ED 091 230 SE 017 811

AUTHOR Poole, Richard L.
TITLE Writing and Improving Classroom Tests, A Teacher's Guide. Booklet 1.
INSTITUTION State Univ. of New York, Buffalo. Office of Computer Services.
PUB DATE 73
NOTE 35p.
AVAILABLE FROM Mr. Raymond Volpe, SUNY at Buffalo, Documentation Center, Office of Computer Service Press, 4250 Ridge Lea Road, Buffalo, New York 14226 ($1.00)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS Achievement Tests; Cognitive Tests; *Evaluation; *Item Analysis; Measurement Techniques; Science Education; *Test Construction; Test Interpretation; *Test Reliability; *Test Validity

ABSTRACT

This booklet was designed to provide inservice teachers with a self-contained manual for use in writing and/or improving classroom achievement tests. Following a preface and an introduction, the contents are divided into four sections: Practicality, preface and an introduction, the contents are divided into four sections: Practicality, Interpretability, Validity, and Reliability. The section on reliability contains a discussion of the technique of item analysis. Two appendices (Sample Item File Sheet, Sample Item Analysis Data) are also included. (Author/PEB)

WRITING AND IMPROVING CLASSROOM TESTS

A Teacher's Guide

WRITING AND IMPROVING

CLASS ROOM TESTS

Booklet #1

RICHARD L. POOLE

OHIO STATE UNIVERSITY

State University of New York
at Buffalo
Office of Computer Services Press

1973

SUNY/AB O.C.S. Publication No. D 0001 1

Published by permission of Richard L. Poole.

PREFACE

The purpose of this booklet is to provide the inservice teacher with a self-contained manual so that he may either write better or improve his existing classroom achievement tests.

To this end, the criteria of a good test are enumerated and developed to provide the necessary detail for actual test use and improvement. For example, the section treating reliability contains parts dealing with interpreting item analysis data and the techniques of writing test questions or items.

Therefore, the author would appreciate receiving in person, by telephone, or in writing, any comments which a teacher has regarding the use of this manual. Such comments should be made to Richard L. Poole, Center for Vocational and Technical Education, Ohio State University, 1960 Kenny Road, Columbus, Ohio, 43210 (614) 486-3655.

TABLE OF CONTENTS

Page

I Introduction

It is frequently the case that tests spring into existence because of the immediacy of the situation and similarly are used again and again because of the same exigency. Accordingly in such situations the purpose of the test builder was getting a test ready so he may have something on which to evaluate or grade his students. Hence the focus of the test was not to measure the students achievement, mastery, aptitude, speed or power in the course material, but rather something on which to grade him. Moreover, such things as consistency, practicality and interpretability were not considered. Although few good tests are written in one night, many, no doubt, bad ones are. But what is a good test? What are the criteria of a good test?

In general good tests can be distinguished from bad tests by the extent to which they represent the body of knowledge treated in and perhaps out of the classroom, give consistent results, are easy to administer and score, and are inexpensive to produce, as well as providing norms and if possible alternate forms. That is, a good test is one that is valid, reliable, practical and interpretable.

Note also that these criteria would apply to any type of test, but that in this particular case we are concerned about classroom achievement testing. In addition it is worthwhile to keep in mind that test writing is a continuous event, in that, although it was initially written for grading, had low reliability, little validity, and was impractical and uninterpretable, it does represent the first step in the construction of a good evaluative teaching instrument. To be sure, some of the test questions are bad, but conversely some are also good, and we all have to start improving our tests somewhere and sometime. Hence think optimistically - think "wait till next time."

It is true that the criteria of validity and reliability are the most important for test selection and evaluation, but this does not mean that the criteria of practicality and interpretability are to be ignored. Furthermore, it is not unusual for the test writer or user to devote the lion's share of his efforts to seeking evidence of the former, and thus having exhausted himself disregard the latter. The reasons for this are vague, particularly when one examines what it takes to make a decision regarding these four criteria. Consider, on the one hand that in order to ascertain validity one may need an analysis of the test, a blueprint for the "ideal" test, and perhaps statistical as well as logical data. And for reliability one needs item analysis data or statistical data for each student. On the other hand to ascertain the practicality of a test, one needs to examine the test itself for such things as administration ease, adequacy of directions, ease of scoring, and cost, while to establish interpretability statistical data for the group or individual are needed.

Therefore this manual will treat the criteria of practicality first because it can be determined from a common sense, a priori, viewpoint without recourse to difficult, complex or sometimes unattainable data. Following this the criteria of interpretability, validity, and reliability will be discussed.

I PRACTICALITY

The practical aspects of any test are those concerned with its administration, scoring, and economy.

For the teacher-made type of test, cost is minimal in that it is usually typed and run off on a mimeograph or duplicating machine.

Scoring should be convenient if an answer sheet is used, be it locally constructed or a standardized form designed for use on an electronic scoring machine. If on the other hand the student has to mark his answers on the test booklet itself, then the test directions should be changed to accommodate an answer sheet, or the student should be instructed to put his answers in a designated place; for example, to the left of the question number. The point of this is to increase scoring accuracy and reduce the amount of time needed for scoring.

The remaining facet of practicality is administrability which has to do with the character of the actual test administration. The character of the administration is usually set by the completeness of the test directions, and the resulting student activity immediately prior to the beginning of the test. In terms of the former, the basic concern is to make them as simple and as complete as possible without delaying the start of the test. Furthermore, there is empirical evidence that good directions contribute to reliability.

For the teacher-made type of test, adequate and complete directions should include the following:

1. The number of items or questions composing the test.
2. The number of pages making up the test booklet.
3. The amount of time in the testing period.
4. A statement pertaining to how the test will be scored.
5. A statement indicating how the student is to indicate his answer.
6. A statement indicating the arrangement of the questions on the the answer sheet.
7. A statement telling the student to inspect the booklet before beginning the test.

The reason for this is that the responsibility of the candidate or student is to take the test, while the responsibility of the teacher includes providing him with a complete test. In those situations where the teacher is not directly concerned with the production of the test, these directions are of particular importance. Why?

Consider the consequences of production or collation errors on the test for both the student and the teacher. For example, suppose the last several pages of the test are omitted from one test booklet. How is the student to know that there were 75 questions on the test, and not 50 as he had in his booklet? How is the teacher to score and interpret such a test?

In addition, it is recommended that the directions be put on a separate page which can then be used as a title or cover page for the test booklet. The advantage of this is that if the examination is changed, then the original test booklet may be used in full or partially, and the new items just added on.

Sample 1, contains an example of a test cover or title page.

Moreover every test booklet should be numbered so that none may be "lost". And the test writer may wish to use some code to identify when the test was first written. Inspection of Sample 1, has a place for booklet number, and under that a space for coded information (XXXXXX).

Name: ______________________________ Test Booklet No.____
XXXXXX

F-502 Mid Term
Directions

Sample 1 - Test Directions

In the questions below, choose the best answer. Indicate your choice by filling in the space under the appropriate number or letter on the answer sheet. Be sure that the number of the question you are answering is the same as the question number on the answer sheet where you are indicating your choice. Answer every question, even if you are not completely certain that the answer you are giving is the correct one. Each correct answer is worth one point. There are ____(86) questions on this test, and the test booklet contains ____ (13) pages. Examine your booklet to see that it has the required number of questions and pages.

Write your name on the line provided on the test booklet, and record your booklet number on the answer sheet in the space labelled GRADE. You have 2 hours to complete the test. Are there any questions on what you are to do? Turn the page and begin.

Sample 1 - Test Cover or Title Page

II INTERPRETABILITY

As the name implies this is concerned with the meaning of the numbers or scores. The raw score, that is the quantitative end product of counting the number of questions answered correctly is of very little value. Consider a score of 20 or 60. What does it mean? Obviously without additional information very little can be said about the score itself. Other information might include such things as the total number of questions on the test, or the number or percentage of students getting this or a lower score, or it might be the test mean and standard deviation. Accordingly, with these different bits of information different interpretations become possible. If for example we knew that the total number of questions asked were 20 or 60, then we know that our student received a perfect paper. But suppose that the total number of questions was 80. In this situation we might say that the student with the score of 20 was not very well off, or that he did poorly, whereas, the student with the score of 60 did somewhat better. However to do so might be an over-generalization, for in the first case it might be that this was the highest score in the class, while in the second, a score of 60 might be the lowest score in the class.

Clearly then in order to interpret test scores, additional information is needed. Traditionally test score interpretation usually starts with referencing the group or class of students. That is, the test results for the class are ranked and tallied from low to high to form a frequency distribution. Based on this frequency distribution it is then possible to ask two types of questions:

(1) What percent of the students get a score (some score) or less?

(2) Seventy-five (or some) percent of the students get what score or less?

These two types of questions point out the essential difference and similarity of a percentile rank and a percentile. Both of them are derived or converted raw scores resulting from focusing upon the two aspects of the frequency distribution. Notice in the first question we were given a score, and are looking for a percent, hence we are seeking a percentile rank. In the second question we are given the percent and are looking for a score, hence we are seeking a percentile. That is, a percentile is a test score below which a certain percent of the students fall. A percentile rank is a percent representing the relative portion of students getting at least a given score.

The basic factor in the determination of percentiles and percentile ranks is just the number of students to a point on a distribution of test scores. Therefore, because they are determined from an ordering process they should not be added, subtracted, or subjected to any arithmetic manipulation as is usually needed for the determination of a student's final grade. This being the case, they then have limited classroom use.

STANDARD SCORES

As indicated previously another type of derived score is based on the use of the test mean and the standard deviation There are several forms of standard scores, but the most basic one, "z" is defined as the ratio of the difference between the raw score from test mean to the standard deviation of the test. Symbolically "z" is defined as:

$$z = \frac{X - \bar{X}}{S}$$

Where X represents a test score

$\bar{X}$ represents the test mean

S represents the standard deviation of the test.

It can be seen that z scores do focus upon the actual test scores, and not how the students distribute themselves. Furthermore, they may be subjected to arithmetic manipulation as is needed in the determination of a students final grade.

Some limitations in the use of standard scores are that they may be initially troublesome to work with in that they yield both positive and negative values. Another limitation deals with equalizing the contribution of any one test in the determination of a final grade. Frequently, teachers inform students (wrongly) that their final grade will be determined from a equal weighting of two or more tests. However, what these teachers should say is that they are going to average the students test scores. These two statements are not equivalent, for the weight of a test psychometrically is determined by it's standard deviation. And what this means is that if a teacher subscribes to the first statement standard scores should be determined and used in the grading process. If he subscribes to the second statement then he averages the raw scores and then uses them in the grading process.

Consider the following situations using two students and two tests. To use more of either would be to complicate the matter, and this example should point out the distinction that we want to make.

Situation #1	Test		Student & Raw Score	
Test #	Mean	Standard deviation	John	Jim
1	50	10	60	40
2	60	5	55	65
Total point count			115	105
Average point count			57.5	52.5

Grading Strategy:

Because John's point count or average point count exceeds Jim's John gets a higher grade than Jim.

ISSUE:

Test 1 counts twice as much as test 2, since the standard deviation is twice as large as test 2's. Notice that if the scores of test two are added in twice, the total sum of the three scores will be equal - (60+55+55 = 170; 40+65+65 = 170).

Situation #2 Test #	Test Mean	Standard deviation	Student & Raw & Standard Scores John	 Jim
1	50	10	60 + 1z	40 - 1z
2	60	5	55 - 1z	65 + 1z
Total z			0z	0z
Average z				

Grading Strategy:

Because both John & Jim have identical z's, then they both get the same grade.

ISSUE:

- Standard scores reduce all test standard deviations to the same number, therefore both tests are equally weighted. However, if there is evidence that the tests are not of equal value (validity, reliability, etc.) then it may be illogical to weigh them equally.

STATISTICAL DATA:

Fundamentally there are two types of statistical indices which are helpful in interpreting the test. One of these indices is concerned with the topic of central tendency or location, and the other with score scatter, dispersion, or variability.

Arithmetically, the mean, is the quotient resulting from dividing the sum of the scores by their number. This is equivalent to summing across for the entire group or class the number of test questions answered correctly and dividing by the number of students in the class. In effect, then, what the mean does is to distribute the number of points or the number of questions answered correctly by the group, equally among the group. Viewed in this fashion, the mean is the "democratic" score, or the base score, or if you will - the anchor in reality. That is, it is the representative score for the group.

For classroom achievement tests, it is desirable to have the mean be approximately equal to between 40% to 60% of the number of questions asked. Bear in mind that we are trying to determine the level of achievement for each student, and with a mean that is about in the middle of the score distribution, each student will be able to more or less clearly identify himself. Similarly if the mean is either too high or too low, we in effect reduce the possible score variability by putting a ceiling or floor on it. Hence the student can not clearly indicate his knowledge status. Generally, the difference between the lowest and highest score should be equal to 4 to 5 standard deviations.

III VALIDITY

The most important criteria for test development and selection is validity. But validity is a generic term which includes all the different types and their associated purposes. Therefore a common synonym for validity is "purposeful".

The most basic form of validity is content validity, and its purpose is to represent the material on which the test is based. It is obtained by sampling from "the material", and determined by the adequacy of the sampling. The second type of validity is called criterion related validity and its purpose is to determine the degree of relationship existing between test performance and other kinds of student performance either now or in the future. It is obtained by correlating the two measures. The third and last type of validity is called construct validity which has as its purpose the determination of what the test is measuring. This type of validity is determined by gathering logical as well as empirical data.

In terms of achievement test construction the only type of validity of concern to us is content validity. As indicated above the establishment of content validity is determined by the adequacy of the sampling of the course material. However because it usually takes several repetitions to get the final test, we must go back and forth between what was desired and what was acquired.

We may begin therefore with the theoretical ideal approach, or an actual pragmatic one, and because most of us already have most of our tests built, it seems more realistic to start there.

To determine the content validity of the existing test, it has to be analyzed item by item in terms of content and item or question type. That is, is the item factually or non-factually oriented, and also what is its source or what part of the course does it focus upon.

A convenient form to use when analyzing a test is to use a two-way table with the subject matter content along one side and item type along the other. After each item has been classified it is usually possible to collapse specific item content into a smaller number of general areas. An example of such a test analysis table (TAT) for a hypothetical elementary arithmatic test is found in Table 1.

After a test analysis has been done, it is possible on one hand to compare the subject matter with the course outline, and the item types with your feelings and recollections of the skills taught for. On the other hand, it would be better if you could compare your test analysis with an independent outside source which would provide you with a blueprint of the ideal test. Indeed, it would be beautiful if you were the outside authority who produced the blueprint for the ideal test. To produce such a blueprint it is necessary to build a two-way table which treats or outlines the subject matter along one dimension and various behaviors or skills, which come from the educational objectives, along the other. This time however, the classification of the test items can be more detailed than factual or non-factual, in fact, hopefully the items measure what is called for by the educational objectives. In addition, the teacher in a subjective or personal fashion assigns a weighting or emphasis, usually expressed as a percent, but sometimes as a count, to each topic and behavior. This weighting is an indication of the relative worth of each entry in the table. Since very few teachers make tests containing 100 or more items, it is more convenient to express this weighting as a count.

A two way table which contains not only the subject matter and cognitive behaviors or skills taught, but also the partitioning of the number of test items to each dual entry of content and behavior is called a table of specifications (TOS). An example of a table of specification is found in Table 2.

Now that both the TAT and the TOS are produced we can compare what was desired as indicated in the TOS with what was actually acquired. Moreover, the TAT was constructed using every item in the test, not necessarily every statistically or pedagogically acceptable item. That is, some of the items in this table might be poor, and hence should be discarded or set aside for the moment. In any event the discrepancy between the real test and the ideal test is evident.

In this example the ideal test has partitioned the items equally, whereas the real test has the items distributed unequally with the topics of addition and subtraction receiving three times as many items as those of multiplication and division. In terms of cognitive behaviors or educational objectives, the ideal test calls for the emphasis to be placed on the higher forms of thinking behavior, while the test itself places the majority of items in the factual or perhaps lowest level of cognitive behavior.

If one also considers item quality, then in all probability the disproportionality may become even more severe. Therefore before constructing the TAT with all the test items, it is desirable to perform an item analysis and then build the TAT with only those items which "pass" it.

Without getting quantitative, for content validity is more qualitative than quantitative, it might be said that content validity is determined by the coalescence of these two tables.

It is to be expected that because of item attrition one or more categories in the TOS will not be represented. But, taking the item analysis data into consideration, hopefully one can salvage the unacceptable items by rewriting them so that next time these areas will be covered. Also it should be noted that it is usually more economical to endeavor to rewrite a discarded or unacceptable item than it is to write a completely new item.

TABLE 1 - Test Analysis

Table for a Hypothetical Elementary Arithmetic Test

Item Type Content	Factual	Non-Factual Thinking	Total
Addition	~~1111~~ ~~1111~~ ~~1111~~		15
Subtraction	~~1111~~ ~~1111~~ 1	1111	15
Multiplication	1111	1	5
Division	111	11	5
Total	33	7	40

TABLE 2:

A Sample Table of Specification for an Elementary Arithmetic Test

A Hypothetical Arithmetic Test

Behavior (objectives & Items)

Behavior Content	Defining	Knowledge	Comprehension	Problem Formulation	Problem Solving	Total
Addition	1	2	2	3	2	10
Subtraction	1	2	2	2	3	10
Multiplication		1	3	3	3	10
Division		1	3	3	3	10
Total	2	6	10	11	11	40

IV RELIABILITY

Reliability like validity is a gerᵔᵔic term and therefore includes within it all of the different types. But regardless of the type they all focus on the consistency of the test performance. Hence, a synonym for reliability is "consistency." However, as there are different instances of consistency of test performance, correspondingly there are the different types of reliability. There is consistency of performance on the test as a whole, consistency of performance on different tests, and also consistency of performance on a test over time, which respectively account for internal consistency, equivalence, and stability reliability.

It follows that each type of reliability estimate requires a different set of procedures for its determination, and these in turn no doubt contribute to its use or disuse. Consider that in order to establish stability reliability it is necessary to administer the same test twice to the same group of students. In terms of testing, this is somewhat inconvenient, but the educational aspects present more of a problem. Consider the ramifications in terms of marking or grading, student, teacher and possibly administrator behavior. Practical matters such as these as well as certain technical matters preclude the use of this type of reliability for the classroom tests. While to establish equivalence reliability, the task for the teacher is more onerous, in that now two very similar but not identical tests are needed. Indeed most teachers have enough trouble writing and producing one test, so that requiring two tests would be "out of the question." Hence, again another type of reliability is precluded from classroom use.

Therefore since both the stability and equivalence types of reliability are precluded from classroom use because of practical problems, then if we are to have reliability at all, it will have to be of the internal consistency type. Basically, there are two approaches to this type of reliability - the split half or Spearman-Brown, and the Kuder-Richardson. With the Spearman-Brown approach it is necessary to divide the test into 2 equal halves - usually the odd number versus the even number items - and score each half separately. These two scores are then correlated. This correlation coefficient is then substituted into the Spearman-Brown "prophecy formula" to yield the reliability coefficient for the entire test.

To eliminate the necessity of dividing a test in half and scoring each part separately, Kuder & Richardson devised a method to estimate reliability from item analysis data. In this approach, the reliability coefficient is determined from the item difficulties, the test mean and standard deviation, and the total number of items making up the test.

Interpretation of a reliability coefficient

Although there are several ways of interpreting a reliability coefficient, the three that seem most important for our purposes are those of comparability, ranking and expected chance variability.

The comparative approach, of course, focuses on the reliability results that are obtained with standardized tests. Using this criterion we usually find coefficients ranging from .80 to .93 or higher. However, for a teacher made type of test a reasonable estimate of reliability is at least .70. This value is to be viewed as an arbitrary one, and therefore not absolute. Depending on certain conditions, or the situation and the terrain, this value may be lower. But in general, the reliability coefficient of a teacher made test should be of the order of .70 to be acceptable.

Regardless of the type of reliability under consideration all of them indicate the consistency with which a test ranks the students. Hence our second method of interpreting a reliability coefficient will focus upon this.

To the extent that the reliability coefficient deviates from a maximum value of 1.00 and approaches a minimum value 0.00, there are changes in the relative positions or ranks of the students tested.

For education and the students tested the analog of these ranks are grades or marks which are isomorphic to them. Hence, if there is inconsistency in the ranking then there will be inconsistency in the grading or marking. That is, it's possible for a student to receive a mark or grade that is too high or low for him. It has been shown that with a test that has a reliability coefficient of .90, and with a 5 level grading distribution of A,B,C,D, & F, divided hierarchically as 5,25,40,25,5 percent, that about 23 percent of the class were mis-marked. In other words, with an unreliable test, the students receive grades which may be more a function of chance than achievement.

These two approaches do not give us any indication of the amount of variation that can be expected for everyone's score. That is, theoretically if we were to retest this student, by how many points would his score change? In general we can answer this question by knowing the standard error of measurement. For example if a student's score is 46 and the standard error of measurement is 2, then we can expect his score to vary between 44 and 48 about 68% of the time. That is, with repeated testing, and without any additional learning by the student, his score can vary by this much as a function of chance errors.

By way of a summary, what these interpretations mean for the classroom teacher, is that a student's score can vary somewhat as a function of chance errors, and also with an unreliable test his grade may be a function of chance.

Improving Test Reliability

Test reliability can be improved in general by:

1. Increasing its length by the addition of more test questions of about the same quality.
2. Replacing the items which are either too hard or too easy; that is replacing those items which have either high or low item difficulty.
3. Replacing the items which have low or negative item discrimination.
4. Increasing the number of alternatives or options for each test item.
5. Writing adequate and clear test directions.

For the most part these rules apply when we are considering giving the test for a second time. But what can a classroom teacher do for his test, which now has a poor or low reliability. If the test is long, say about 75 questions, then it is possible to eliminate the questions which have negative and low discrimination, and then rescore the papers. Sometime ago we did this for a 75 item test which had a reliability of about .47, and after eliminating 23 items, the same test had a new reliability of .67.

ITEM ANALYSIS

Item analysis is a statistical process designed to yield several indices which assist us in determining the statistical properties of the individual items. It is one of the two processes which should be used in evaluating the test items. The other process doesn't have a universal label, but for lack of anything else,let's call it "educational soundness or appeal." Far too many test makers concern themselves only with the item analysis data and not with the items' educational appeal. That is, for them the final filter is item analysis, and if the item has "good" statistics it is permitted to go into the final form of the test. However, we would like to see classroom teachers use both criteria for item selection. What this means is that "good or acceptable" items are those that the teachers feel are appropriate for their classes, and also those whose statistics are not too bad. The item statistics to which I am referring are those of item difficulty, item discrimination, and option distracting power. It will be recalled that an option is one of the possible choices that an individual has in responding to a test question.

ITEM DIFFICULTY:

By definition, item difficulty is the percentage of students getting an item correct; hence, it has values from zero to 100. Accordingly, items which have low difficulties, that is, very few students get it correct are called hard. But items which have high difficulties, that is a large proportion getting it correct, are called easy.

In order to maximize reliability one would write or include in his test only items which had 50% difficulty, but because item difficulty is empirically determined this does not happen that frequently. Nevertheless, the condition does exist that with an item bank, one could generate such a test.

For classroom tests however, its not unusual for the item difficulties to range between 25% and 75%. In addition items which have difficulties beyond these limits may be used by individual teachers because they value the item educationally. The issue of course is that if there are a great many of these kinds of items, reliability will suffer. Hence, if one is desirous of building an achievement test then one can entertain the criterion of educational soundness until it begins to seriously weaken the statistical properties of the entire test. In other words, it may be good educational policy to include a few very easy items at the beginning of the test to give the student a good start, but if the test is to be composed of all easy items, it would probably be best to view it as a mastery and not achievement test.

ITEM DISCRIMINATION:

Item discrimination refer to the ability of an item to separate the group into high and low achievers. It may be expressed by any one of several indices, which means that the one finally selected is done probably because of convenience. Regardless of how it was calculated it may range from - 1.00 to + 1.00. In general the more positive the discrimination the better.

Those items which have a discrimination of .30 or higher are said to be good and acceptable. Those between .20 and .29 are called marginal, and those of .19 to zero or negative, are poor or terrible.

It follows that item discrimination is related to item difficulty, in the sense that if the item is very easy or hard the class can not be "equally" separated into high and low achievers. Hence with very hard or easy items, the labels might be shifted to the lower ranges of values. That is, for an item with a difficulty of 20%, a discrimination value of say .25 could be considered as acceptable.

OPTION DISTRUCTION POWER

It will be recalled that item discrimination deals with the ability of a test question to separate the examinees into high and low achievers. Accordingly, it is determined by dividing the class into two groups, sometimes called the high and the low achievers. These two groups also provide us with information on how the class responded to the item as a whole, and also to each possible answer.

Option Distracting Power is by definition the difference between the number in the high group and low group who chose each option. That is, for a four choice multiple choice question, each possible answer or option has a distracting power. The keyed response should be the only option with a positive distracting power. All the other options should have a negative distracting power. The logic of this is that more students in the high achievers should get the item correct than in the low achievers. Conversely, fewer students in the high group should choose any of the wrong answers than those in the low achievement group. Based on these three indices it is possible to evaluate an item statistically, and also if need be to help with its rewriting by pointing out options which are not functioning. It is also possible to eliminate some poor questions by just rereading them to see if they conform to the characteristics of good item writing procedures. Some characteristics of good item writing will be listed below, but before listing these rules, let it be said by way of definition that the stem of a test question is that sentence or phrase which is supposed to set the stage for selecting an answer.

SOME RULES FOR WRITING BETTER TEST QUESTIONS:

1. Does the stem clearly indicate the problem, or establish a basis for an answer?
2. Is the stem short and simple and does it contain words necessary only for the communication of its intent?
3. Does the stem provide any clues to the correct answer?
4. Are key words in the stem underlined or printed in capital letters to make them conspicuous?
5. Are all the options plausible?
6. Is the keyed response unique in any way?
7. Are words repeated in each option which may otherwise be put in the stem?

Appendix A: Item File Sheet

Test Title: ______________________ Class: ____________

Item # ________ Content: ______________________

Cognitive Level:

Reference:

Date	OPTION										OMITS	TOTAL N	DIF	DIS	PT BIS
	1		2		3		4		5						
	H	L	H	L	H	L	H	L	H	L					

Comments:

Appendix B: ITEM ANALYSIS DATA

ITEM...........	79	KEYED RESP...........	4
DIFFICULTY.....	72	DISCR.................	.34
PT. BISRL......	.45	T....................	3.96

MEAN SCORE	RIGHTS..............................	57.5
	WRONGS..............................	50.6

RESPONSE PATTERN

	1	2	3	4	5	OMIT
	1	2	0	29	0	0
Upper Freg PCT	1.5	3.1	0.0	44.6	0.0	0.0
Lower Freg		10	3	18	0	0
PCT	3.1	15.4	4.6	27.7	0.0	0.0

<u>Item</u>.......... 79 - Particular item or question under consideration.
<u>Keyed Resp</u> 4 - Particular option or possible answer which is the answer to this item.
<u>Difficulty</u> 72 - Percent of the students getting the item correct.
<u>Discr</u> .34 - A measure of discrimination.
<u>Pt. Bisrl</u> .45 - A more sophisticated measure of discrimination.

It is recommended that most teachers use the DISCR index for their classroom tests; however for publication and perhaps research purposes the more satistically rigorous PT. BISRL index should be used.

<u>T</u> ...... 3.96 - Is the Student's t Test statistic with N-2 degrees of freedom. It is a statistic to determine if the point biseral correlational coefficient differs significantly from zero. Also it is a test of the difference between the mean score rights and mean score wrongs. (Most teachers should ignore this).

<u>MEAN SCORE RIGHTS</u> 57.5 - This is the average or mean score for the 47 students or 72% of the class that answered the item correctly.

<u>MEAN SCORE WRONGS</u> 50.6 - This is the average or mean score for the 18 students or 28% of the class that answered the item incorrectly.

RESPONSE PATTERN As its name implies this depicts both the number of students and also the percent of students that select each option. For example 29 students or 44.6% of the students in the upper half of the class and 18 students or 27.7% of the students in the lower half of the class selected the keyed (correct) answer.

The more discerning reader may note that the upper half of the class contains 32 students, while the lower half contains 33 students. If the class contains an even number of students then each group will be equal, but if it contains an odd number of students the odd one is placed in the lower group to yield a conservative estimate.