Chapter 8

Judging the Quality of Test Items: Item Analysis

INTRODUCTION

An item merely written is not complete. Once an idea for a test item has been conceived and articulated according to the rules for writing good items, the important task of determining its quality remains. If it is flawed, it must be improved or discarded. To accomplish such analysis for items, the writer must employ a precise methodology to systematically uncover information about an item, and then make judgments about it based on that information. In educational and psychological assessment, such judgments revolve around detecting and reducing errors in measurement, which can be in the form of either systematic bias or random error. This chapter explores the concept of measurement error first by explaining it and then by discussing strategies for determining the degree to which it may exist and how the sources for error can be reduced. By alleviating the causes of measurement error, the quality of particular test items correspondingly increases.

Two basic approaches can be used to unearth errors in measurement: through judgments by knowledgeable people using established criteria and through appropriate statistics. A variety of procedures exists for both, and both can address systematic bias as well as random error. This chapter describes the most widely practiced forms of these approaches for detecting both kinds of measurement error.

Throughout this chapter, the reader should keep in mind that many of the techniques of item analysis are grounded in either classical or modern test theory. In most instances, describing the theory in detail is beyond the scope of this book; therefore, this chapter will frequently refer the reader to other sources devoted to measurement theory for more thorough explanations.

Additionally, throughout this chapter various statistics will be cited as useful for analyzing particular data. The formulas for these statistics are not described in this book, because in most test-development contexts, the calculations required by the mathematical algorithms are accomplished with the aid of a computer. Many statistical programs are available for both large and small computers that will compute these statistics. While computers are not required for figuring most of the statistics presented in this book—in fact, many can easily be done either by hand or with the aid of only a pocket calculator—they are commonly used and do make the task more convenient. Instead, this chapter focuses on understanding the aims and reasons for a particular analytical look at items, as well as learning the procedures needed to accomplish it.

The following topics are covered in this chapter:

- measurement error
- item analysis
- validating the content of items
- judgmental approaches to item analysis
- item statistics
- item parameters
- item bias

MEASUREMENT ERROR

The reader has already been introduced to the concept of measurement error in Chapter 2, where it was discussed in relation to describing the purpose for test items. It was explained there that error is inherent in measuring educational achievements and psychological constructs. Here, the description of measurement error focuses upon techniques for detecting its presence and reducing the sources from whence it arose. These unwanted sources could be wording in items that is confusing to examin-

ees, information in items that is not factually accurate, lack of congruence between an item and the objective it is intended to assess, and more. While measurement error cannot be completely eliminated in psychological assessment as it is currently practiced, keeping the sources of error to a minimum will contribute to making items better.

Description of Measurement Error

Measurement error can be simply described as the amount of deviation an examinee's score on a set of test items would exhibit if the test was administered to that examinee an infinite number of times, under identical conditions. The more those scores disperse, the greater the error of measurement. Of course, in real life, no examinee is given a set of test items an infinite number of times, so the measurement error must be estimated from a single administration. But, it can be estimated with precision. The precision with which a score is estimated is expressed in the term *standard error of measurement.* Theoreticians conceive of this relationship with the following equation:

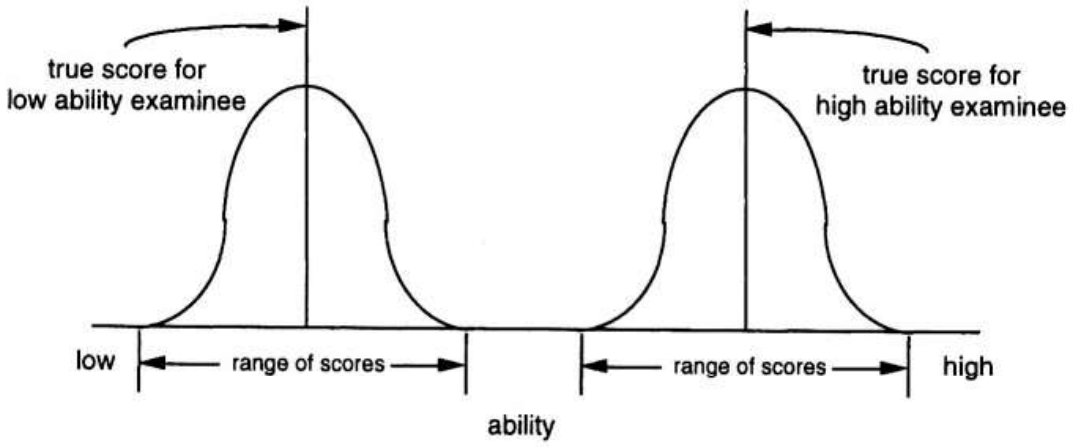
True score = Observed score – Measurement error

where: *True score* is the score an examinee would obtain if no error was present, and

Observed score is the score an examinee actually received during a real-life test administration.

The true score is conceptualized as the mean score the examinee would have received by averaging his or her score from the infinite number of test administrations theoretically done. The standard error of measurement may be graphically represented as the distribution of scores around the true score for an individual. Figure 8.1 displays this graphical representation for two examinees, one of low ability and one of high ability.

Figure 8.1 Display of standard error of measurement for different abilities.



These highlighted points about measurement error are important to understand if one wishes to use properly the techniques of item analysis. The reader should be aware, however, that this description of measurement error scarcely touches the surface of the theory behind this topic. Researchers have written extensively about the theory of measurement error, from Thorndike's 1904 classic *An Introduction to the Theory of Mental and Social Measurements*, to the excellent 1986 text by Crocker and Algina, *Introduction to Classical and Modern Test Theory.* In between, dozens of fine books were written that describe in detail test theory and errors of measurement (e.g., Cronbach, 1984; Ebel & Frisbie, 1986; Gulliksen, 1950; Lindquist, 1936; Lord, 1952; Lord & Novick, 1968; Nunnally, 1978) to which the interested reader is referred.

Keeping Measurement Error in Perspective

One should realize, however, that the presence of error in psychological assessments is not so serious a problem as it first appears. There are a number of reasons why this is so. First, it is easy to overstate the importance of error in mental measurements because seemingly little is understood about psychological processes and how they may be assessed. This is in itself not too great a problem when one considers that our understanding of many aspects of the physical realm is similarly limited. In fact, psychological measurements are probably no more plagued by error than are measurements in other areas of science. Furthermore, some small amount

of measurement error probably does not harm most scientific investigations, whether in the physical realm or the psychological domain.

Another reason why the importance of error in measurement is often exaggerated is that the terminology and methods used for describing it can have an imposing appearance to persons unfamiliar with the statistics involved. Since many aspects of measurement error can be conveniently expressed in mathematical terms—and later in this chapter we will explore some of the techniques for doing so—it is difficult for untutored persons to regard such numbers with perspective. Mathematical models for describing measurement error, with their specialized nomenclature and elaborate, Greek-lettered symbols, are seemingly impressive. This makes it easy to overstate their importance.

While it is worthwhile to keep the issue of measurement error in perspective, the skilled item developer identifies and reduces sources of error throughout the item-construction process. By identifying and reducing the sources for measurement error, the quality of the items will be correspondingly improved. Therefore, it is important to learn about error in this context and discover strategies that will help to reduce it, while simultaneously keeping perspective on the concepts of measurement error.

UNDERSTANDING ITEM ANALYSIS

Item analysis is the process by which test items are examined critically. Its purpose is to identify and reduce the sources of error in measurement. Writers routinely perform item analysis so that they may gauge the quality of items and discard those which are unacceptable, repair those which can be improved, and retain those which meet criteria of merit.

Item analysis is accomplished in either of two ways: through numerical analysis on by judgmental approaches. In numerical item analysis, the statistical properties of particular test items are examined in relation to a response distribution. This definition requires that someone has already prepared a numeric description of test items *after* they have been administered to a group of examinees, as is done in a field trial for examination development. The primary purpose for field trials of items is, of course, to collect appropriate data for reviewing them.

While the field trial of items is necessary to gather the data for analysis of individual items, this text does not discuss the process of conducting a

field trial. A field trial is itself a procedure involving myriad considerations, such as how items should be ordered on a test form or distributed among various test forms, how to determine an appropriate sample, and necessary considerations of the size of that sample. Although these issues are important to constructing a good test instrument, they are more properly concerns for test developers rather than item writers, and therefore they are not addressed here. The reader can easily identify sources in which these issues are fully addressed, such as Allen and Yen (1979), Crocker and Algina (1986), Nunnally (1987), Thorndike (1982), and many others.

Judgmental approaches to analyzing items involve asking people to comment on particular items according to some criteria. The persons asked for comment might be content-area experts, editorial specialists, or even examinees. Judgmental reviews have two guiding principles: each reviewer must be qualified for the task, and the task itself must be a systematic process. In this context, *a systematic process* means that a methodology is defined and that criteria for the review are available. Both numerical analysis and judgmental review are important ways for writers to learn about the items they have written. Each of these avenues to analyzing items will be explored in appropriate sections of this chapter.

VALIDATING THE CONTENT OF ITEMS

It was explained in Chapter 3 that evidence for valid test-score interpretations is not inherent in the item-construction process but must be gathered through a systematic validation study. Such a study will typically be an examination of content-related evidence for validity. The procedures used for gathering content-related evidence for validity can also be of enormous help in determining the quality of test items. The item writer can use the information uncovered through this systematic study to examine and improve items.

A content-validation study usually seeks to establish a consensus of informed opinions about the degree of congruence between particular test items and specific descriptions of the content domain that is intended to be assessed by those items. This typically requires convening a panel of expert judges who rate the item-to-content congruence according to some established criteria.

Two principal methods are used to gather the opinions of experts about

the congruence between a content domain and specific test items. In the first method, judges are given the objective descriptions and test items that have already been matched by the intention of the item writer. In other words, the item writer, who will have constructed an item intentionally to reflect a particular domain or objective, will tell the judges which item is supposed to be matched with which description. The role of the judges, then, is to confirm or reject the opinion of the item writer.

Of course, an appropriate form for recording the opinions of the judges will be needed. The form for recording judges' ratings should allow for more than just a yes-versus-no matching by permitting judges to record degrees of congruence—typically, one designation recording a strong match, a second designation noting a moderate match or uncertainty, and a third designation denoting a poor match or no match at all. These three categories are all that is necessary; finer discriminations are of little practical value and could needlessly complicate the consensus-building process. Also, a space on the form should be provided where judges can comment on the rationale for a given opinion. An example of such a form is given in Table 8.1.

Another, and stronger, approach to gaining consensus among expert judges of item-objective congruence is to refrain from informing the judges beforehand of the item writer's intended match. In this approach, the judges are simply given the items and the objectives without any indication of which item is meant to be matched with which objective. Each judge will indicate his or her perceived match on a rating sheet, and a project coordinator will tally the responses. A consensus of judges' opinions as to a particular item-objective match is considered content-related evidence for validity. Table 8.2 offers a sample of a rating form that can be used for this data-gathering technique. Variations may be made in the forms displayed in Tables 8.1 and 8.2 to suit particular circumstances.

As a further means of enhancing the rigor of this judging process, the judges may be assigned to a "blind" panel, that is, they do not meet in face-to-face sessions, nor do they know the identity of the other panelists. Opinions of the panelists are gathered by a project coordinator through telephone or mail contacts. Because the panelists do not meet, the consensus would presumably be uncontaminated by specious persuasion, or by the effects of personal prestige, rank, or charisma.

In most instances four or five judges rating each test item will suffice; however, if a large number of items are to be rated, the items may be split

into two or more groups, and four or five judges for each group of items will be needed. For tests that have a cut-off score and significant consequences for examinees, then more judges—possibly as many as ten or even fifteen—are recommended. Under no circumstance should one person be the sole judge in a content-validation study, especially the person who wrote the test item. Unrecognized prejudices, chauvinistic perspectives, or other biases can too easily go unnoticed when one judges one's own work.

The judges should be both expert in the subject area they are assessing and trained for the matching task. A judge's subject-area competence is requisite to his or her selection because any subject has many details that could go unnoticed by a novice. Training the judges for the matching task is equally important, for the task requires more skill than may be imagined at first glance. Chapter 4 included a thorough discussion of strategies for arriving at an item-objective match and could be used as a training guide.

Quantifying Results of Judge Opinions

Once the relevant data from the judges' evaluation of items has been gathered, the information must be analyzed and interpreted. This means determining whether a consensus of opinions has been achieved. The item writer, seeking to get feedback about the quality of items, must decide how many judgments matching a particular test item to a specific objective must coincide in order to declare that a consensus of opinion has been achieved. There is no precisely established number, but the consensus should be quite evident; e.g., if there are five judges, four must agree, or if there are ten judges, eight must agree.

Although simply tabulating the number of opinions that agree is undoubtably the most popular method, other procedures are sometimes preferable. Some of these methods require quantitative approaches and may be more precise than tallies. For example, if there is the presumption (often made) that for every test item there should be one, and only one, clear match to a skill or objective, an index of the item-objective congruence may be derived (Rovinelli and Hambleton, 1977; Hambleton, 1980). For this procedure, judges would be instructed to assign a +1 if there is a strong match between an item and an objective, a 0 if a judge is uncertain whether congruence exists, and a -1 if the item does not match the objective. The rating form displayed in Table 8.1 is an example of a form

Table 8.1 Item-Objective Congruence Rating Form

Name _____

Instructions: Read Objective #1 below. Next, read the first item in the test booklet. Consider carefully the degree to which the item is congruent with the skill. Rate the congruence according to this scheme:

H = high degree of congruence
M = medium degree of congruence
L = low degree of congruence or uncertainty

If you have comments about the congruence of this item, record them in the space provided. After you have finished with this item, proceed to the second item, and thereafter to all subsequent items, rating each in the same manner.

Objective #1
Use mathematical techniques to solve real-life problems.

	Rating	**Comment**
Item #1	_____	_____
Item #2	_____	_____
Item #3	_____	_____
Item #4	_____	_____
Item #6	_____	_____
Item #9	_____	_____

Objective #2
Use the properties of two- and three-dimensional figures to perform geometrical calculations.

	Rating	**Comment**
Item #3	_____	_____
Item #5	_____	_____
Item #8	_____	_____
Item #10	_____	_____
Item #11	_____	_____

Table 8.2 Judge's Summary Sheet for the Items/Objectives Matching Task (Sample)

Items/Objectives Matching Task

Reviewer: _____ Date: _____

Content Area: _____

First, read carefully through the lists of domain specifications and test items. Your task is to indicate whether or not you feel each test item is a measure of *one* of the domain specifications. It is, if you feel examinee performance on the test item would provide an indication of an examinee's level of performance in a pool of test items measuring the domain specification. Beside each objective, write in the test item numbers corresponding to the test items that you feel measure the objective. In some instances, you may feel that items do not measure any of the available domain specifications. Write these test item numbers in the space provided at the bottom of the rating form.

Objective	Matching Test Items
1	
2	
3	
4	
No Matches	

From R. K. Hambleton "Validating the test scores" (p. 225) in R. A. Berk (Ed.), *A Guide to Criterion-Referenced Test Construction,* 1984, Baltimore: The Johns Hopkins University Press.

that accommodates this quantification scheme.

A formula indicating that any particular item, k, is congruent with a specific skill or objective, i, can be applied to the judges' ratings. This formula is:

$$I_{ik} = \frac{(N-1)\sum_{j=1}^{n} X_{ijk} + N\sum_{j=1}^{n} X_{ijk} - \sum_{j=1}^{n} X_{ijk}}{2(N-1)n}$$

In this formula, I_{ik} is the index value, i and k are as described above, N is the number of skills or objectives, and n is the number of judges. The X_{ijk} is simply the rating assigned by a particular judge for the congruence between a given item and a specific objective. The Σ is of course the symbol for summation.

Although rather imposing at first glance, this formula is actually straightforward and can be easily worked through with a set of data. For example, suppose a test has 36 items that are intended to assess five specific objectives. For this example, suppose the item-objective congruence rating of interest is between the test's first item and the test's second objective. (In other words, "How well does item #1 match objective #2?") Now, imagine nine judges have rated the item for its congruence to the objective. One of the judges rated the item as a poor match (or -1), one of the judges rated the item as a moderate match (or 0), and seven of the judges rated the item-objective match as strong (or +1). The sum of the nine judges' ratings is 6 (i.e., (-1) + 1(0) + 8(+1) = 6). Applying these numbers to the item-objective congruence formula yields the following:

$$I_{ik} = \frac{(5-1)6 + 5(6) - 6}{2(5-1)9} = \frac{(4)6 + 30 - 6}{2(4)9} = \frac{24 + 30 - 6}{(8)9} = \frac{48}{72} = .67$$

This formula will yield an index score from +1 to -1. A +1 would be obtained if all the judges agree that there is a strong item-objective match. Conversely, if none of the judges agree that an item is matched to one and only one skill or objective, the formula will yield an index of -1.

For the item writer, this index can provide information useful for gauging the quality of an item in either of two ways: by using the item-objective congruence index either as a relative standard or as an absolute

standard. When the index is viewed as a relative standard, the statistic for any particular item is judged in relation to its power (i.e., its proximity to +1) compared to the power of the index for the other items considered. For example, if the index value were computed for each of 50 items and it proved to be comparatively low for, say, six, of the items, then the six would be suspect and should be reviewed.

To use the index in an absolute sense, the item writer must establish a criterion level for the index, above which items would be passed, and below which items would be reviewed for their success in fulfilling a test's objective. This criterion level may be set by deciding the poorest level of judges' rating that would be acceptable. In the computational example above there were nine judges. The criterion might be that at least seven of the judges should rate the item as strongly congruent to a given objective. This standard would yield a criterion for the index of .78. Hence, .78 would be the floor value for accepting the congruence for any particular match of an item with an objective. In the example above, the index was .67—below the .78 criterion—indicating that the content of that particular item should be examined by the item writer for its appropriateness as a measure of the objective.

Alternative Quantification Schemes

Klein and Kosecoff (1975) describe a variation of this tally method which includes examinees' performance data in the congruence process. This procedure may be slightly more rigorous than merely tallying judges' ratings; however, it is probably less precise than computing an item-objective index. Also, Polin and Baker (1979) offer an item-review scale consisting of six dimensions: domain description, content limits, distractor limits, distractor domain or response criteria, format, and directions plus a sample item. A rater reviews an item and scores it on each of the six dimensions according to prescribed criteria. This procedure is intended to provide information useful to discovering the relationship between "what test writers have wrought and the original test specification" (p. 2). Although an exploratory study of this methodology produced mixed results, the approach does represent a useful attempt to judge content-related evidence for validity.

USING LEADING QUESTIONS IN ITEM ANALYSIS

Judgmental approaches to analyzing items can be very effective methods for improving items. Obviously, the item-congruence index discussed in the preceding section is a judgmental approach, but there are other methods involving judgment as well, such as the technique of asking leading questions to appropriate people. For example, if the items are be included on a test for assessing school children, asking teachers to review and discuss items informally with the item writer is often worthwhile. No particular form is needed for this discussion, but some guiding questions might focus the task. Such questions might include the following:, "In your opinion, will this item be confusing to students?"; "Do you notice any language that might be offensive on stereotyping to students?"; "Do you believe the distractors are plausible?"; "Are there any vocabulary words in the item that you imagine will be unfamiliar to students at this age or grade level,"; "Are the graphics clear?"

Another group to whom the item writer can turn is persons in the same category as future examinees. If a test is being developed for, say, fifth-graders, the opinions of students who are in the fifth grade can be valuable. Again, the same guiding questions used with teachers can be used for discussion with students.

There are two principal disadvantages to this technique of informal discussion. First, it is often difficult to manage the logistics for such visits. They may be hard to arrange in schools, because teachers and administrators could view this as an interruption of valuable instructional time. Or, if the test is to be used for licensing, one frequently does not know beforehand who will take the test and an appropriate group cannot be easily identified.

The second disadvantage to the technique of informal discussion is that it consumes enormous amounts of time. It is the author's experience that in a typical 50-minute classroom period, perhaps only three or four test items will be discussed. This drawback might be alleviated if the writer were to select from a pool of freshly prepared items only a few for discussion, and then consider the other items based on what has been said by students about the selected items.

Regardless of the logistical difficulties in managing an informal discussion of items, the practice can be extremely valuable to improving items and has been used with students as early as the second grade. In fact,

students of all ages frequently demonstrate remarkable insight into the reasons for asking a particular question, and they can diagnose specific flaws in items and suggest improvements.

While informal discussion of items with an appropriate group is not a widely practiced technique for analyzing items, it is highly recommended. It can yield specific improvements in items and—perhaps even more importantly—it can help the writer become sensitive to examinees in ways that may be difficult to describe but are nevertheless extremely valuable.

ITEM STATISTICS

A number of statistics can be used to indicate particular features of test items. Researchers (Crocker & Algina, 1986) categorize these indices by the parameters which are commonly examined:

1. Indices that describe the distribution of responses to a single item (i.e., the mean and variance of the item responses),
2. Indices that describe the degree of relationship between response to the item and some criterion of interest, and
3. Indices that are a function of both item variance and relationship to a criterion.

Some of the commonly used statistics for describing these item parameters are *p*-values, *variance,* and a variety of *item discrimination* indices, such as the point-biserial estimate of correlation, the biserial correlation coefficient, and the phi correlational estimate. Each of these statistical indices is important for a specific purpose in item analysis, and each will be described.

The Proportion Correct Index

Probably the most popular item-difficulty index for dichotomously scored test items is the *p*-value. The *p*-value is merely a shorthand way of expressing the proportion of examinees who responded correctly to a particular item. It may be most clearly explained by using a few simple examples. Suppose a given item was administered to 100 examinees, and 80 of them responded correctly. In such a case, the *p*-value is .8, indicating that 80 percent of the examinees responded correctly to the item. If another item were administered to a group of 311 examinees and 187 of them

responded correctly, the *p*-value would be .60 (i.e., 187 ÷ 312 = .60). Table 8.3 presents a simple table of *p*-values for a hypothetical five-item test.

Table 8.3 *P*-Values for a Five-Item Test

Item Number	*p*-value
1	.68
2	.45
3	.91
4	.36
5	.48

The *p*-value for an item does not by itself indicate whether an item is good or bad; rather, it merely represents a difficulty index relative to the particular group of examinees to whom the item was administered. If the same item were administered to a different group of examinees, the *p*-value would probably not be identical.

The group-referent aspect of the *p*-value is termed *sample dependence,* meaning that any given *p*-value is dependent upon the particular group to whom the item was administered. This sample dependence characteristic for the index can be easily seen by imagining that a given item is administered to third-grade students and again to sixth-graders. Since these two groups are not from the same total population, the *p*-values yielded by the two groups would probably be different. The item would have two *p*-values, one indicating its difficulty relative to third-graders, and the other showing its difficulty for students in the sixth grade. Each *p*-value depends upon which sample of examinees is tested. This is what is meant by saying the *p*-value is sample dependent.

Of course, if a sample of examinees is carefully selected to represent a larger population, then the *p*-value for the sample can be interpreted as similar to a value that would have been obtained had the item been administered to the entire population. (Subject to the limits of the sampling design.) Conversely, if a group of examinees is not selected from a larger population by a method that allows generalizations, then the index is meaningful only to that particular group of examinees.

It is important to understand the sample-dependence feature for *p*-values because it is one of the most salient aspects of *p*-values. As we shall

see later, other statistics have been developed which attempt to define a difficulty index independent of a particular sample of examinees.

Using *P*-values for Item Analysis

P-values are of enormous help to a writer during item analysis. By understanding *p*-values and interpreting them correctly, the writer can see how an item is performing in relation to a given group of examinees, as was displayed in Table 8.3. However, they can also be used to give the item writer a more complete description of an item's performance. For instance, *p*-values can aid in detecting some common writers' mistakes, such as making apparent wording in an item that is evidently confusing to examinees, recognizing flawed distractors, and identifying inadvertently miskeyed items. The index can allow the writer to see how an item is performing for examinees within the same population who differ in ability. Each of these uses for the *p*-value will be explained in turn, but first, the *p*-value must be displayed in a manner that facilitates such interpretations.

While Table 8.3 presented *p*-values for a group of items in a test, in item analysis work, *p*-values are typically displayed singly. Additionally, for item analysis, *p*-value is computed for each response alternative: the correct option and all of the distractors. Further, when *p*-values are displayed in this manner, it is customary to also report the number of examinees who omitted the item. Table 8.4 presents an example of *p*-values reported for every response alternative to a single test item as well as the number of examinees who omitted the item.

Table 8.4 P-Values for a Single Test Item

	A	B	C*	D	omits	Total
Number	28	17	197	41	3	286
p-value	.10	.06	.69	.14		

*correct response

As can be seen in Table 8.4, the *p*-values reveal much data useful to writers trying to improve items. In this instance, the correct response is indexed at .69, indicating that for most tests of general achievement the item is neither too difficult nor too easy for this group of examinees. Also, as a group, examinees have selected all of the distractors, which may signify that none is rejected out-of-hand. Often, this is a good indicator of the quality of an item; however, since comparatively few examinees (six percent) selected option B, it may be worthwhile to examine this distractor for possible improvements.

It is common in most general assessments of academic achievement for a test developer to specify limits for item difficulty. Often, such limits for *p*-values are that the set of items on the test can range from a low of .40 to a high of .80. There are, of course, many instances when other limits for item difficulty will be appropriate. The writer should be aware of the test developer's limits for item difficulty values, as they can guide the writer in deciding which items should be reviewed.

Now, note the *p*-values for another item, displayed in Table 8.5. In this instance, *p*-values reveal several flaws in the item. Apparently, examinees find this item confusing since their responses are widely scattered among the response alternatives. The correct response, A, attracted fewer examinees (27 percent) than did one of the distractors, C (45 percent). And distractor B was apparently so implausible than no one responded to it. Also, a comparatively large number of examinees omitted responding to the item, which may be another indication that the item is confusing. Taken together, these findings suggest a seriously flawed item.

Table 8.5 *P*-Values for a Poor Test Item

	A*	B	C	D	omits	Total
Number	77	0	130	63	16	286
p-value	.27	.00	.45	.22		

*correct response

Lest the reader imagine that an item so fraught with error would not be produced by serious item writers, the reader should be aware that flaws in items frequently do not become apparent until they are subject to the scrutiny of item analysis. Remember, constructing good test items is not a process that ends with the initial writing; it demands rigorous scrutiny of the item as well.

Miskeyed Items

Another useful feature of *p*-values for item development work is for identifying items that have been miskeyed. Regretfully, miskeying items is all too common in item preparation. In many instances it is an understandable lapse. Sometimes the tedium of producing many items can cause writers to be lax in attending to detail, and an item will be miskeyed. At other times, the ambiguity of ill-conceived or poorly worded items is not immediately noticed, and the correct response may not be apparent. Occasionally, when a particular item is designed to assess a complex cognitive-processing skill, the subtleties of language or the difficulty of the content result in miskeying items.

Miskeyed items often become obvious when the item writer examines a table of *p*-values that display widely different results from what had been anticipated. For example, Table 8.6 displays statistics for a sample test item in which the item writer perceived one correct response (B), but examinees uniformly selected another response alternative (A). In this case, the item is a good item, but one that had been miskeyed.

Table 8.6 *P*-Values for a Miskeyed Test Item

Response Alternative	A	B*	C	D	omits	Total
Number	202	31	28	25	0	286
p-value	.71	.11	.10	.9		

*correct response

Comparing *p*-values for High- and Low-Achieving Subpopulations

Sometimes it is useful to compare subgroups of the examinee population to determine how an item is performing. For this analysis, the population is often divided into two groups, a high-achieving group and a low-achieving group. Typically, the groups are examinees whose total score on a test comprise the top 27 percent of all examinees, and those whose scores place them in the bottom 27 percent of the examinees. The figure 27 percent is chosen because it is used in some computational algorithms for determining internal reliability indices and Kelly (1939) demonstrated that this number will provide a stable index of differences between high and low ability groups. For this analysis, the principal focus is on determining how well the item is functioning for the extremes of the ability range. Table 8.7 displays item data for this analysis.

Table 8.7. *P*-Values for Examinee Subgroups on a Single Test Item

Response Alternative	A	B*	C	D	omit
upper 27%	.29	.61	.08	.02	1
lower 27%	.31	.27	.31	.11	6
difference	-.02	.34	-.23	-.09	

*correct response

Note that the item statistics displayed in Table 8.7 reveal that the difference between the examinee subpopulatioris is 34 percent for the correct response. This is a rather large difference which may signify that the item was not especially difficult for high-ability examinees (the top group) but was quite difficult for low-ability examinees (the bottom group). For many kinds of assessments, this difference is a desirable feature for an item. Also, heed the fact that the difference between the groups for all other response alternatives is a negative value. Such negative values indicate that fewer high-ability examinees selected the distractor than did low-

ability examinees. Again, usually, this is a positive feature for an item. Still, writers should pay attention to the fact that the difference between the groups for response alternative A is a scant .02 percent. This suggests that option A should be reviewed.

Comparing Several Subpopulations

Frequently in item analysis several segments of a population of examinees are compared. Each segment, or subpopulation, of the total group of examinees represents an ability stratum. Ideally, the examinees are grouped into segments, or subpopulations, by their performance on an outside criterion, such as another measure of analogous content with similar reliability. In practice, however, such external measures are rarely available; hence, the test itself is usually used as a measure of examinees' ability. For this purpose, the total test score is used.

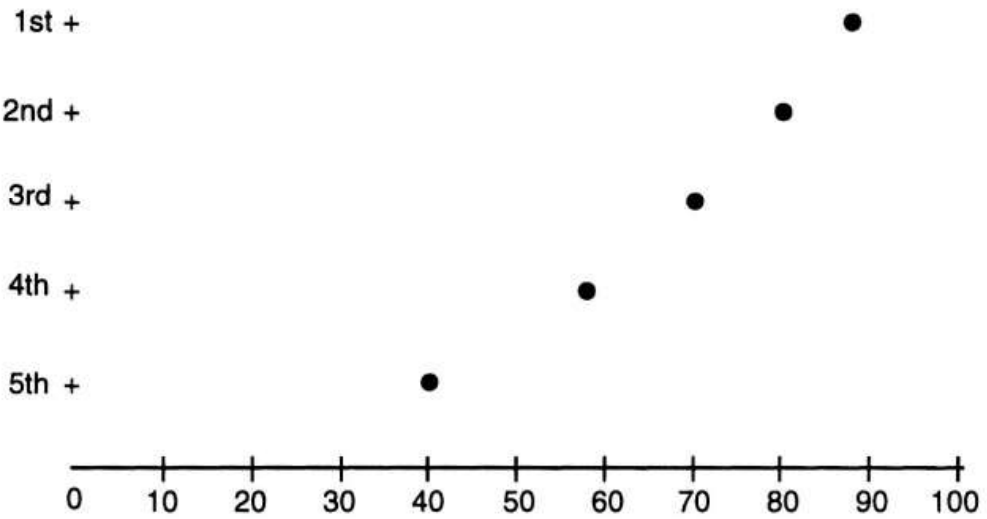
When the total population to whom the test has been administered is large (about 200 examinees or more), typically five groups are formed, each representing about 20 percent of the distribution of scores on the test. A sample of responses reported by fifths of the population distribution is presented in Table 8.8, which will be described momentarily.

The procedure for splitting the population is straightforward. First, a frequency distribution of scores is prepared. Then, exact scores are noted at the 20th, 40th, 60th, and 80th percentile points, yielding five ranges of scores. Test scores within each of these ranges become the criterion into which particular examinee scores are grouped. Because most populations exhibit skewed distributions, not every group will have precisely 20 percent of the examinee population, but most groups should be fairly close to 20 percent.

Often, when the examinee population is divided into fifths, it is useful to display *p*-values graphically. Such a representation makes it easy to identify the relative position of each segment of the examinee population. Flaws in items that may go otherwise unnoticed, are revealed by viewing the number of examinees for each distractor for the various subpopulations, as well as the *p*-values. Table 8.8 displays a graphical representation of an item with the corresponding numbers for each subpopulation and overall *p*-values.

Notice in 8.8 that the top fifth of the population (that is, the highest scorers on the total test) also achieved the highest number of correct

Table 8.8 Graphical Representation of an Item Including Item Statistics.



	A	B	C *	D	o
top fifth	8	17	207	11	0
next fifth	11	39	223	25	0
next fifth	16	53	183	32	0
next fifth	17	61	113	29	0
lowest fifth	43	70	106	39	2
p-value	.07	.18	.64	.10	

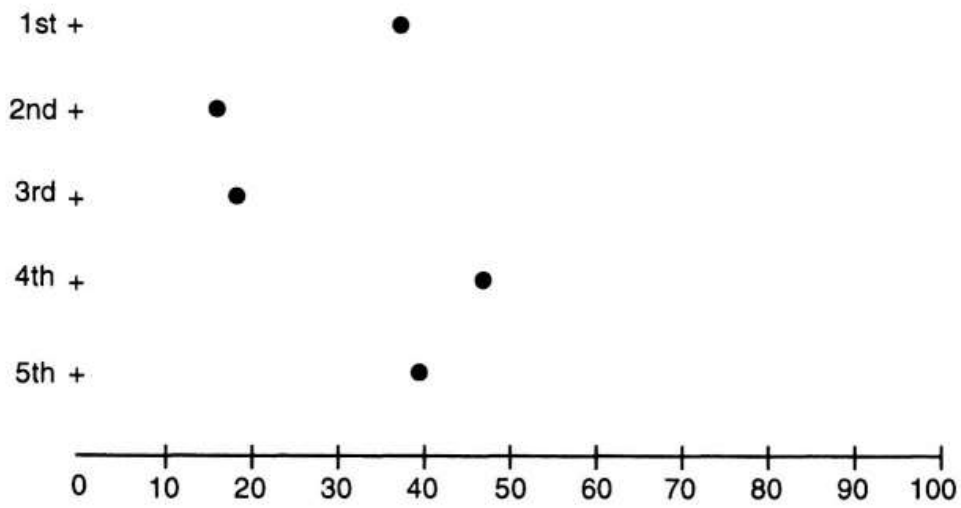
* correct response

responses on this particular item, followed by the next fifth who achieved the next-highest number of correct responses, and so forth, until the lowest fifth is shown achieving the lowest number of correct responses. For a group-referenced interpretation, this item seems to be behaving very well since one would anticipate that the examinees who are most able in the tested construct would also have the greatest proportion correct on any particular item. Notice also that the lower-ability groups increasingly chose an incorrect response alternative.

Table 8.9 similarly displays data for a different item. Notice, here that the item is not performing very well and needs to be revised. As can be

seen, no examinee, regardless of ability, selected distractor A. It adds no information to the test and should be dropped or modified. Also, more examinees at every level of ability selected response alternative C rather than the intended correct response, B. This could indicate poor wording. And, as shown in the graphical representation, the two lowest achieving groups outperformed the highest achieving group. This is a clear sign to examine the item for confusing wording or for observing the phenomenon of proactive inhibition (a concept discussed in Chapter 4). In either case, this item needs major repair. It does show, however, an example of using graphical representation and numerical analysis for improving items.

Table 8.9 Graphical Representation of a Poorly Performing item.



	A	B*	C	D	or
top fifth	0	27	111	21	0
next fifth	0	23	116	21	0
next fifth	0	16	123	13	0
next fifth	0	10	83	32	0
lowest fifth	0	8	97	24	2
p-value	.00	.11	.73	.15	

* correct response

Item Discrimination Indices

Discrimination is another important concept for judging the quality of items. Actually, we were examining discrimination for items in the preceding section, but it may not have been conceptually understood. Discrimination for items may be conceptually understood as the relationship between the difficulty of an item and the ability of the examinees. Simply put, *item discrimination* is an index for determining differences among individual examinees on the subject matter or psychological construct being assessed. It relies upon a fundamental assumption, which is that examinees who exhibit mastery of the subject or high ability in the construct are presumed to be more likely to answer any particular item about that subject or construct than examinees who exhibit low mastery or ability. Conversely, items that either all examinees responded to correctly or all examinees missed do not discriminate. Items that do not discriminate yield no information about differences between individuals.

Mathematically, item discrimination defines an item's difficulty as a function of the examinee population's ability in the construct being assessed. In other words, discrimination is related to difficulty for a particular ability. The relationship may be readily seen by examining the graphical representation of four items. Figures 8.2 to 8.5 display such graphical representations of items at four levels of item discrimination: high dis-

Figure 8.2. Characteristics of a highly discriminating test item.

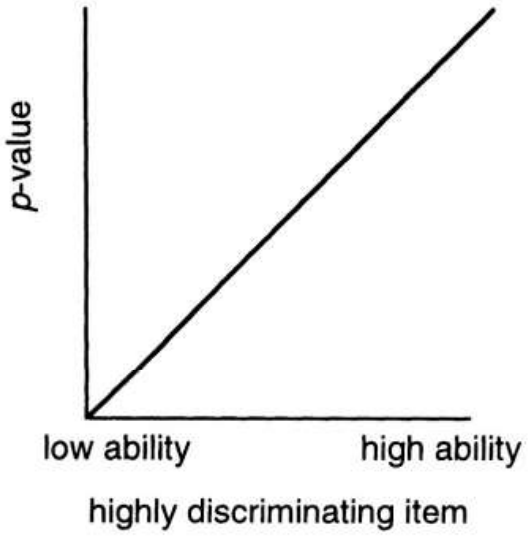
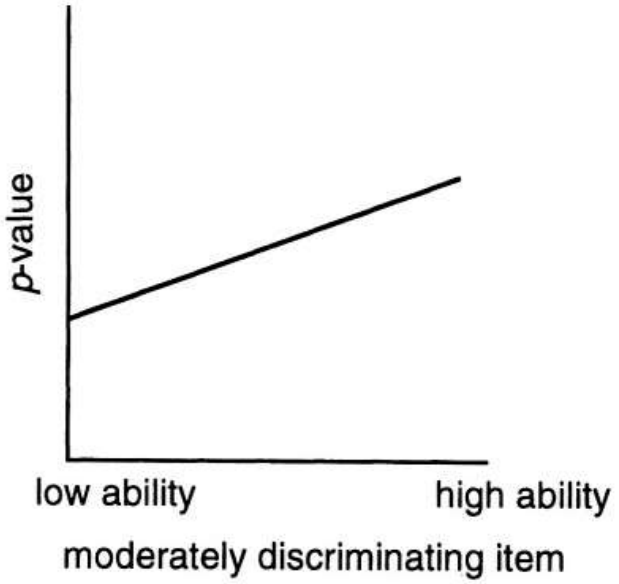


Figure 8.3. Characteristics of a moderately discriminating test item.



crimination, moderate discrimination, no discrimination, and negative discrimination.

The item represented in Figure 8.2 would usually be considered to be a good item because it is highly discriminating. It distinguishes among examinees who are of high ability and got the item correct and those who are lower in ability and did not respond correctly to the item. Notice in 8.2 that as examinee ability increases, there is a corresponding increase in the difficulty. The data shown earlier in Table 8.4 would be for this highly discriminating item.

Figure 8.3 presents data for an item which discriminates moderately well. It shows differences among examinees but not as sharply as the item in 8.2.

Figure 8.4 displays a graphical representation of an item that shows no differentiation between high and low achievers. Probably, this is an item that either all examinees got correct or all incorrect. Typically, such items add little or no information to a test and are rejected. However, zero discriminating items should not be rejected summarily. There are some items that assess skills so important to a particular test's objective that one would expect all examinees to respond correctly.

Figure 8.4. Characteristics of a non-discriminating test item.

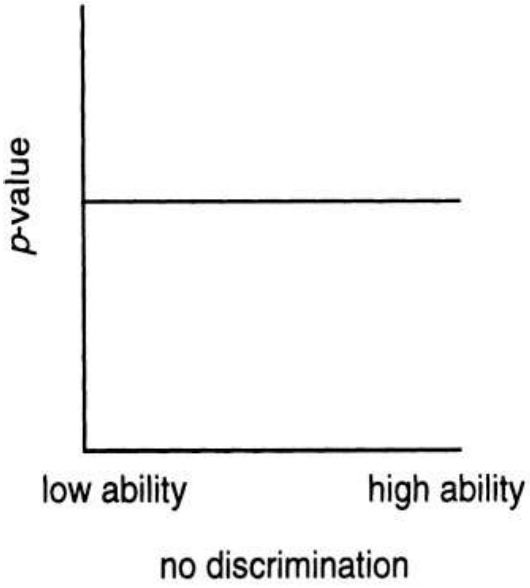
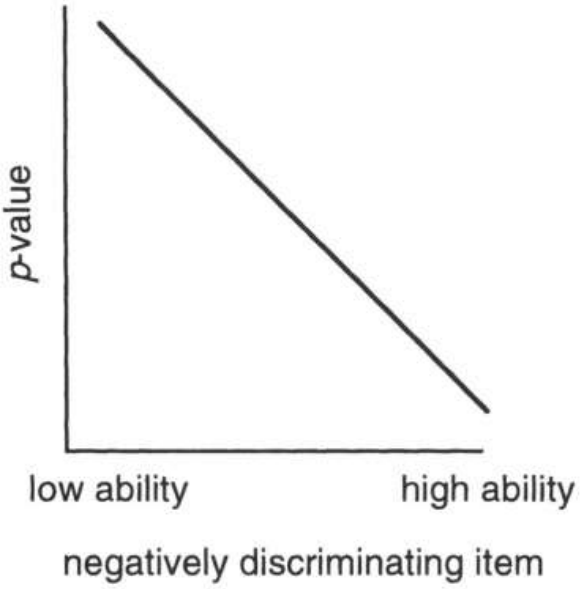


Figure 8.5. Characteristics of a negatively discriminating test item.



Finally, Figure 8.5 is a graphical representation of a negatively discriminating item. In items of this type, more low-ability examinees answered this particular item correctly than did high-ability persons. This usually means that the item is poorly worded or in some way confusing to examinees who have greater knowledge of the content. Here, it is likely that proactive inhibition is operating for the item. Negatively discriminating items are almost always in need of repair by the writer, or they should be discarded.

Although, generally speaking, discrimination is a positive item attribute, judgment must be used in deciding when an item discriminates optimally. In some easy mastery-type items, it is appropriate for items to be highly discriminating at just one level of ability, as was seen in the earlier example of an item on an airline pilot's certification test in which examinees are requested to identify the radio broadcast frequency used by the control tower at Los Angeles International Airport. At other times, very good items may discriminate less restrictively.

Still, such judgments about optimal discrimination could be arbitrary without guidance. Fortunately, several statistical procedures are available that can quantify the discrimination of an item. These are especially useful statistics in item analysis because they often will guide the item writer to specific items needing improvement.

The Point-Biserial Measure of Correlation

One index of discrimination is the *point-biserial correlation coefficient.* As a measure of correlation, the point-biserial coefficient estimates the degree of association between two variables: a single test item and a total test score. As before, for most purposes of item analysis, the total test score is considered a reasonable measure of examinees' ability. It is often the only measure of ability available to the item writer when he or she is considering the quality of items. When the test item is inherently dichotomous (i.e., scored in only one of two possible categories, such as correct or incorrect) and the total test score is inherently continuous (that is, the scores range from low to high), the point-biserial statistic is most useful for examining the relative performance of an item between two groups.

The point-biserial estimate of correlation is a *product-moment correlation coefficient.* To understand this term and its advantages for analyzing items, one must realize that moments are thought of as standard score deviations about a mean. The deviates themselves are referred to as the

first moments of a distribution; the squared deviates are the second moments; cubed deviates are the third; and so forth. Since standard scores have a mean of zero, one standard score away from the mean is the first deviate. By this conceptual approach, and by calculus, point-biserial coefficient of correlation is the association of two sets of standard scores. The advantage of this arrangement of data is that the relationship between an item score and a total test score is on a common base so that they may be meaningfully correlated.

With this statistic, the item writer can gauge the discriminating value of a test item. For example, consider the data displayed in Table 8.10. Here, several test items are arranged in descending order of their difficulty. The discrimination value for each item, as established by the point-biserial statistic, is also cited. Generally speaking, items with higher point-biserials are more highly discriminating. Conversely, items with relatively low point-biserials are less discriminating. As a general practice, items with negative point-biserials are either dropped from further consideration or revised.

Item writers will use the information yielded by the point-biserial correlation in conjunction with *p*-values to examine the quality of particu-

Table 8.10 Difficulty and Discrimination for a Hypothetical Test

Item	*P*-Value	Point-Biserial
3	.94	-.09
1	.86	.06
16	.75	.12
9	.73	.45
2	.68	.15
11	.62	-.21
15	.60	.31
8	.55	.46
8	.51	.45
14	.28	-.21

lar items. When an item has a difficulty index within a range deemed appropriate for the goals of a test (often, $.40 \leq p \leq .80$) and is relatively highly discriminating, the item is judged to be sound. For example, notice in Table 8.10 that the first item (test item #3) is too easy for our criterion and also discriminates poorly. Hence, this item is diagnosed as flawed and needs review. Item #14, the most difficult item in the set, also needs review. It is apparently confusing to most examinees since the few examinees who got the item correct came from no apparent ability group. Item #9, however, seems to be operating well; its difficulty is within the acceptable range, and its relative discriminating power is high.

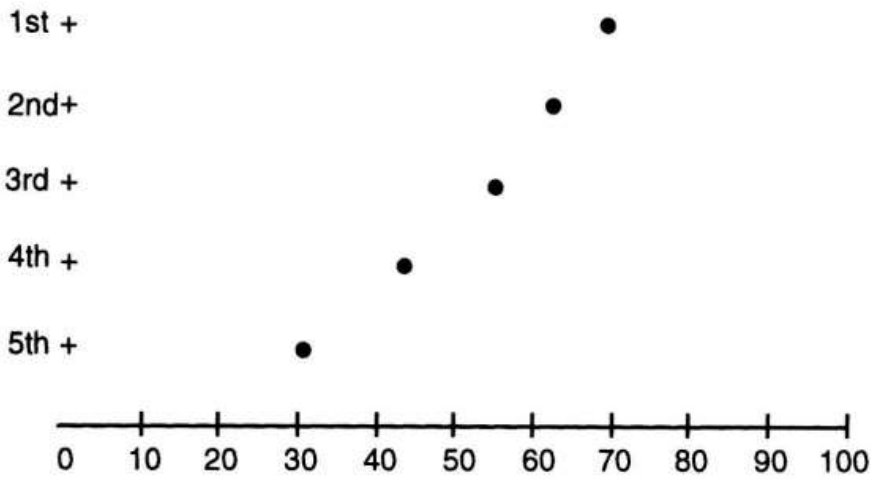
In using the point-biserial coefficient of correlation for item analysis, it is often helpful to contrast the coefficient for the correct response alternative with that of the distractors. Accordingly, computing this statistic for each of the response alternatives—the correct response as well as the distractors—is commonly done. Discrimination for a single item is presumed to exist if the coefficient for the correct response is a positive number while the same statistic for the incorrect response alternatives is negative. Table 8.11 displays the statistics for a test item that exhibits this circumstance. By this criterion, it is a good test item since it exhibits a high level of discrimination.

Shortcomings of the Point-Biserial Estimate

Despite the fact that the information yielded by the point-biserial correlation is often used in item analysis, the statistic is not problem-free. With only casual analysis, one can realize that the particular item score being analyzed has itself contributed to the total test score, or ability measure. This leads one to wonder whether the information is actually spurious and may therefore be misinterpreted. Logically, this point is correct, but in practice it is not a serious problem since the effect of a single item on the total score for a set of items is minimal, especially when the number of items is comparatively large, say, 25 or more.

When very precise estimates are required, or when the number of items is fewer than 25, the point-biserial estimate may be corrected for spuriousness. The formulas for calculating the correction are not especially complex, but they are computationally long, and they are seldom employed for the purposes of item analysis. (They are used in some multivariate analyses.) Therefore, they are not described here; instead, the

Table 8.11 An Item Displayed with *P*-Values and Point-Biserial Indices



	A	B	C*	D	omits
top fifth	21	3	193	47	0
next fifth	15	4	145	59	0
next fifth	30	11	161	83	0
next fifth	46	23	131	87	0
lowest fifth	50	38	84	81	1
p-value	.12	.06	.54	.27	
point-biserial	-.14	-.20	.29	-.11	

*correct response

interested reader may readily find them in any of several sources (e.g., Allen & Yen, 1979; Henrysson, 1963; Nunnally, 1978; Thorndike, 1982).

A further problem with the point-biserial correlation coefficient is that when the distribution of scores in the total test group is continuous, the range for the statistic is restricted to less than +1 and greater than -1. In fact, the point-biserial range is a function of the point at which the ability groups are split. This anomaly of numbers can be most easily comprehended when one considers the point mentioned above that very easy items and very difficult items provide relatively little differentiation between high-ability examinees and low-ability examinees. Despite these technical

limitations (which will be elaborated upon momentarily), the statistic remains useful for item analysis and is the generally preferred statistic for examining item discrimination by traditional item analyses.

The Biserial Estimate of Correlation

Another statistic that can be used for analyzing items for discrimination is the *biserial estimate of correlation.* It is closely related to the point-biserial correlation, with an important difference. The distinction between these two measures exists in the assumptions. Whereas the point-biserial statistic presumes that one of the two variables being correlated is a true dichotomy, the biserial estimate of correlation assumes that both variables are inherently continuous. Further, the assumption is made that the distribution of scores for both variables is normal. For computational purposes, however, one of the variables has been arbitrarily divided into two groups, one low and the other high. In item analysis, the two groups are examinees who responded correctly to a given item and those who did not.

For practical purposes in item analysis, the strength of this statistic lies in its ability to overcome the limitations of the point-biserial statistic discussed above. When it can be assumed that the construct being assessed is normally distributed among the examinee population, the biserial range is limited from -1 to +1 absolutely. This means that examinees of either very low ability or very high ability are better represented in the correlational estimate. Thus, if the writer is considering items at the extremes of the difficulty range, the biserial estimate of correlation is preferred to the point-biserial statistic.

The Phi Coefficient

The *phi coefficient of correlation* is another estimate of a correlational relationship that can be used for analyzing test items. Like other correlation coefficients, it yields an estimate between +1 and -1. However, it differs from the two previously discussed correlation estimates because it assumes a genuine dichotomy in both variables to be correlated. The principal focus of the phi coefficient is to determine the degree of association between an item and some criterion, like some program feature, gender, or some other demographic characteristic.

During item analysis, it is often convenient to correlate two items, giving the effect of treating each item as a criterion for the other item. For this analysis, it is necessary to present the data for the two items in terms of

the joint proportion-response distribution for the two items. Table 8.12 displays this data for two items in which 30 percent of the examinees got both items correct and 20 percent answered neither item correctly.

For analyzing items, the value of the phi coefficient lies in its utility for comparing the degree of stability in responses to the same item by examinees at different points in time. For example, if the item writer wishes to consider whether some variable such as gender is correlationally related to how a group of examinees perform mathematics items from differing test administrations, the phi coefficient is the appropriate statistic. This information is especially useful to writers wishing to improve items on a test that is undergoing revision.

Table 8.12 A Four-fold Table Presenting Responses to a Single Item Administered Two Times.

Item # 2 \ Item # 1	pass	fail	
pass	.30	.17	.47
fail	.33	.20	.53
	.53	.47	

Using the Phi Coefficient With Pre- and Post-Instructed Groups

One technique of analyzing items is to compare the performance of two groups on the same items. One group is selected as the "criterion group"—that is, an appropriate group whom one expects to be able to respond correctly. In the case of mastery-specific learning, the criterion

group would be the group who has received instruction. For any item, then, the performance of the uninstructed group is compared against that of the criterion group. Ideally, both groups would be tested simultaneously, and the only difference between them would be the specific instruction. The phi coefficient can be applied to analyzing differences between the groups.

This technique has been in use for some time (e.g., Cronbach & Meehl, 1955; Klein & Kosecoff, 1976; Millman, 1974b); however, it has substantial limitations. First, and perhaps most obvious, is the fact that rarely are instructional programs so well organized around a single set of clearly elaborated objectives that the item writer could identify two appropriate groups. This is a problem of establishing a criterion.

A second disadvantage is the procedural difficulty of testing two similar groups, one of which has received the instruction while the other has not. This can lead to an adjustment in which the same group is tested before and again after instruction. However, this adjustment introduces a lengthy, and often impractical, delay between testings. Despite these difficulties, when a criterion can be reasonably set and appropriate groups are available, this is a strong techniques for analyzing items.

Shortcomings of the Phi Coefficient of Correlation

One especially significant limitation of the phi coefficient occurs because this estimate of correlation, like its relatives the point-biserial and biserial measures of association, is derived from the traditional Pearson coefficient of correlation. Since all product-moment correlations are expressed in the form of a standard score, two variables with identical standard scores will necessarily correlate perfectly. Hence, the phi coefficient will always be exactly +1.00 when the p-value for the two groups are equal.

ITEM PARAMETERS

It has been emphasized throughout this book that the modern view of test items considers not only the particular subject content addressed by an item, but also the psychological construct that examinees must employ to solve the problem. The discussion of the purpose for test items in Chapter 2 provides a thorough treatment of the rationale for this approach, and most of Chapter 3 was devoted to an explanation of putting this theory into

practice. This theory of a latent trait approach to testing has a serendipitous effect for item analysis: the ability to present test items graphically in a way that makes it easy to view several important characteristics simultaneously. This is accomplished by mapping *item trace lines,* or functions, for test items. Technically, Figures 8.2 to 8.5 in this chapter can be considered item trace lines, but in practice item trace lines are more often computed for items analyzed with latent trait approaches.

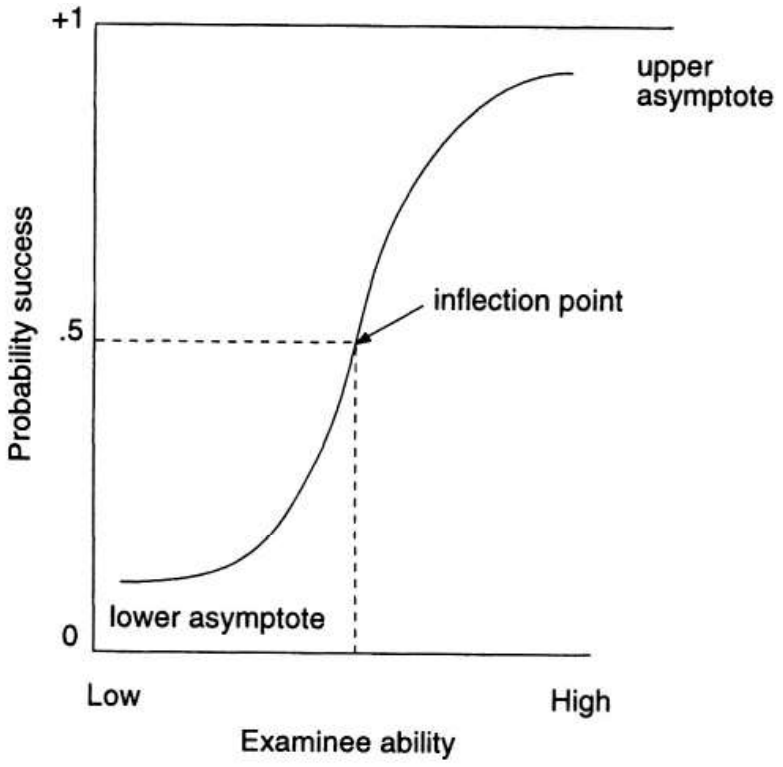
Item trace lines are typically called *item characteristic curves* (ICC), and they present information about one, two, or three *parameters,* or mathematical boundaries, for each item. Generally speaking, these parameters are

1. Parameter A, indicating the "steepness" of the item trace line and representing the probability of responding correctly to an item increasing as one goes up the scale as a measure of discrimination among varying ability levels,
2. Parameter B, defining the difficulty of the item by noting the point at which a latent variable (e.g., psychological construct) falls—this is also the left-to-right shift of the curve—and, sometimes,
3. Parameter C, showing the beginning, or base, of the curve, suggesting the probability of guessing (also called "chance" or "pseudo-chance") a correct response on the item for very-low-ability examinees.

Although item characteristic curves were briefly mentioned in Chapter 2, it will be useful to display several curves here for the purpose of examining these particular item parameters. Figure 8.6 presents a item characteristic curve. Note that the graph plots "percent success" along the ordinate (Y axis) and the examinee attribute (viz., ability) along the abscissa (X axis). Three other features of note are: 1) the slope of any curve is monotonic, that is, it always rises and is never exactly horizontal; 2) an "inflection point" (which can be shown by drawing a horizontal line from a point on the curve to the Y axis) is determined by the left-to-right shift of the curve (Parameter B); and 3) the two asymptotes, lower and upper, may approach but never actually reach 0.00 and 1.00 respectively. An item trace line is technically termed a *monotonic normal ogive.* Ogives are merely a specialized graphical representation of a frequency distribution.

Because all three features—discrimination, difficulty, and guessing (or pseudo-chance)—for an item can simultaneously be displayed graphi-

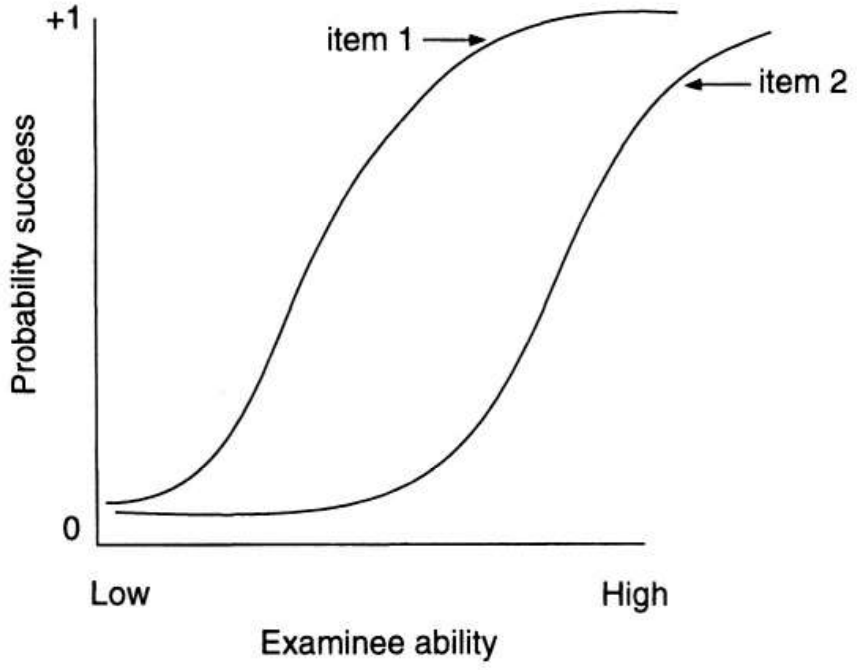
Figure 8.6. Trace line for a single item.



cally, ICCs are especially useful for analyzing items. Accordingly, it will be instructive to examine the curves for several items so that differences among them can be noted and their utility discussed. Figure 8.7 displays ICCs for two items which are similar in many respects but differ in difficulty.

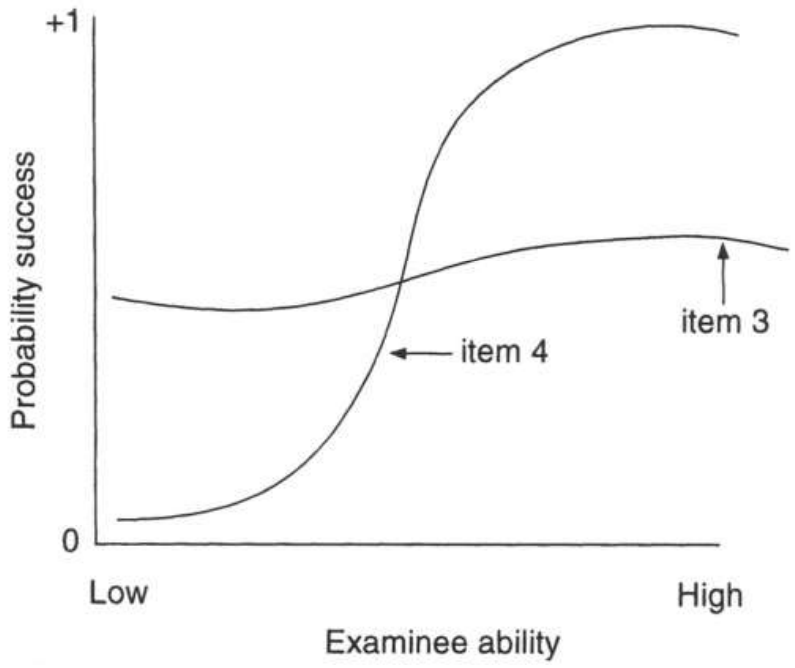
Notice in the figure that items 1 and 2 have similar shapes, indicating that the two items discriminate at about the same rate; however, since the curve for item 2 is shifted further to the right than that for item 1, item 2 discriminates at a higher level of ability. One can conclude, therefore, that these two items have equal discriminating power but that item 2 is a much more difficult item. The informed item writer could use this information to decide whether item 1 or item 2 would be appropriate for a particular group of examinees. For example, item 1 might be appropriate for use with average-achieving third-graders, whereas item 2 may be appropriate only for especially able third-graders or perhaps for fourth-graders.

Figure 8.7. ICC of two similar items of different difficulty.



Next, examine ICCs for two different items, items 3 and 4, in Figure 8.8 Notice that the ICC for item 3 is very flat, indicating that the item discriminates very little, regardless of the ability level of examinees. Typically, items that display the characteristics shown in item 3 are poor and need repair or elimination. Last, notice the very steep slope of the ICC for item 4. This item discriminates very well, but at only one point along the ability continuum. Under certain circumstances such sharp discrimination is appropriate, but more often item writers will consider this slope too steep for making distinctions among examinees and will repair or discard the item. For most tests, item writers will seek items whose ICC is of the smooth, lazy-S form displayed in Figure 8.7.

Figure 8.8. ICCs for two different items.



Examining Item Characteristic Curves with Item Response Theory

It should be clear from the discussion thus far about item characteristic curves that they are merely a specialized form of a frequency function and that they can be computed using the methods described; but, this is not the only way to compute item characteristic curves. A much more elegant, but technically complex, approach is to use the rationale and methods of latent trait theory to examine ICCs. Simply put, latent traits are examinee characteristics, or hypothetical constructs, that cause a consistent performance on a test of any given cognitive skill or achievement or ability.

Latent trait theories have been developed and applied under several rubrics, but we shall use the one that most clearly emphasizes the psychologically based nature of latent trait theories, *item response theory* (IRT). It is from IRT that the item ICC may be most meaningfully used in item development. However, while theoretically satisfying, ICC techniques are also the most difficult to understand conceptually and are extremely com-

plex procedurally. Computer processing of data, often involving large samples of examinees, is the only practical way IRT methods may be exploited. Nonetheless, in those instances when IRT is available to the item writer, it can provide powerful data for item analysis.

The issues involved in item response theory are too encompassing to detail here, nor are computational algorithms germane; rather, the focus here is on understanding how to use ICCs that have been computed by IRT methods for item development. Several excellent descriptions of the theory and methods of IRT are available. A technical introduction is given by Birnbaum (1968), and developments in latent trait theory and related issues are discussed by Hambleton (1979), and by Hambleton and Swaminathan (1985). A primer of IRT is given by Warm (1978); also, an excellent introduction to the models of IRT is provided by Crocker and Algina (1986). As one can easily imagine, a vast and growing body of literature is devoted to this important topic.

ITEM BIAS

Item bias is a particularly significant topic in reviewing test items for quality because it is used by those who argue that tests are unfair, inconstant, contaminated by extraneous factors, and subject to misuse and abuse. For this reason, in addition to the focus on improving items generally, the careful item writer will pay special attention to bias in items. As with many other concepts in writing items, bias is not something inherent in test items; rather, it arises from specific sources of error variance. Hence, addressing bias in items involves searching for a particular kind of error variance and then seeking to eliminate or reduce the sources of error.

Further, bias can be either internal to a test or external. Internal bias is concerned with particular characteristics of items, and it shall be our only focus. External bias is more a matter of determining the appropriate uses for test scores, such as for selection or placement into programs, and usually involves an entire test rather a particular item. External bias, by definition, sets the score of a test in a statistical comparison with a criterion. While a very important topic to test developers and test users, external bias is not a matter that the writer of items can directly address without reference to the broader issues of entire test scores and their reference to outside measures; therefore, it will not be covered in this discussion. For a thorough treatment of external bias, the interested reader

may refer to Jensen's *Bias in Mental Testing* (1980) and to a special issue of the *Journal of Educational Measurement* (1976) devoted to the topic.

For the writer to examine bias in individual test items in a meaningful way, he or she must understand fully the concepts covered by the term and must also realize what is not included. To begin, item writers should understand that the term *bias,* when used in item analysis, is conceptually distinct and operationally different from the concepts of fairness, equality, prejudice, preference, or any of the other connotations sometimes associated with its use in popular speech. In item analysis, *bias* is a technical term whose meaning will become evident momentarily. The reader interested in exploring the sociological aspects of bias in tests can consult any number of books, journals, and periodicals for discussion (e.g., *Ability Testing: Uses, Consequences and Controversies, Parts I and II*, National Research Council, Committee on Ability Testing, 1982; and *Bias in Mental Testing,* 1980).

As has been mentioned, in item analysis bias has a precise, mathematical definition. According to Osterlind (1983), "bias is defined as a systematic error in the measurement process. It affects all measurement in the same way, changing measurement—sometimes increasing it and other times decreasing it.... It is a technical term and denotes nothing more or less than the consistent distortion of a statistic" (p. 10-11).

This mathematical definition for bias may be readily understood when one examines a common occurrence of systematic distortion at the U.S. Bureau of Standards, the official store for U.S. measurements. At the Bureau, measurements of weight are kept in two metrics: the kilogram and the K20. It has been empirically determined that these two measures are not precisely equal. The K20 is estimated to be 19 parts in a billion lighter than the kilogram. Therefore, all measurements at the Bureau done by K20 are systematically off (or biased) by this very small amount. Since some measurements require extreme accuracy, the Bureau compensates for this measurement bias by revising K20 measurements up by 19 parts in a billion. Regardless of compensating remedies, the example shows a systematic error in measurement, or bias.

In test theory, an item is said to be *unbiased* when the probability of success on the item is the same for equally able examinees regardless of a particular subgroup's membership. In other words, if an item is designed to assess reading comprehension for all fifth-graders, any two children from this population *who are of the same ability* should have an equal chance of

responding correctly to the item. Characteristics beyond the fact that they are both fifth-graders should not matter, whether gender, ethnic heritage, or whatever. If members of one subpopulation consistently score lower or higher than members of another subpopulation (assuming, of course, that individual members between the groups have equal ability), then there is a consistent distortion, and bias exists.

One naive but widely held notion concerning bias should be quickly dispelled: Bias is not the mere presence of a score difference between groups. It is grossly inappropriate to simply compare *p*-values between two groups, note that the *p*-value for one group is higher than for the other, and conclude that bias is present. Were this true, every item on every test could be "biased" against or in favor of one subpopulation or another, and "bias" could be repeatedly inferred by merely redefining the groups. The logic of this argument would have every item "biased" for tall persons, or overweight persons, or either of the two genders, or persons of one or another ethnic heritage, or any other variable that could be named. This thinking confuses the issue of bias either with the fact that real differences between groups are extant, or with concerns about curricular validity of the instrument, equal opportunity to learn the subject materials, violations of standardizations of testing conditions, and the like.

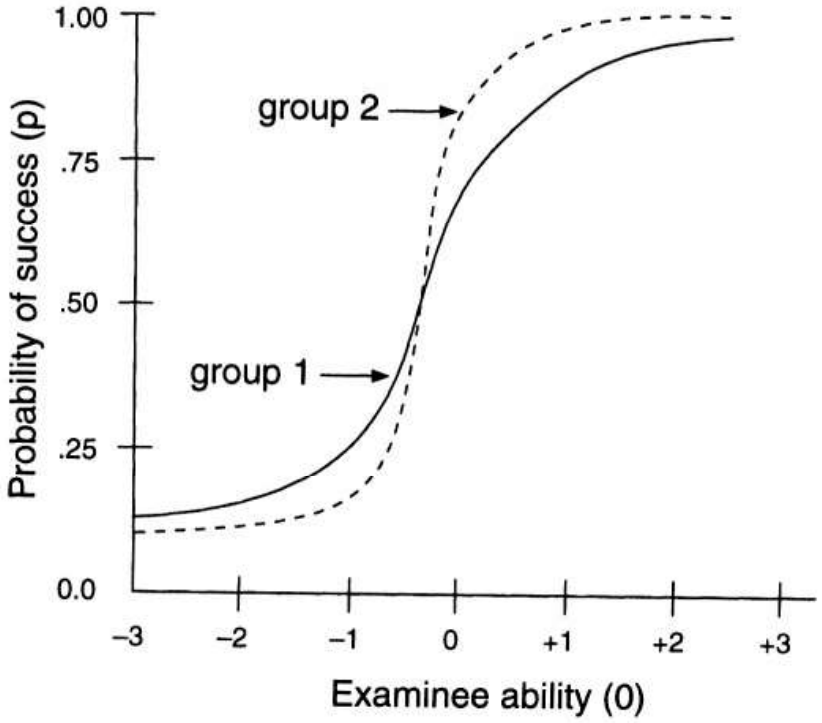
The techniques of bias detection have evolved considerably in a short period of time. But the most significant advancement of bias-detection strategies accompanied the rising interest in IRT in the late 1970s and into the 1980s. Today, techniques involving item response theory are generally considered the most robust, or technically meritorious, approach to detecting items that exhibit a systematic distortion (Lord, 1980).

Unfortunately, these procedures involve exceedingly complex statistics, require very sophisticated computer programs which must perform vast numbers of calculations, and are very difficult to implement because the mathematical algorithms need enormous sample sizes from each subpopulation to produce stable item-parameter estimates. This final condition means that for bias detection work, IRT can be used in only a few very large-scale testing programs because when the variable to be investigated is ethnic heritage (the usual case), it is rare to have a population with sufficient numbers of examinees in each of the subpopulations. Nevertheless, item-bias detection using these techniques is important, even if only conceptually available to most item writers. Therefore, the technique will be described, if only briefly. Osterlind (1983) offers a more complete

discussion of this technique, as well as surveying a variety of other bias-detection strategies.

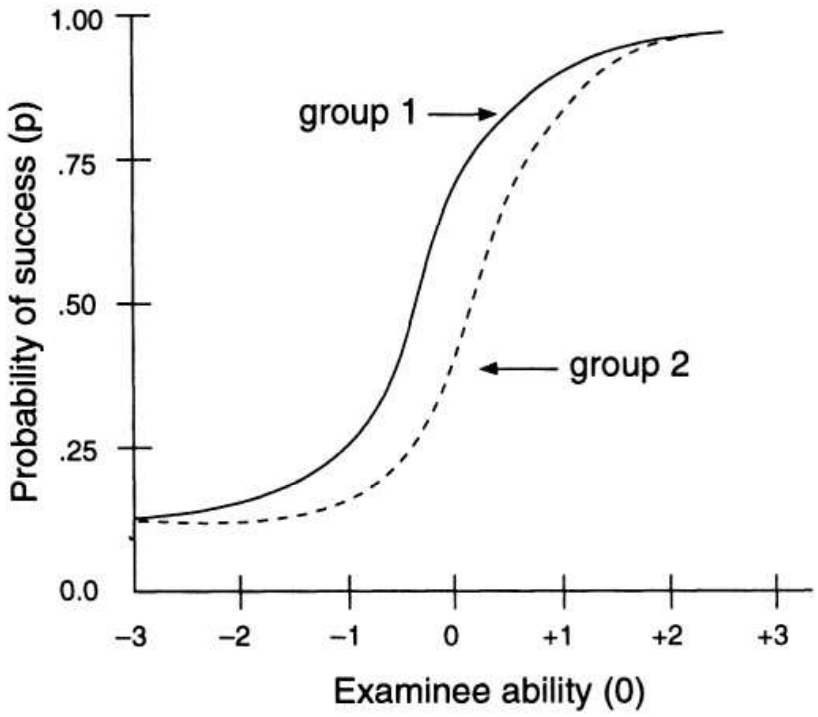
The technique for IRT item-bias detection is to compare the differences in the ICCs for groups. The area between the equated ICCs is an indication of the degree of bias present in an item. In other words, for a particular item, an ICC is computed for each group. The two ICCs are placed on the same scale by a simple linear transformation, and then compared. This method can be easily presented graphically, as in Figures 8.9, 8.10, and 8.11.

Figure 8.9. Hypothetical equated item characteristic curves for two groups different in discrimination.



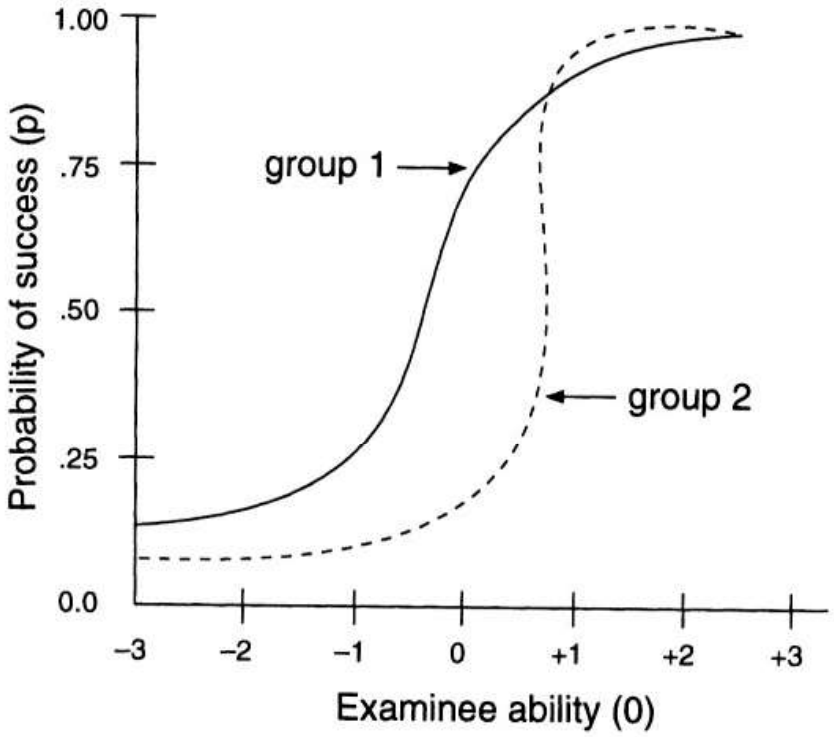
Notice in Figures 8.9, 8.10, and 8.11 that in each instance only a single item is considered. There are two ICCs, one from each subpopulation. Each figure displays an item that operates differentially between groups for different reasons. In 8.9, the differing slope for the item reveals differential performance in discrimination for each group; in 8.10, the left-to-right shift shows differences in difficulty for each group; and, in 8.11, differences are displayed in all three parameters for each group.

Figure 8.10. Hypothetical equated item characteristic curves



This information is very useful to item writers. Not only is it known that an item performs differently for various groups, but the nature of the differential performance is revealed. With such knowledge, writers can eliminate items or improve them, knowing where they need to focus their efforts—whether on making the item easier or more difficult, or on trying to produce an item that is more highly discriminating among ability levels, or on attempting to reduce the effects of guessing for very-low-ability examinees.

Figure 8.11. Hypothetical equated item characteristic curves for two groups different in discrimination, difficulty, and pseudochance.



Simple but Incomplete Bias-Detection Strategies

One very simple, albeit incomplete, technique for detecting bias in items compares the rankings of item difficulty values between two groups. If the items for the two groups do not rank similarly, a differential performance may be inferred. It should be understood, however, that comparing rank order of item-difficulty indices between groups is an incomplete strategy for concluding bias exists in test items. It is, nevertheless, a useful technique as an early indication of whether particular items behave differently between groups. And, it is one that item writers can use for preliminary examination of particular items.

To set up the procedure of relative rankings for items, the *p*-value rankings for each of two or more groups are computed. These values are placed side-by-side to facilitate comparisons (Table 8.13).

Table 8.13 Rank Order of Item Difficulty for a Hypothetical Test

Item	Rank Order for Item Group I (*p*-value)	Rank Order for Item Group II (*p*-value)
1	3rd (p = .62)	2nd (p = .64)
2	1st(p = .93)	1st(p = .81)
3	4th (p = .55)	3rd(p = .51)
4	2nd(p=.71)	5th(p=.19)
5	5th (p = .28)	4th (p = .38)

p=.40.

Suppose the data for a five-item test are distributed as in 8.13. Notice in the table that item 2 is the easiest for both groups, regardless of the fact that a substantially higher percentage of examinees in Group I responded correctly than did examinees in Group II (i.e., PI= .93 versus PII= .81). The pattern for responses, however, is not continued for item 4. This item is only second in difficulty ranking for Group I, but it is the most difficult of all the items from Group II. Thus, the suspicion is raised that Item 4 does not behave similarly between the two groups. Bias may be present in the item to a degree that consistently underestimates the performance of Group II. All other items appear to rank in a pattern similar for both groups,

so no other item is suspected of aberrance.

A rank-order correlation coefficient, typically Spearman's rho, between the two sets of values can be computed for further confirmation of aberrance. For correlations of this kind, one would look for a correlational estimate of .8 or higher to judge the similarity in rankings of item-difficulty values between the two groups. In the example, the coefficient of correlation is .40. The comparatively low level of correlation supports the suspicion of bias by this method. (Of course, this data is computed only for illustrative purposes, and is distorted by the very few items considered.)

Mantel and Haenszel (1959) suggested a procedure with more technical merit that involves applying the chi-square statistic to matched groups. Although their work appeared early in the literature of bias-detection approaches, its value was not fully recognized until recently. This may be due to the fact that although the Mantel-Haenszel procedure is quite simple, it provides a powerful approximation of the IRT methods described above.

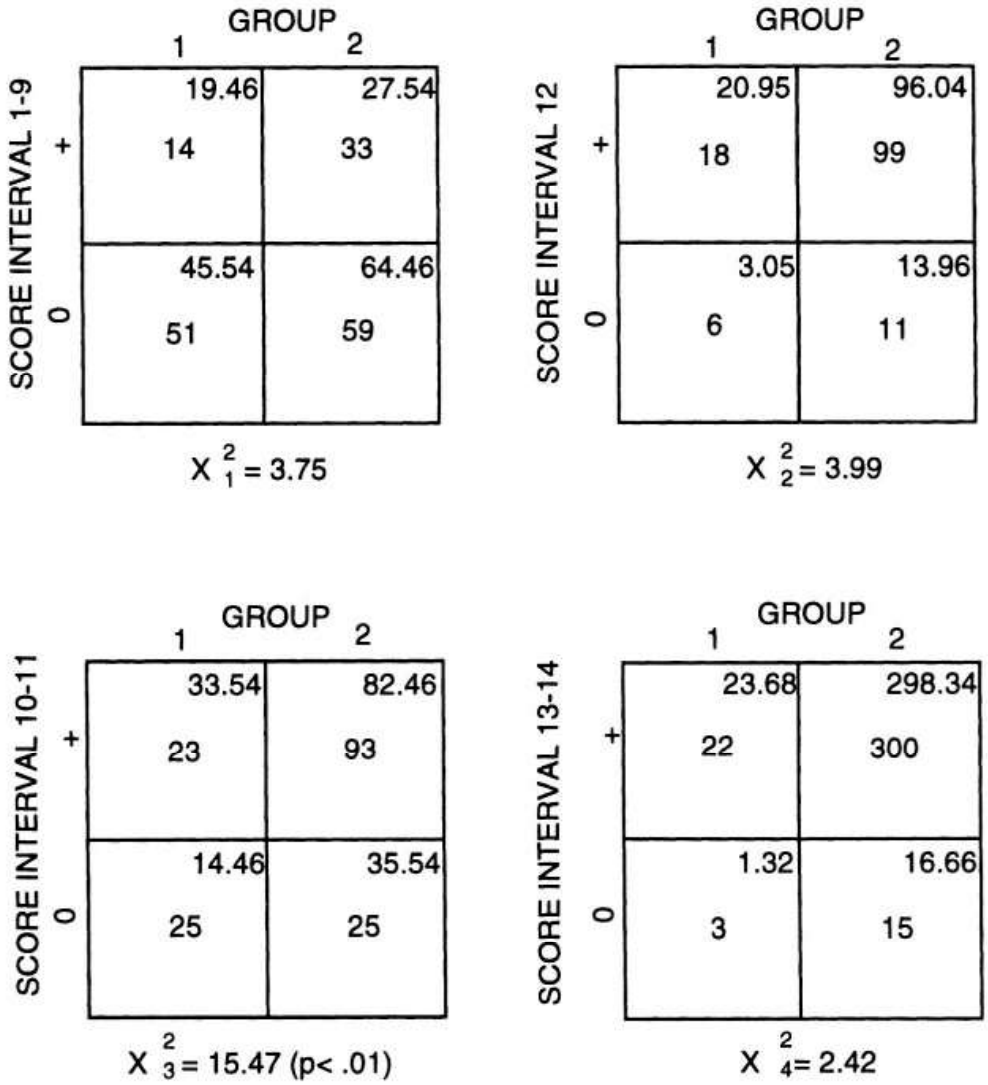
Another advantage of the Mantel-Haenszel procedure (cf. Holland and Thayer, 1986) is that it involves the computation of only a chi-square statistic and is, therefore, not limited like IRT to use only with very large groups of examinees. The chi-square approach proposed by Marascuillo and Slaughter (1981), which is also based on a chi-square statistic and is very similar to the Mantel-Haenszel, offers this same advantage.

Procedurally, one establishes strata of ability groups based on the total test score by considering natural breaking points in the total population's distribution of scores. Typically, three or four ability strata are established. The number of persons from each subpopulation to be considered who passed and failed the item is then determined. These frequencies are then set in a series of 2 X 2 contingency tables. Table 8.14 displays data for a hypothetical distribution of scores that have been broken into four ability strata.

The chi-square statistic is then computed and tested for significance. If a value significantly above chance is attained, differential item performance is inferred. The item should then be discarded or reworked.

Thus, examining items for bias is important to the item writer. It provides information that can be helpful in identifying poor items, and it may provide clues as to how a particular item can be improved. The skilled item writer will use these bias-detection strategies to advantage when conducting item analysis.

Table 8.14 Contingency Tables for Two Groups at Four Total Score Intervals on One Test Item



\+ = success on item
0 = failure on item

Judgmental Approaches to Bias Detection

In addition to the mathematical definition for bias described in the preceding section, the writer should be sensitive in his use of language to gender, cultural, ethnic-heritage, and other differences. Language that may

offend persons of either gender or persons from any ethnic-heritage or religious group may not manifest numerical differences and could go undetected by statistical approaches to detecting bias in items. Further, changing roles for persons force changes in language. For example, at one time it was nearly universal to refer to medical doctors with the masculine pronoun *he* and nurses with the feminine *she.* Such distinctions are no longer applicable. Or, persons with physical disabilities were once called *handicapped.* Today, more enlightened attitudes generally prevail, and terms such as *physically challenged* are preferred. Chapter 5 cites sources to which one can turn for more information on reducing offensive and stereotyped language.

It is important that item writers gather differing opinions representative of the two genders, of persons from differing ethnic heritages, and of other groups to address the issue of sensitivity in language. Such review process does not require a complex methodology. In most cases, for the purpose of improving items, it is adequate to have persons representing the relevant viewpoints review items informally, preferably in honest, open discussions. This does not mean such a review should be haphazard. The reviewers will need criteria. The criteria might be a list of guiding questions, such as the following: "Does the language offend either gender, or persons of various ethnic heritages, or others?"; "Does the language stereotype either gender or persons of various ethnic heritages, or others?"; "Does the language set a tone that reflects out-of-date attitudes for either gender or persons of various ethnic heritages, or others?."

This kind of review will likely yield the writer, who is trying to improve the language of the items more useful information if it is gathered through discussion, rather than asking reviewers to complete a form. When such a discussion is conducted in the spirit of improving language, the item writer will usually find these reviewers invaluable in reducing this kind of bias in test items.

CONCLUSION

It was mentioned at the outset of this chapter that constructing a test item is not complete until the item has been thoroughly scrutinized for its quality. This chapter provides the writer with the tools necessary for such scrutiny. As can be seen, these tools include the techniques of conducting a validation study relevant to the preparation of the items, as well as methods

for defining items in mathematical terms so that they may be examined by appropriate statistics.

When these tools are understood, the item writer will be prepared to gauge his or her work; when the tools are properly used, the writer will have taken a very important step in identifying good test items and poor ones. When the good items are recognized and the poor items are removed or improved, the writer will have taken the final step in the long and difficult journey of constructing good test items.

The next chapter is the concluding one. It focuses less on specific constructing test items than on discussing some overarching considerations, especially ethical and legal concerns for item writers.