 30 students, .30 should be used as a desirable standard. These rules of thumb are intended only for upper and lower groups determined as suggested in this section.

Frequently an instructor will have scores which do not conveniently divide into high and low groups of 10. For example, consider the following 23 scores obtained on a ten-item essay test:

Score	Frequency	Score	Frequency	Score	Frequency
29	1	23	4	17	1
27	1	22	4	15	1
26	2	19	4	13	1
25	2	18	2		

The top ten scores include the students who scored from 23 to 29. However, the bottom group of ten consists of the nine scores from 13 to 19 and one of the four scores of 22. If each group is to have ten papers, one paper must be selected at random from the four papers with scores of 22. Of course, it is not necessary that both groups have an equal number of papers. The important thing is that each group has about 10 papers. In the above example, it probably would be convenient to use only the nine lowest papers in the lower group.

### USING THE INDICES

Use of these indices, at least tabulating the individual question scores as shown above, may help reveal some flaws in the question or in the scoring method. If a negative (or very low) discrimination index occurs for an item, the question may not logically "fit" with the other items contributing to the total test score (it may measure another kind of achievement), or the question may not have indicated clearly enough what kind of responses were desired. Another explanation for a negative D-value is that the better students somehow read into the question something different than was intended (say a more complex, but unintended, interpretation of a key phrase).

D-values near zero may indicate that the question had little relationship to the other questions or that it was a very easy (or very difficult) question so that nearly all students got a large number of points (or very few). In this case, estimating difficulty P will help determine whether the latter is an appropriate explanation.

Difficulty values near 0% or 100% raise some questions about the efficiency of the item. That is, if very few students earn any points on the item (or if nearly all students earn full credit), the item has very little effect on the total test score differences among students.

## SUMMARY OF PROCEDURES FOR ITEM ANALYSIS

1. Obtain the total test score for each student.
2. For classes with less than 40 students, select the 10 highest and 10 lowest total scores to identify "upper" and "lower" groups. For larger classes, select the highest and lowest 25% of the students.
3. For each question in the test:
  - a. compute difficulty indices for the upper and lower groups ( $P_U$  and  $P_L$ ) using formula (2),
  - b. compute discrimination:  $D = P_U - P_L$ ,
  - c. estimate overall difficulty:  $P = 1/2 (P_U + P_L)$ , and
  - d. record each question and the obtained P and D values.
4. Examine these values for each item to evaluate item quality and to look for needed improvements. Refer to students' answers for clues when the D values are negative or very small or when P is close to 0% or 100%.