Chapter 14

ITEM ANALYSIS

In test construction, a general goal is to arrive at a test of minimum length that will yield scores with the necessary degree of reliability and validity for the intended uses. This is typically accomplished by field-testing a large pool of items and selecting a subset of items from that pool that make the greatest contributions to reliability or validity. In constructing a new test (or shortening an existing one), the final set of items is usually identified through a process known as *item analysis*. Item analysis is a term broadly used to define the computation and examination of any statistical property of examinees' responses to an individual test item. Item parameters commonly examined fall into three general categories:

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
3. Indices that are a function of both item variance and relationship to a criterion

In the following sections definitional formulas for some of the most popular indices of each category will be presented. Later a data example will illustrate how information from such item parameters is considered in item selection or revision decisions for norm-referenced tests. Item analysis methods for criterion-referenced tests are considered in the final sections of the chapter.

ITEM DIFFICULTY, MEAN, AND VARIANCE

When an item is dichotomously scored, the mean item score corresponds to the proportion of examinees who answer the item correctly. This proportion for item i is usually denoted as p_i and is called the *item difficulty*. This parameter was dis-

cussed in some detail in Chapter 5. Recall that the value of p_i may range from .00 to 1.00. As we shall see later, nearly all total test score parameters are affected by item difficulty. The total test score mean is directly related to the item difficulties because

$$\mu_X = \sum_i p_i \tag{14.1}$$

Furthermore, if we are interested in the difficulty level of an average item on the test, it can be obtained by

$$\mu_p = \frac{(\mu_X)}{k} \tag{14.2}$$

where k is the number of items on the test. When describing an examinee group's performance on several tests of different length, the average item difficulty may be preferred to the raw-score means (which vary as a function of test score length).

As shown in Chapter 5, item difficulty level controls item variance because

$$\sigma_i^2 = p_i q_i \tag{14.3}$$

There it was also established that assuming a constant degree of correlation among items, total test score variance will be maximized when $p_i = .50$. Thus it might seem surprising to learn that for most published aptitude and achievement tests designed for norm-referenced score interpretation, item difficulties typically fall in the range of .60 to .80. The reason for this lies in the item format commonly used in such tests.

To understand how item format may affect p-value, consider a format such as

$28 \times 7 =$ _____

Obviously there is little chance that an examinee who does not know the answer can supply it by guessing. Thus the observed p-value for this item and consequently its variance are primarily functions of examinees' knowledge (i.e., their true scores on this item). Now consider the same item in this format:

$28 \times 7 =$ _____
a. 186
b. 196
c. 287
d. 554

Because this item format allows some examinees to mark the correct response by guessing, the observed p-value is affected by both the examinees' true scores on the item and by guessing. Under the random guessing assumption, the observed p-value is expected to be the sum of the proportion of examinees who *know* the answer (the "true" p-value) and $1/m$ of the proportion who do not know the answer, when m is the number of choices. To maximize the total true score variance for the test it is

TABLE 14.1 Response Distributions and p_o Values for Objective Items with Different Numbers of Possible Responses

Number of Choices	Proportion Who Know Answer	Proportion Who Guess Answer	p_o	Lord's p_o
4-choice item	.50	.50/4	.50 + (.50/4) = .62	.74
3-choice item	.50	.50/3	.50 + (.50/3) = .67	.77
2-choice item	.50	.50/2	.50 + (.50/2) = .75	.85

necessary to maximize the item true score variances. Item true score variance will be maximized when half the examinees can answer the item based on knowledge and half cannot; however, among the half who do not know the correct answer, each examinee has a $1/m$ probability of answering the item correctly by guessing at random among the m choices. Thus the proportion of examinees who may be expected to mark the correct answer by random guessing is $.50/m$. Thus the observed difficulty of an item with maximum true score variance would be expected to be

$$p_o = .50 + .50/m$$

Table 14.1 demonstrates how p_o values are computed for items with different numbers of choices. Values calculated in this table are still probably somewhat lower than the optimal p-value for a norm-referenced test. In a simulation study Lord (1952) demonstrated that reliability is improved by choosing items with p-values even higher than those computed by adjusting for random guessing. Lord's recommended p-values are also reported in Table 14.1. From a practical standpoint, even though Lord's values are based on simulated response data, they are probably still more reasonable than those computed from the random guessing model because many examinees have some partial knowledge that enables them to eliminate one or more foils before taking a guess, thus making their probabilities of a correct response greater than $1/m$.

ITEM DISCRIMINATION

The purpose of many tests is to provide information about individual differences either on the construct purportedly measured by the test or on some external criterion which the test scores are supposed to predict. In either case the parameter of interest in selection of items must be an index of how effectively the item discriminates between examinees who are relatively high on the criterion of interest and those who are relatively low. At times there is no more adequate measure of that construct available than the total test score itself. (The classroom achievement test is a primary example of this situation.) In this circumstance, total score on all the items is used as an operational definition of the examinee's relative standing on the construct of interest. With this internal criterion, the goal is to identify items for which high-scoring examinees have a high probability of answering correctly and

low-scoring examinees have a low probability of answering correctly. In achievement testing, for example, we would say that such an item discriminates, or differentiates, between examinees who know the material and those who do not. In contrast, we would be suspicious of an item on which both high and low scorers were equally successful. Such an item would not seem to measure the same construct tested by the other items. It would be even less desirable to have items that are missed by many high-scoring examinees but are answered correctly by low-scoring examinees. Such items are said to show *negative discrimination*.

Five parameters used as indicators of the item's discrimination effectiveness will be described in the following section. One of these is based on the concept of differentiating between groups of examinees defined by imposing cut scores at specific points on the criterion score distribution. The remaining four are various types of correlation coefficients. It is important to recognize that each of these parameters is equally applicable to the situation in which the item score is being related to total test performance or to the situation in which the item score is being related to performance on some external criterion variable. Although the latter practice is somewhat controversial for reasons to be discussed later, the use of an internal criterion (i.e., total test score) or an external criterion does not alter the defining formulas or computation procedures.

Index of Discrimination

One simple discrimination parameter, called the *index of discrimination*, can only be applied to dichotomously scored items. Its computation requires designating one or two points on the criterion score distribution as cut scores and separating the examinees into groups who scored below and above these cut scores. For example, if the test developer is interested in selecting items that discriminate on the internal criterion of total test score, the groups could be composed of the upper 50% and the lower 50% of the examinee group, based on total test score. If the examinee group is large, it may not be necessary to use the entire upper and lower 50%. A classic study by Kelley (1939) demonstrated that under certain conditions, a more sensitive and stable item discrimination index can be obtained by using the upper 27% and the lower 27% of the examinee group; however, when sample size is reasonably large, virtually the same results can be obtained with the upper and lower 30% or 50% (Beuchert and Mendoza, 1979; Englehart, 1965).

Once the upper and lower groups have been identified, the index of discrimination (D) is computed as

$$D = p_u - p_l \tag{14.4}$$

where p_u is the proportion in the upper group who answered the item correctly and p_l is the proportion in the lower group who answered the item correctly. Values of D may range from -1.00 to 1.00. Positive values indicate that the item discriminates in favor of the upper group; negative values indicate that the item is a reverse discriminator, favoring the lower-scoring group.

TABLE 14.2. Illustrative Response Patterns of Ten Examinees to Three Selected Items

	Examinee									
	1+	2+	3−	4+	5	6	7	8−	9	10−
Item 1	0	1	1	1	0	1	1	0	1	0
Item 2	1	1	1	1	0	0	0	1	0	1
Item 3	1	1	1	1	0	1	1	1	1	1
Total score on 30 items	25	27	15	24	20	18	16	14	22	10

Table 14.2 displays responses of 10 examinees to selected items from a 30-item test. Normally we would require a larger sample for item analysis, and use this example only for purposes of illustration. Examinees in the upper 30% are identified with a "+" and those in the lower 30% with a "−" beside their identification numbers. Responses of these three high-scoring and three low-scoring examinees have been used to calculate D for items 1, 2, and 3 in Table 14.3. Note that although both items 1 and 2 have the same p-values (.60), their D-values differ considerably. Based on practical experience Ebel (1965) offered the following guidelines for interpretation of D-values when the groups are established with total test score as the criterion:

1. If $D \geq .40$, the item is functioning quite satisfactorily.
2. If $.30 \leq D \leq .39$, little or no revision is required.
3. If $.20 \leq D \leq .29$, the item is marginal and needs revision.
4. If $D \leq .19$, the item should be eliminated or completely revised.

One drawback of using D is that it has no well-known sampling distribution. It is not possible to answer questions such as what D-value is significantly greater than zero, or how large a difference between D-values is statistically significant? Nevertheless, because of its ease of computation and straightforward interpretation, the index of discrimination continues to be one of the most popular methods of reporting item discrimination effectiveness. It is especially appropriate for classroom test item analysis performed with hand computations.

Correlational Indices of Item Discrimination

In Chapter 2 we discussed the Pearson product moment correlation coefficient as a measure of the degree of linear relationship between two variables. If the test items undergoing development have a possible score range of 1 to 4, 1 to 5, or greater (such as items from a Likert attitude inventory), this formula is commonly used to estimate the degree of relationship between item and criterion scores. Although the

TABLE 14.3. Illustrative Computations of Item Discrimination Indices for Items from Table 14.2, where $\mu = 19.10$ and $\sigma_X = 5.17$

Item	p	μ_+	Index of Discrimination $D = (p_u - p_l)$	Point Biserial Correlation $\rho_{pbis} = \frac{(\mu_+ - \mu_t)}{\sigma_X}\sqrt{pq}$	Biserial Correlation $r_{bis} = \frac{(\mu_+ - \mu_t)}{\sigma_X}(p/Y)$
1	.60	20.33	$(.67 - .33) = .34$	$\frac{(20.33 - 19.10)}{5.17}(1.22) = .29$	$\frac{(20.33 - 19.10)}{5.17}(.60/.3867) = .37$
2	.60	19.17	$(1.00 - 1.00) = .00$	$\frac{(19.17 - 19.10)}{5.17}(1.22) = .016$	$\frac{(19.17 - 19.10)}{5.17}(.60/.3867) = .021$
3	.90	19.00	$(1.00 - 1.00) = .00$	$\frac{(19.00 - 19.10)}{5.17}(3) = -.06$	$\frac{(19.00 - 19.10)}{5.17}(.90/.1714) = -.10$

same formula for the product moment correlation can also be used with dichotomously scored variables, special formulas have been developed that are easier to use in hand calculations when one or both variables are dichotomously scored. Four such correlational indices will be described in the following sections.

POINT BISERIAL CORRELATION. One situation which occurs frequently in item analysis is when the test developer is interested in how closely performance on a test item scored 0 to 1 is related to performance on the total test score (or some other continuously distributed criterion). A simplified computational formula for the Pearson product moment coefficient in this situation is called the *point biserial correlation*, denoted as

$$\rho_{pbis} = \frac{(\mu_+ - \mu_X)}{\sigma_X}\sqrt{p/q} \tag{14.5}$$

where μ_+ is the mean criterion score for those who answer the item correctly, μ_X is the mean criterion score for the entire group and σ_X is their standard deviation, p is item difficulty, and q is $(1 - p)$. Table 14.3 demonstrates the calculation of the point biserial correlations between total score and the item scores presented in Table 14.2.

The value of the point biserial correlation between an item score and total score is somewhat spurious because the item score has contributed to the total score of each examinee. If the number of items is reasonably large (perhaps 25 or more), this fact is seldom a problem. However, with a small number of items, this problem may be corrected by

$$\rho_{iX-i} = \frac{\rho_{iX}\sigma_X - \sigma_i}{\sqrt{\sigma_i^2 + \sigma_X^2 - 2\rho_{iX}\sigma_X\sigma_i}} \tag{14.6}$$

where ρ_{iX-i} is the correlation between an item score and the total score with that item removed, and σ_X and σ_i are the total and item standard deviations respectively.

BISERIAL CORRELATION COEFFICIENT. If we wish to assume that the latent variable underlying item performance is normally distributed, it is possible to derive a formula for the correlation between this variable and a continuously distributed criterion such as a test score. This statistic, first derived by Pearson (1909), is called the *biserial correlation coefficient* and may be computed by the formula

$$\rho_{bis} = \frac{(\mu_+ - \mu_X)}{\sigma_X}(p/Y) \tag{14.7}$$

where μ_+ is the criterion score mean of those who answered the item correctly; μ_X is the criterion score mean of all examinees and σ_X is their standard deviation; p is the proportion of examinees who answered the item correctly; and Y is the Y ordinate of the standard normal curve at the z-score associated with the p value for this item. For item 1 in Table 14.2, $p_1 = .60$. Because p_1 is greater than .50, turning to the standard normal curve table in Appendix A, we examine the column of probabil-

ities (or areas to the left) of positive z scores. The area value closest to .60 is .599, and the ordinate associated with this point on the normal curve is .3867. From data in Table 14.2, we can also compute the values of μ_+, μ_X, and σ_X. Using these values in Equation 14.7, we obtain $\rho_{bis} = .36$. The calculations of biserial correlation values for items 1, 2, and 3 are demonstrated in Table 14.3. Notice that these values are somewhat different than those obtained from the point biserial correlation formula.

Mathematically, the relationship between the biserial and point biserial correlations is

$$\rho_{bis} = \frac{\sqrt{pq}}{y}\rho_{pbis} \tag{14.8}$$

Because the value of the Y ordinate on a normal curve is always less than $\sqrt{pq}$, the value of a biserial correlation will always be at least one-fifth greater than the point biserial correlation for the same variables (Lord and Novick, 1968). This fractional difference in magnitude remains fairly moderate for items of medium difficulty; however as p-values drop below .25 or increase above .75, the difference between biserial and point biserial correlation increases sharply. Magnusson (1967) graphically demonstrated that in extreme difficulty ranges, the biserial correlation may be four times greater than the point biserial correlation between item score and total score. Notice in Table 14.3 that the ratio of the biserial to point biserial correlation is greater for item 3 (1.67) than for item 2 (1.23) or item 1 (1.28). This occurs because low item variance operates to restrict the value of the point biserial correlation. Thus test users who are comparing reports of item analyses from different studies where different correlational formulas were used should remember that biserial correlations may be systematically higher than point biserial correlations, and therefore, apparent differences in the magnitudes of the item discrimination parameters may be due to the choice of correlational formula rather than to qualitative differences in the items.

PHI COEFFICIENT. When scores from a dichotomously scored item are to be correlated with scores from a dichotomous criterion (e.g., success or failure in a rehabilitative program or a demographic characteristic such as gender), the phi coefficient described in Chapter 5 may be used. Another use of such a coefficient is to determine the degree of stability in responses to the same dichotomous item by the same examinees on different occasions. Use of phi is most appropriate when the variables involved are true dichotomies. When criterion groups are formed by imposing an artificial cutoff point on a continuous distribution, this statistic does not permit full use of the information available. In other words if all examinees whose scores fall above a cutoff score are simply classified as 1 for pass and 0 for fail, quantitative information about the differences between the scores in the passing and failing groups will be lost. Another possible limitation of phi is that its value can only be 1.00 when the p-values for the two variables are equal (because the phi coefficient, like the point biserial, is derived directly from the Pearson product moment correlation).

TETRACHORIC CORRELATION COEFFICIENT. At times the test developer may be interested in how strongly two dichotomous variables are correlated when each variable is created through dichotomizing an underlying normal distribution. In such instances the tetrachoric correlation coefficient may be appropriate. Computation of this statistic is complicated and is seldom undertaken unless the test developer strongly believes that another correlational index, such as phi, is inadequate for the purpose. One such situation is when an intercorrelation matrix of dichotomously scored items will be submitted to a factor analysis. Because values of the phi coefficient are restricted, except when both p-values are equal, such correlation coefficients are less appropriate for factor analysis than tetrachoric coefficients. The formula for the tetrachoric coefficient is not easily presented, but computer programs for estimating this coefficient are available in several standard computer packages (see, for example, Dixon et al., 1981).

Comparison of Item Discrimination Indices

Thus far five different methods for investigating item discrimination have been presented. Obviously in some situations, because of the scoring of the variables, one technique may be more appropriate than others. At other times it might be feasible to use more than one of the methods described. Thus it seems reasonable to question the similarity of results obtained from the various methods. Several empirical studies have addressed this issue. Englehart (1965) used responses from 210 examinees taking a state high school equivalency examination. Using total score as criterion, he computed a variety of item discrimination indices, including D, phi, biserial ρ, point biserial ρ, and the tetrachoric ρ. When the values obtained for each of these item indices were correlated, the correlations between pairs of discrimination statistics ranged from .85 to .99 on one form of the test and from .90 to .99 on a parallel form. Similar studies have been reported by Beuchert and Mendoza (1979), Findley (1956), and Oosterhof (1976). In most of these studies the greatest discrepancies occurred for items at extreme difficulty ranges.

In summary, the following recommendations can be offered concerning the choice of an item discrimination procedure for dichotomously scored items:

1. When items are of moderate difficulty, it makes little difference which discrimination statistic is used. If ease of computation is a major concern, D is recommended. If a test of statistical significance is desired, one of the correlational procedures should be used.
2. If the goal is to select items at one extreme of the difficulty range, biserial ρ is recommended (if it is reasonable to assume a normal distribution of the trait underlying item performance).
3. If the test developer suspects that future samples will differ in ability from the present item analysis group, biserial ρ is again recommended since the relative order of item discrimination for this statistic should remain more stable from sample to sample when samples vary in ability. Another way of saying this is that a low biserial ρ value for a sample of any ability level indicates that the item is low in discriminating power; but a low point biserial value for a sample of low ability (or

high ability) may simply be a function of the item difficulty and does not necessarily indicate that the item is a poor discriminator.

4. If the test developer is fairly confident that future samples will be similar in ability to the item analysis sample, and the goal is to select items that will have high internal consistency, it may be preferable to use the point biserial correlation (Lord and Novick, 1968). Although this has not been conclusively demonstrated, it seems logical in the view of the fact that point biserial values would be higher for items of medium difficulty, and such items would allow maximum item covariances and hence a maximum value of alpha.

5. In cases where both the item and criterion variable are scored dichotomously, the phi or tetrachoric coefficient should be used. Phi is easier to compute but will be artificially restricted when the proportions in the two dichotomies are not equal. The tetrachoric correlation is based on the assumption that the item and criterion scores arise by dichotomizing two normally distributed variables. Use of this coefficient is seldom warranted as a measure of item discrimination because of its computational complexity, but its use is strongly recommended if the correlations are to be used in a subsequent factor analysis.

ITEM RELIABILITY AND VALIDITY INDICES

The third class of item parameters often examined during an item analysis study may be typified by the *item reliability index* and the *item validity index*. Each of these is jointly a function of item score variability and item score correlation with a criterion. If the internal criterion of total test score is used, the index is defined as $\sigma_i \rho_{ix}$, where ρ_{ix} is the correlation between item and total test score. This index is called the *item reliability index*. For dichotomously scored items this formula is more commonly written as $\sqrt{p_i q_i} \rho_{ix}$, where ρ_{ix} is the point biserial correlation between item and total test score. If an external performance criterion is used, the index is defined as $\sigma_i \rho_{iy}$, where ρ_{iy} is the correlation between item score and the external criterion. When the goal of item selection is to improve test score reliability or validity by selection of items which discriminate on the criterion of interest, it has sometimes been suggested that the item reliability index (or the item validity index) should be used in lieu of the simple correlation between item and criterion because the item variance actually weights the relative contribution of a particular item to overall test score reliability or validity. If, for example, two items have equal correlations with total test score, but one item has greater variance than the other, the item with greater variance makes a greater contribution to test score reliability. Although this argument has some merit, it should be noted that as long as items with medium difficulties are chosen, there is little practical advantage in using the item reliability index instead of the item total score correlation.

There are, however, some test construction situations in which the item reliability index or the item validity index may be useful. It has been shown that total test score variance can be expressed as the sum of the item reliability indices, so that

$$\sigma_X^2 = (\Sigma \sigma_i \rho_{ix})^2 \tag{14.9}$$

This may be useful in an item analysis where the test developer has set a desired minimal value for the total test score variance. As the process of item selection begins, the sum of item reliability indices is incremented with the selection of each additional item until the desired minimum level for total score variance is achieved. This is much simpler computationally than rescoring the test for each examinee and recomputing the variance from examinees' total raw scores with the addition of each item.

Similarly, if the test developer has set a minimum value for the internal consistency coefficient, as measured by coefficient alpha, the value of alpha can be reestimated with the addition of each new item, using only item-level data, by the formula

$$\rho_\alpha = \frac{k}{k-1}\left[1 - \frac{\Sigma \sigma_i^2}{(\Sigma \sigma_i \rho_{ix})^2}\right] \tag{14.10}$$

Where k represents the number of items selected to this point.

Finally, if the test developer wishes to specify the minimum value for a validity coefficient between a selected subset of items and some external criterion, as each new item is added, the validity coefficient can be estimated from item-level data by

$$\rho_{Xt} = \frac{\Sigma \sigma_i \rho_{iy}}{\Sigma \sigma_i \rho_{ix}} \tag{14.11}$$

It should be noted that the values of variance, coefficient alpha, and validity coefficients estimated from item reliability indices are approximations when the values of ρ_{ix} are obtained from an item analysis of all items in the field-tested item pool rather than just for those items in the selected subset of items. Derivations for these formulas showing the relationships between item parameters and total test score parameters are given by Gulliksen (1950) and Lord and Novick (1968).

CONDUCTING AN ITEM ANALYSIS STUDY

Field-Testing and Item Analysis: Overview

By now it is apparent that total test scores can have no properties that are not a function of the items that comprise the test. Thus if the test developer wants test scores that have minimal measurement error or that have strong relationships to performance criteria or measures of other constructs, it is not enough to draft a set of items and "hope for the best." Once all the items have been written and revised from results of formal item review and preliminary tryouts, it is standard practice to field-test the items on an appropriate examinee sample. Results of this field test are used in the item analysis. Through item analysis the test developer identifies those items that are functioning as intended and those items that are not. In most cases the former items will be retained and the latter will be revised or eliminated from future versions of the test.

In a typical item analysis the test developer will

1. Decide what properties of the test score are of greatest importance
2. Identify the item parameters most relevant to those properties
3. Administer the items to a sample of examinees representative of those for whom the test is intended
4. Estimate for each item the parameters identified in step 2
5. Establish a plan for selection of items (or identification and revision of malfunctioning items)
6. Select the final subset of items
7. Assess whether the desired results have been achieved by conducting a cross-validation study

In the preceding sections of this chapter we identified and defined item parameters which are usually of interest because of their relationships to important total test score parameters. Formulas were presented for estimation of these item parameters. In the following sections, we focus on steps 3, 5, 6, and 7 of the previous sequel.

Sample Size

There is no absolute rule for the minimum number of examinees to use in an item analysis study. Certainly the item analysis for a test which will be widely used, such as the Graduate Record Examination or a commercially published aptitude or achievement test, should be based on a sizable, representative sample, perhaps of thousands of examinees. In contrast, a doctoral student who develops an instrument for dissertation research may have to rely on a much smaller sample. As a general rule, most item parameters described in this chapter can be estimated with relative stability for samples of 200 examinees, and so this might be considered the minimum number desired. Another longstanding rule-of-thumb (Nunnally, 1967) is to have 5 to 10 times as many subjects as items. If the test developer relies on this latter guideline, at the minimum, 20 items and 100 subjects probably should be used. Sample sizes required for parameter estimation based on item response theory (described in Chapter 15) may vary from 200 to 1,000 subjects, depending on the particular model chosen. Thus it is vital for the test developer to know what item analysis procedures will be applied when planning the item field test so that a sufficient number of examinees can be tested. Another related consideration is the need for cross validation, which will be discussed shortly. This final phase of the item analysis study requires testing subjects in addition to those whose responses were used to estimate item parameters.

Establishing a Plan for Item Selection

Historically, a controversial issue in test development has been whether it is more important to select items that correlate with total test score (resulting in test scores with higher internal consistency) or to select items on the strength of their relationship to an external criterion. On the one hand, if a polyglot of items is chosen which correlate highly with an external criterion but have little relationship with each other, the meaningfulness and interpretability of the test scores as measures of a construct will be questionable (see Travers, 1951). On the other hand, the more that items correlate with each other, the less additional variance in the criterion is likely to be accounted for by the collection of items. A key point for the test developer to remember in designating the criterion to be used for item discrimination is that the appropriateness of the criterion should be dictated by the intended purpose of the test and the likely audience it is to serve. The more it is true that the sole purpose of the test is to predict performance on a single well-defined criterion in a specific setting, the more reasonable it may be to select items which correlate with that criterion. However, such singleness of purpose in test usage is relatively rare. More commonly a test is developed to represent a domain of behavior or to predict a future performance which cannot be adequately defined by a single criterion variable. In such cases, selection of items which maximize the predictive power of the test for one particular criterion in one local setting, without requiring the items to have evidence of relationship to a broader construct of interest, will likely result in a test which will have limited usefulness in other settings, where the criteria or examinee populations differ even slightly. Thus many test developers write items to assess a trait of interest; select items from item analysis based on total score on the test; and later assess how test scores, as measures of the construct, are related to one or more external criterion variables.

In selecting field-tested items to retain for the test, the test developer usually encounters one of the following situations. In the first there are many more items than can be reasonably administered during routine test use, and to save on testing and scoring time, test users will want to administer as few items as necessary. Therefore the task is to select a subset of items which make the greatest contributions to the desired level of test score reliability or validity or with as few items as possible. The usefulness of the item reliability index (or the item validity index) for this purpose has already been discussed. In the second situation the item pool is not overly large, and the test developer wants to retain every item which makes a positive contribution to the test score parameter of major interest. In this case, a minimum level must be established for the item discrimination parameter, and all items with a discrimination parameter greater than that minimum may be retained. Ebel's (1965b) suggested criterion may be employed for assessing the index of discrimination. For correlational indices the test developer often elects to keep every item that has an item criterion correlation significantly greater than .00.

For phi and point biserial correlations, a convenient approximation for the standard error for the Pearson product moment correlation can be used to establish this level by computing

$$\hat{\sigma}_\rho = \frac{1}{\sqrt{N-1}} \qquad (14.12)$$

where N is the sample size. (In using this formula we are assuming a sample size of

at least 50.) Usually the minimum critical value is set at 2 standard errors above .00. Thus for a sample of 101,

$$\hat{\sigma}_r = \frac{1}{10} = .10$$

and .00 + 2(.10) = .20. Thus we would want to retain items with point biserial values of .20 or greater.

The standard error for the biserial correlation can be estimated by the formula

$$\hat{\sigma}_{bis} = \frac{\sqrt{pq/(N-1)}}{Y} \tag{14.13}$$

where p is the proportion answering the item correctly; $q = (1 - p)$; N is sample size, and Y is the ordinate of the normal curve at the value of p (Kurtz and Mayo, 1979). Again, this standard error may be used to generate the upper bound of an interval around .00, so that items with correlations greater than those expected by chance can be identified. The standard error of the biserial correlation will be smallest when $p = .50$ and will increase as p-values become extreme. Thus the stability of this statistic from sample to sample is strongly affected by item difficulty.

In developing an item selection strategy, the novice test developer may wonder how item difficulty data should influence the decision. Several points are relevant to this issue. First, item difficulty is rarely a primary criterion for item selection. For norm-referenced tests, it is generally less important than item discrimination. Second, p-values will vary from sample to sample. (The amount of sampling error in item difficulty can be estimated by using the standard error of a proportion, $\hat{\sigma}_p = \sqrt{pq/N}$, where N is the sample size.) Finally, for a test expected to discriminate reliably across a broad range of ability, items should be of uniform moderate difficulty. For years commercial test publishers constructed tests so that they consisted of a nearly even mixture of low-, medium-, and high-difficulty items. Gradually, however, this practice has been replaced by the strategy of selecting items from the medium difficulty range, with adjustments for guessing on multiple-choice items.

Two studies influential in this practice were reported by Lord (1952) and Cronbach and Warrington (1952). In general both studies showed that when items are moderately correlated with total test scores (as is usually the case), items of uniformly medium difficulty will permit more reliable discriminations among examinees of nearly all ability levels than will a collection of items with a wider spread of difficulties. More specifically, Henryssen (1971) suggested that when the average biserial correlation between item and total test score is in the range .30 to .40, the ideal item difficulty level should be between .40 and .60; but as the average biserial correlation increases above .60, a wider range of item difficulties may be acceptable. The one recognized exception to this strategy is when the test scores will be used exclusively for decision making for examinees at the upper or lower end of the distribution. For example, if a test is to be used to select applicants into a competitive professional training program and only a handful of the most qualified

applicants are to be chosen, it would be appropriate to select items with relatively low p-value (difficult items), as these are more likely to result in a test that discriminates among examinees in the ability range of interest. When a test is constructed for a select segment of the population, it is important for the test developer to state this intended usage clearly since the usefulness of the test for examinees at other ability levels will be lessened.

Using Item Analysis Data in Test Revision

The following case-study in item analysis is presented to illustrate (1) detection of flawed items and (2) use of examinee response data in diagnosing potential causes of item malfunction. Data are based on responses of 50 examinees to a 35-item classroom test in an undergraduate course where a norm-referenced grading policy is used. The instructor chooses the internal criterion of total test score in computing the item discrimination indices. Data for items 21 to 35 are shown in this example.

Table 14.4 exemplifies the output information from a typical computer program that performs item analysis. Two types of item discrimination statistics, item difficulty, and the proportion of examinees selecting each response are presented. For this situation, item discrimination statistics will be useful in identifying "problem" items, but difficulty level and response distribution pattern will be useful in diagnosing item construction flaws which may have resulted in the poor discrimination.

TABLE 14.4. Illustrative Item Analysis Results from 50 Examinees on Items 21 to 35 of a 35-Item Test

Item	Item Responses (%) 1	2	3	4	Omit	Diff. p	Index Disc.	Point Biserial Corr.
21	24	4	52	16+	4	.16	.00	−.06
22*	4	40	56+	0	0	.56	.67	.48
23*	0	76+	12	12	0	.76	.50	.45
24	4	28+	28	32*	8	.28	−.17	−.12
25	16	12	0	72+	0	.72	−.17	−.29
26	0	4	52	44+	0	.44	.00	−.11
27	92+	0	8	0	0	.92	.33	.45
28*	8	68+	0	20	4	.68	.83	.61
29*	24	12	56+	8	0	.56	.50	.46
30	88+	0	0	8	4	.88	.17	.31
31	68+	12	4	16	0	.68	.17	.15
32*	20	20	8	52+	0	.52	1.00	.73
33	8	16	60+	16	0	.60	.00	.06
34*	20	20	8	52+	0	.52	.83	.59
35*	80+	0	0	4	16	.80	.50	.43

+: the keyed response
*: an effectively discriminating item
_: a poorly discriminating item

Applying Ebel's criteria for D-values, we would consider items 22, 23, 28, 29, 32, 34, and 35 to be "good" items because their D-values exceed .40. If a minimum point biserial value is set at .00 + $2\hat{\sigma}_{\mu}$, the minimum acceptable point biserial value for this sample of 50 is .29. Thus we would select the same items chosen by the D-index plus items 27 and 30. An asterisk beside the item number in Table 14.4 indicates that the item passed on both discrimination criteria. Such items would be retained without revision. To identify items which require revision, we may use Ebel's criterion of having a D-value less than .20. These items are identified by a line under the item number in Table 14.4. On this test the six "problem items" are 21, 24, 25, 26, 31, and 33. (Note that item 30 is the only item which has an acceptable point biserial value, but an unacceptable D. In this case we are more inclined to be guided by the point biserial, which is based on all examinees' responses.)

Item 21 is a negative discriminator according to the point biserial correlation. Its difficulty level (p = .16) appears to be very atypical among the items on this test. In examining the distribution of responses we see that 52% of the examinees chose response 3 instead of response 4, which is keyed as the right answer. Since this is an abnormally high percentage of examinees choosing the same incorrect answer, one logical possibility is that the item has actually been miskeyed. A check of the content revealed that this was indeed the case. Item 24, also a negative discriminator, appears to be more difficult than would be ideally expected (p = .28). The responses are distributed almost equally across three possible choices, so it appears that examinees might have been responding at random. The relatively large number of "omits" also indicates that some examinees were confused about what was being asked. Such a pattern of responses suggests three possibilities. The wording of the item stem was so ambiguous that examinees could not understand the question, the item covered unfamiliar content for these examinees, or there is no correct response for this item. This item should probably be eliminated or rewritten completely.

Items 25 and 26 seem to have problems created by the content or construction of a particular response option. Item 25, the poorest discriminator, has a difficulty level (p = .72) which is only slightly greater than the ideal. From the response distribution, however, we see that option 3 has not been marked by a single examinee. The inclusion of a foil that is so obviously incorrect increases the chances that less-able examinees will select the correct answer by guessing. In revising this item, option 3 should be replaced by a more reasonable choice in the hope of attracting examinees who are uncertain of the correct response. It would also be advisable to check the answer sheets of several high-scoring examinees who missed the item to determine if they were attracted to a particular foil. Item 26, with p = .44, has two nonfunctional foils (options 1 and 2). A greater problem with this item, however, is that 52% of the examinees chose a single incorrect response. Obviously the content of the "correct" option should be reviewed to insure its accuracy; in addition, option 3 should be revised to make it less desirable. If option 3 is altered to make it less attractive, options 1 and 2 could also be rewritten to be more attractive. It should be noted that sometimes an ambiguity in the item stem causes large proportions of examinees to be attracted to a particular incorrect response. This possibility, too, should be checked.

For items 31 and 33, the item analysis results do not offer a particular clue to the cause of the poor discrimination. Careful examination of these items' content and perhaps a check of the incorrect responses chosen by high-scoring examinees may be necessary to identify the problem.

Finally, even though item 35 appears to have been a highly effective discriminator, this was achieved because a substantial percentage of the group did not answer the item. Because it is the last item on the test, we might suspect that the time limit should be extended slightly to allow all examinees adequate time to complete the test.

In classroom testing, most authorities recommend that item analysis data should be used as a basis for future test revision but do not advise discarding items in computing scores for the current class. When students took the test they assumed that all items would be counted, and they would probably not regard the grading policy as "fair" if the instructor elects to base their grades on only a subset of the items after the test is given. Furthermore, Cox (1965) demonstrated that a classroom test constructed only on the basis of item discrimination indices "would not validly measure the instructional objectives specified in the planning stage." In development of an experimental form of a test to be used for evaluation or research purposes, the scores of individual examinees who participated in the item analysis study are not of interest in themselves. In this case, the test developer may simply want to discard flawed items. For this reason it is advisable to produce and field-test more items than will be needed for each objective or content area, thus ensuring that an adequate number of good items will be available to construct a final test that is balanced appropriately in content. Finally, if the test constructor is responsible for ongoing maintenance of an item bank, the data may be used to identify items which are ready for immediate use. In addition, however, these data may be used to revise faulty items that can be field tested again. Lange, Lehmann, and Mehrens (1965) demonstrated that in such cases construction of new items required almost five times longer than revision of existing items. Thus it would be inefficient to discard flawed items without attempting to revise them if future test forms over the same material will be needed.

Cross Validation

When items are selected on the basis of a statistical criterion using the responses of a given sample, the test thus constructed should be quite effective for that particular sample—more effective than it would be for any other sample of examinees. This was dramatically demonstrated by Cureton (1950), who described the development of a test in which he selected items on which students with high GPAs had fared better than students with low GPAs. Then using the same set of item responses, he recomputed total scores for each student on only the selected items and correlated these with their GPAs. The resulting correlation was .80. Only later did Cureton

reveal that the "items" in this test were label tags in a cocktail shaker, and each student's responses were obtained by dumping the tags onto a table and awarding the student points for each tag that landed "face up." On each toss, there was an equal chance that a tag would land face up or face down. By random chance, some tags landed face up for a greater proportion of high-GPA students than for low-GPA students. These were the "items" retained from the item analysis. When only these items were counted and the students' total scores were computed with the same data, performance on these items appeared to be closely related to GPA. This apparently strong relationship would vanish, however, if the study were replicated and the tags were thrown again. Cross validation is analogous to putting the selected items back into the shaker and dumping them out again to see if the same items will function effectively a second time. To conduct a cross-validation study, the test developer uses only items that have been selected from the item analysis. These items are administered to a second, independent sample of examinees, and the reliability and/or validity of their scores is determined by using procedures described in Chapters 7 and 10, respectively.

Because of the effort involved in test administration, it is common in an item analysis/cross-validation study to try to collect data for both phases of the study in one testing session. This is done by administering all items in the item pool to all available examinees. Each test paper is randomly assigned to the item analysis or cross-validation condition. If, for example, 400 examinees were tested on 30 items, their answer sheets would be randomly divided into two groups of 200 each. The test developer would then use one set of 200 answer sheets to conduct an item analysis. On the basis of this analysis, suppose that 20 items were selected for the final version of the test. For the second set of 200 answer sheets, only 20 items would be used for cross validation in determining the reliability and validity estimates of interest. At times the test developer may wish to study whether similar results would be obtained regardless of which group was used for item analysis and cross validation. This can be done by using sample 1 for item analysis and sample 2 for cross validation and then replicating the study by using sample 2 for the item analysis and sample 1 for the cross validation. This is known as *double cross validation*.

Obviously the need for cross validation means that more subjects must be available than those used in the item analysis. One question which sometimes occurs in item analysis/cross validation concerns the proportions of the examinee groups assigned to the two samples. Although a 50/50 split (used in the previous example) is most common, in some cases other divisions might be more sensible. If the original item pool consisted of 50 items and only 400 examinees were available, it would be preferable to use a larger proportion of the group for item analysis and a smaller proportion for cross validation. In this case, at least 250 examinees could probably be used for the item analysis, leaving 150 examinees for the cross validation. (Note that this split was determined by following the rule of having 5 examinees per item in the item analysis study.)

ITEM ANALYSIS FOR CRITERION-REFERENCED TESTS

The usual purpose of a criterion-referenced test (CRT) is to assess performance on a set of tasks representative of a well-defined domain. For this reason CRT developers invariably employ some techniques for assuring the content validity of test items through expert judgments, which have been discussed in Chapter 10. Yet examination of empirical item response data may also be appropriate in development of a criterion-referenced test. In examining item response data the test developer is seldom looking only for flawed items. Instead the entire instructional process, the test development plan, and the item itself are under scrutiny. When an examinee group does not perform as expected on a particular item, this failure may be due to inadequacies in the instruction, the test specifications or objectives, or the construction of the item. The purpose of item analysis for CRTs is usually to investigate whether factors extraneous to the specified domain have contributed to performance on the test items. For this reason, early proponents of CRT were quick to point out that item statistics should not be a function of the item score variance for a single examinee group (see Millman and Popham, 1974; Popham and Husek, 1969). Thus the item discrimination indices described in earlier sections of this chapter seem inappropriate for CRTs.

Just as for norm-referenced tests, item analysis for CRTs should be undertaken with a clear purpose in mind. Specifically the test developer should know why information on item responses is needed and how it will be used. In certain situations, one or more of the following questions may be appropriate:

1. What is the item difficulty level?
2. Is the item sensitive to instruction (i.e., does it discriminate between those who have had the instruction and those who have not)?
3. Is there agreement among response patterns for particular items, as would be hypothesized from the test specifications?

A variety of item statistics have been suggested to answer these questions. One or more of each type will be presented here; the procedures described were chosen for their general applicability, ease of computation, and unambiguous interpretation. Readers interested in more detailed discussions of these and other analytic methods of CRT items should consult Berk (1980b); Harris, Pearlman, and Wilcox (1977); Harris, Alkin, and Popham (1974); and Lord (1980).

Item Difficulty

The difficulty level of a CRT item is generally defined as the proportion of examinees who answer the item correctly (p). Earlier discussion of this concept is relevant here except that the CRT developer should not be concerned with selecting items to maximize variance. The examination of item difficulty may, nonetheless,

still be important. It is probably reasonable to determine the average or median difficulty level for each cluster of items that measures a common objective. This value can be useful for assessing the effectiveness of instruction on that objective or the adequacy of the item specification. For example, items that are extremely easy on a pretest should cause the test developer to ask whether instruction in this particular content is necessary or redundant for these examinees. Items that are extremely difficult for a group after instruction may indicate that the instruction was ineffective or that the item specification includes content or processes not covered by the instructional objective. A single item that is far easier or harder than others based on the same objective should be examined for technical flaws, unintentional clues, miskeying, or ambiguities in wording that may affect difficulty regardless of subject content. The variability among difficulties for items measuring a common objective may also provide useful information. When the difficulty levels of items tapping one objective are highly variable (and this is not because of one or two technically flawed items), review of the objective or item specification seems in order.

Instructional Sensitivity

A measure of the instructional sensitivity of an item is basically a measure of how well that item discriminates between examinees who have received instruction and those who have not. Cox and Vargas (1966) suggested a procedure whereby the same group is pretested before and posttested after instruction. The discrimination statistic is defined as

$$D = p_{post} - p_{pre} \tag{14.14}$$

where p_{post} is the proportion who answer the item correctly on the posttest and p_{pre} the proportion who answer correctly on the pretest. Values for D may range from -1.00 to 1.00, with high positive values desirable. A variation on this formula uses the responses of two separate groups, one which has received instruction and one which has not. Brennan (1972) proposed a procedure requiring use of a mastery cutoff score

$$B = (U/n_1) - (L/n_2) \tag{14.15}$$

where n_1 is the number who score above the mastery cutoff and n_2 the number below, U is the number above the cutoff who answer the item correctly, and L is the number below the cutoff who answer correctly. Because B is the difference between two proportions, it also ranges from -1.00 to 1.00, with high positive values desirable. Note that B is a measure of an item's ability to discriminate at a particular cut score.

When the test developer wants to identify which items are relatively more or less effective in discriminating between instructed and uninstructed groups, several correlational procedures may also be used. Berk (1980b) suggested use of a correlational procedure derived by Saupe (1966) for selecting items for instruments designed to measure change. Use of this formula requires that each item be administered to the same group in a pretest, posttest design. Each examinee's item change score is computed as 1, 0, or -1, respectively, indicating a gain of 1 point from pre- to posttest, no gain, or a loss of 1 point on that item. In addition, the examinee's total scores on both pre- and posttests are computed along with the total change score defined as

$$D_{total} = Y - X \tag{14.16}$$

where Y is the total score on the posttest and X is the total pretest score. For each item the correlation between item change score and total change score can be computed by the Pearson product moment correlation between item change score and total change score. Saupe (1966) suggested a computational alternative formula for this correlation which yields results that are identical to the product moment correlation between item change score and total change score.

Two other procedures for assessing instructional sensitivity have been described by Millman (1974). For these the test developer must have item response data from separate instructed and uninstructed examinee groups. Through either partial correlation or stepwise regression, both of these methods allow the test developer to begin with a small subset of items and assess the test's improvement in instructional sensitivity with the inclusion of each additional item.

It should be noted that selection (or even revision) of items on the basis of instructional sensitivity indices may be philosophically incompatible with the purpose for which the CRT was originally developed. If the original item pool represented objectives defined by experts and constituted what examinees should know, the items sensitive to instruction may constitute only a subset of that content domain (i.e., what was taught effectively). Suppose, for example, that on one important objective, no instruction and no learning occurred. If items were selected on the basis of high values of D (Cox and Vargas, 1966) or B (Brennan, 1972), all items on this objective would be eliminated from the test. Obviously, the elimination of these items would not improve the content validity of this CRT. Further, a procedure such as Saupe's correlation of item change scores is actually designed to yield a norm-referenced measure of change; such an index results in selection of items on which there is variability in the change scores, which may not be a primary goal for the CRT developer. Thus it may not be reasonable to compute indices of instructional sensitivity routinely for every CRT. It is important to have a sound rationale for use of such information and to recognize that selection of items on the basis of such statistical criteria may not be consistent with domain mastery approaches to measurement. Instructional sensitivity indices may, however, provide useful information about the type of content that is being taught effectively or ineffectively in a particular instructional program. Use of item analysis data in this way presupposes that the test user has confidence in the item quality.

Indices of Agreement

There may be times when the test developer is interested in studying similarity of responses of one group of examinees to each possible pair of items written from the

same specifications. This might be reasonable for situations in which subsequent tests will be developed by selecting items at random from the set of items now being field-tested, and the test developer wants to know if the items can be considered "interchangeable." In this case, data for each pair of item responses can be arranged in a fourfold table such as Table 14.5. Harris and Pearlman (1977) discussed a variety of statistical indices which can be applied to such data. The selection of the appropriate statistic depends on what the test developer wants to know. Some illustrative questions that a test developer might ask and a few of the indices described by Harris and Pearlman will be used to illustrate this item analytic approach. Additional probability models for the examination of item equivalence have been described by Wilcox (1977b). These models are more complex because they take into account the possibility of inappropriate responses by examinees (i.e., that an examinee may know an answer but make an incorrect response and vice versa). The procedures described here do not allow for this possibility.

As an initial example, consider the test developer who wants to know "if two items are measuring the same thing." In this case, it is appropriate to use a test of statistical independence for the responses to the two items. A simple chi square statistic can be computed with the formula

$$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)} \tag{14.17}$$

where n is the number of persons in the sample, and a, b, c, and d are the cell frequencies in Table 14.5. This computation is illustrated in Table 14.6. This computed χ^2 statistic is compared to the value of χ^2 with 1 degree of freedom at the

TABLE 14.5. Fourfold Table for Displaying Responses to a Pair of Items

		Item 1	
		+	−
Item 2	+	a	b
	−	c	d

(a)

		Item 1	
		+	−
Item 2	+	20	5
	−	5	20

(b) Frequencies for 50 examinees.

TABLE 14.6. Illustrative Applications for Indices of Agreement and Association for Item Data from Table 14.5(b)

Illustrative Question	Formula	Computation	Illustrative Interpretation
Do these two items measure the same thing?	$\chi^2 = \frac{n(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)}$	$\frac{50(400 - 25)^2}{(30)(30)(30)(30)} = 8.68$	This computed χ^2 value exceeds the test statistic of 3.84 (for alpha = .05). Therefore the dependence in responses to these two items is significant, supporting the interpretation that the items measure the same skill or content.
What proportion of examinees passed or failed both items?	$P = (a + d)/n$	$(20 + 20)/50 = .80$	There is consistent performance on these two items for 80% of the examinees.
Did examinees perform significantly better on one item than on the other?	$\chi^2 = \frac{(\lvert b - c\rvert - 1)^2}{b + c}$	$\frac{(\lvert 5 - 5\rvert - 1)^2}{5 + 5} = .10$	The computed χ^2 value does not exceed the test statistic (3.84 at alpha = .05). We fail to find a significant difference in item difficulties and thus cannot conclude that examinees have learned the content measured by one item better than the content measured by the other item.

desired alpha level. If the computed value of the test statistic is significant (as we find in our example), the responses to the two items may be regarded as dependent or associated. Failure to find a significant association between the responses to two items written from the same item specification should raise some question about the adequacy of the item specification or the technical quality of one or both of the items involved.

Finding that there is greater similarity in examinees' responses than would be expected by chance, the test developer may next wish to compute a statistic which describes the degree of that similarity. One statistic which is readily interpretable is simply the proportion of agreement, $(a + d)/n$. This is simply the proportion of examinees who responded consistently to the two items. See Table 14.6 for an illustrative computation and interpretation of this statistic. Harris and Pearlman (1977) favor this statistic because it can be averaged and meaningfully interpreted over multiple item pairs and because it is an unbiased estimator. Other statistics suitable for describing the degree of agreement in responses to a pair of items would be coefficient kappa (described in Chapter 8), Yule's Q (recommended by Harris and Pearlman), or the phi coefficient.

A slightly different question is whether the difficulties of two items are equal in the population of examinees. Put another way, the question is whether observed differences in difficulty are small enough to be attributed to sampling errors. Such a question would be appropriate if the test developer hypothesizes that because of the instructional program, the content of the two items should have been learned equally well, and the developer wishes to test this hypotheses. In this case, Harris and Pearlman suggest a χ^2 statistic of the form

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c} \tag{14.18}$$

This computed statistic is again contrasted to the value of χ^2 at 1 degree of freedom at the desired alpha level. A significant value indicates a difference in item difficulties for the sample of examinees that is too large to attribute to chance. (See Table 14.6 for this computation applied to our two-item example.)

In general, indices of item agreement may be appropriate when specific questions arise about how examinees have performed on particular pairs of items. Different questions require use of different statistical indices. It is unlikely that applying one or more of these indices to all possible pairs of items on a large test would yield meaningful information. For small item subsets (e.g., items written from the same specification or matched to the same objective in a content validation study), indices of agreement may provide empirical evidence to support rational judgments about item similarity. On some criterion-referenced tests, examinees' performance is reported separately for a number of different objectives or skill areas, often with only a small number of items measuring each objective. In such cases, investigation of the degree of agreement on item performance within objective or skill seems warranted to support the intended test score interpretations.

As a final comment, which applies to all item analyses methods for CRT, we note

that Harris (1974) and Harris, Pearlman, and Wilcox (1977) took the position that if items are randomly selected from an item pool to represent a carefully structured content domain, no item should be deleted on the basis of item analysis. They contend that careful review of items during the test construction phase and a sound test development plan should be sufficient to achieve valid test scores. The role of item analysis data would be only to provide information about where the instructional system or the test development plan has failed. This view seems to imply that item writers for such a test are immune from producing flawed items, or that valid inferences about examinees' performance can still be drawn from "bad" items. This position has been criticized by Messick (1975) and has led Petersen (1979) to question whether it is reasonable to rely on samples of items in test construction when the goal is to infer to performance on some larger item domain. Thus the test user should be aware that the ways in which item analysis data are used may rest on different philosophical views about test construction and test score interpretation.

SUMMARY

Item analysis is the computation and examination of any statistical property of an item response distribution. For dichotomously scored items, the best-known descriptor is probably item difficulty (p), which denotes the proportion of examinees answering the item correctly. Item p-values are not necessarily an index of item quality; however, for norm-referenced tests, items with p values near .50 will allow total score variance, and hence reliability, to be maximized. When multiple-choice items are used, some allowance for guessing must be made so that p-values greater than .50 are ideal for maximizing total score reliability.

Indicators of item discrimination are used to assess how well performance on an item relates to performance on some other criterion. This criterion may be the total score on the test or some other variable. When total test score is used, selection of highly discriminating items leads to higher internal consistency of test scores. For items not dichotomously scored, the Pearson product moment correlation may be useful in correlating item scores with criterion scores. When the items are dichotomously scored, useful indices may be the index of discrimination (D), the biserial correlation, or the point biserial correlation. These are appropriate when there is a continuous criterion score. When the criterion is dichotomized, the phi or tetrachoric coefficient may be useful. Comparisons of these five indices indicate that their results are usually highly correlated, with greatest differences occurring for items in extreme difficulty ranges. Thus the choice among these procedures should be made on the nature of the variables, the type of information desired, and computational considerations. Formulas for calculating the standard errors of these statistics were also presented.

The item reliability index is the product of the item score standard deviation and item correlation with total test score, and the item validity index is the product of the item score standard deviation and item score correlation with an external criterion.

Total test score variance and coefficient alpha may be estimated from the time reliability indices. The total test score validity coefficient may be estimated from the item reliability and validity indices.

Typically in an item analysis study, the test developer will decide on the properties of total test scores that are of greatest interest and identify the item parameters that have greatest effect on these properties. The items are administered to a sample of examinees (usually 100 or more), and the item responses are analyzed. A plan is established for selecting the items that make the greatest contributions to the test score characteristics of interest (and possibly for identifying how malfunctioning items should be revised). After the final subset of items is selected, it is administered to a second independent sample for cross validation. (In many cases, for convenience half of the original field-test sample is held in reserve to use in the cross-validation study.)

For criterion-referenced tests, traditional item analysis indices may not be appropriate because most item discrimination statistics are designed to favor items on which there is substantial variation among examinees. This is a goal more appropriate for a norm-referenced test. Nevertheless, developers of CRTs may be interested in item difficulties, item sensitivity to instruction, and degree of agreement among specific pairs of items. Some statistical procedures proposed in the literature for investigating these types of properties were presented. Finally, it is important to note that selection of items on the basis of item analysis data may be less appropriate for CRTs than for norm-referenced tests since the representatives of the domain of interest may be reduced by such item selection on statistical criteria. Even for norm-referenced tests, the nature of the construct being measured may be altered if items are selected purely on the basis of statistical criteria without regard to the initial test specifications.

Exercises

1. Consider the following test data for 2 items and 10 persons:

Person	Item 1	Item 2	Total Score
1	0	1	12
2	0	1	15
3	1	1	16
4	0	0	10
5	1	1	7
6	1	0	5
7	1	0	6
8	0	1	10
9	1	1	15
10	1	0	13

A. What are the indices of discrimination for the two items based on a 50/50 split? What are they based on a top/bottom 27% split?
B. What are the biserial $\hat{p}$ values for these two items?

C. What are the point biserial $\hat{p}$-values?
D. How do the underlying assumptions differ in use of a biserial vs. point biserial item/total score correlation?

2. Examine the item analysis information and answer the questions that follow.

Item Number	Item Responses 1	2	3	4	Other	Item Diff.	Item Discrim.	Point Biserial Corr.
21	11	5	149 +	11	0	0.84	0.15	0.23
22	34	11	127 +	4	0	0.72	0.30	0.27
23	11	4	32 +	128	1	0.18	0.13	0.13
24	30	34	1	111 +	0	0.63	0.64	0.52
25	6	0	20	150 +	0	0.85	0.34	0.34

A. Which item appears to be most in need of revision?
B. Items 21 and 25 are nearly equal in difficulty, yet their item discrimination values differ. How can this be explained?
C. If you wanted to improve item 21, how would you try to change it?
D. Suggest at least two plausible hypotheses to account for the negative discrimination of item 23.
E. Estimate the values of the biserial correlation between item and total score for items 24 and 25.
F. What is the minimum acceptable value for the point biserial correlation if the test developer wants to be 95% certain that this statistic is significantly greater than .00?
G. Assuming a random guessing model, what is the ideal difficulty level for these items if the test developer wishes to maximize test score reliability?

3. Assume that item 21 and 22 in the preceding example were drafted according to the same item specification. The bivariate response distribution for this pair of items is

		Item 21 +	Item 21 −
Item 22	+	120	7
	−	29	20

A. What statistical procedure could be used to test the hypothesis that items 21 and 22 measure the same skill or knowledge? Make the computation and state your conclusion.
B. What statistical procedure could be used to test the hypothesis that item 22 is a more difficult item than item 21? Make the computation and state your conclusion.
C. Explain how the statistic used in Exercise 3.A is related to the phi coefficient. Do these two indices provide the same information?

4. For the following situation indicate which item analysis statistic would be most appropriate. Use this key to indicate your response:

1. Phi coefficient between item scores
2. Point biserial correlation between item and total score
3. Point biserial correlation between item and external criterion score

4. Pearson product moment correlation between item and total score
5. Pearson product moment correlation between item and score on an external criterion

A. A counselor or a college counseling center wishes to select valid items on an attitude scale, to be used for predicting the behavior of college males, where each item is scored 1 to 7 and the criterion is joining a social fraternity (students are coded as "members" or "nonmembers").
B. A test constructor desires to improve the test-retest reliability of a behavior checklist where each item is scored 1 point for "yes" and 0 for "no."
C. A college professor wishes to increase the internal consistency of a 50-item unit examination which consists of 4-choice items, scored with 1 point for each item answered correctly.
D. A psychologist wishes to identify Likert-type items on which his subjects' responses are influenced by their ages.
E. A psychologist has developed an inventory to assess racial prejudice. Each item is scored on a five-point scale and was generated from a specific account of a critical incident by a large cross-section of minority group members. The psychologist is unable to identify a single behavior (which it is feasible to observe in practical time limits) that can serve as a criterion.

5. A. Derive the formula for the point biserial correlation coefficient, beginning with the Pearson product moment formula.

$$\rho_{XY} = \frac{\Sigma(X - \mu_X)(Y - \mu_Y)}{N\sigma_X\sigma_Y}$$

where X is a dichotomous variable scored 0 or 1.

B. In light of this derivation, explain why point biserial correlation values of 1.00 are seldom observed.

6. Some authors have argued that in order for scores on an instrument to have any claim to content or construct validity, there should be evidence that each item on the test correlates with total test score. Since item correlation with total score is more directly related to internal consistency, what is the logical rationale for the authors' position?
7. A graduate student has designed a dissertation study which requires measurement of attitudes of the mothers of newborn infants who have certain birth defects. The student will be obtaining subjects from all hospitals in the community with maternity wards but still anticipates having only a small sample (approximately 25 to 30) over a one-year period. Faculty on the four-person advisory committee offer four separate suggestions:
   1. Use the first 15 subjects for item analysis; refine the instrument and use the last 15 subjects in the study itself.
   2. Use the instrument without any item analysis and take the chance that it will provide "good" information.
   3. Administer all items to all subjects; do an item analysis and eliminate bad items; then rescore the instrument for the same subjects using only the "good" items.
   4. Do not conduct the study.

   Discuss the advantages and disadvantages of each suggestion and decide which you favor.
8. Suppose that the test developer is considering creating a subtest consisting only of the 15 items for whom item analysis data are presented in Table 14.4. Estimate the approximate variance and reliability of scores on this subtest.

Chapter 15

INTRODUCTION TO ITEM RESPONSE THEORY

In the preceding chapter we pointed out that item selection is often based on indices of difficulty and discrimination. Although this practice works effectively in many test construction situations, it is conceivable that test development decisions could be improved by using additional information about item responses. Consider the following two items and their respective response patterns:

1. Item A is answered correctly by all examinees who earned a score of 50 points or more on the test; it is answered incorrectly by all those who scored lower than 50.
2. Item B is answered correctly by 20% of the examinees who earned 45 points, by 40% of those who earned 50 points, by 60% of those who earned 55 points, by 80% of those who earned 60 points, and by all of those who earned 65 points or more.

Clearly there is a difference in the observed pattern of item responses to these two items, which might be important for the test developer to know; yet classical item analyses statistics do not provide information about how examinees at different ability levels on the trait have performed on the item. One approach to test development which yields a more complete picture of how an item functions is known as *item response theory* or *latent trait theory*.

With item response theory the test developer assumes that the responses to the items on a test can be accounted for by latent traits that are fewer in number than the test items. Indeed, most applications of the theory assume that a single latent trait accounts for the responses to items on a test. At the "heart" of the theory is a mathematical model of how examinees at different ability levels for the trait should respond to an item. This knowledge allows one to compare the performance of examinees who have taken different tests. It also permits one to apply the results of an item analysis to groups with different ability levels than the group used for the