## Introduction

*The item analysis is an important phase in the development of an exam program. In this phase statistical methods are used to identify any test items that are not working well. If an item is too easy, too difficult, failing to show a difference between skilled and unskilled examinees, or even scored incorrectly, an item analysis will reveal it. The two most common statistics reported in an item analysis are the item difficulty, which is a measure of the proportion of examinees who responded to an item correctly, and the item discrimination, which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not. An additional analysis that is often reported is the distractor analysis. The distractor analysis provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item. Once the item analysis information is available, an item review is often conducted.*

## Item Analysis Statistics

### Item Difficulty Index

The item difficulty index is one of the most useful, and most frequently reported, item analysis statistics. It is a measure of the *proportion* of examinees who answered the item correctly; for this reason it is frequently called the *p-value*. As the proportion of examinees who got the item right, the p-value might more properly be called the item easiness index, rather than the item difficulty. It can range between 0.0 and 1.0, with a higher value indicating that a greater proportion of examinees responded to the item correctly, and it was thus an easier item. For criterion-referenced tests (CRTs), with their emphasis on mastery-testing, many items on an exam form will have p-values of .9 or above. Norm-referenced tests (NRTs), on the other hand, are designed to be harder overall and to spread out the examinees' scores. Thus, many of the items on an NRT will have difficulty indexes between .4 and .6.

### Item Discrimination Index

The item discrimination index is a measure of how well an item is able to distinguish between examinees who are knowledgeable and those who are not, or between masters and non-masters. There are actually several ways to compute an item discrimination, but one of the most common is the *point-biserial correlation*. This statistic looks at the relationship between an examinee's performance on the given item (correct or incorrect)

and the examinee's score on the overall test. For an item that is highly discriminating, in general the examinees who responded to the item correctly also did well on the test, while in general the examinees who responded to the item incorrectly also tended to do poorly on the overall test.

The possible range of the discrimination index is -1.0 to 1.0; however, if an item has a discrimination below 0.0, it suggests a problem. When an item is discriminating negatively, overall the most knowledgeable examinees are getting the item wrong and the least knowledgeable examinees are getting the item right. A negative discrimination index may indicate that the item is measuring something other than what the rest of the test is measuring. More often, it is a sign that the item has been mis-keyed.

When interpreting the value of a discrimination it is important to be aware that there is a relationship between an item's difficulty index and its discrimination index. If an item has a very high (or very low) p-value, the potential value of the discrimination index will be much less than if the item has a mid-range p-value. In other words, if an item is either very easy or very hard, it is not likely to be very discriminating. A typical CRT, with many high item p-values, may have most item discriminations in the range of 0.0 to 0.3. A useful approach when reviewing a set of item discrimination indexes is to also view each item's p-value at the same time. For example, if a given item has a discrimination index below .1, but the item's p-value is greater than .9, you may interpret the item as being easy for almost the entire set of examinees, and probably for that reason not providing much discrimination between high ability and low ability examinees.

### Distractor Analysis

One important element in the quality of a multiple choice item is the quality of the item's distractors. However, neither the item difficulty nor the item discrimination index considers the performance of the incorrect response options, or distractors. A distractor analysis addresses the performance of these incorrect response options.

Just as the key, or correct response option, must be definitively correct, the distractors must be clearly incorrect (or clearly not the "best" option). In addition to being clearly incorrect, the distractors must also be plausible. That is, the distractors should seem likely or reasonable to an examinee who is not sufficiently knowledgeable in the content area. If a distractor appears so unlikely that almost no examinee will select it, it is not contributing to the performance of the item. In fact, the presence of one or more

implausible distractors in a multiple choice item can make the item artificially far easier than it ought to be.

In a simple approach to distractor analysis, the proportion of examinees who selected each of the response options is examined. For the key, this proportion is equivalent to the item p-value, or difficulty. If the proportions are summed across all of an item's response options they will add up to 1.0, or 100% of the examinees' selections.

The proportion of examinees who select each of the distractors can be very informative. For example, it can reveal an item mis-key. Whenever the proportion of examinees who selected a distractor is greater than the proportion of examinees who selected the key, the item should be examined to determine if it has been mis-keyed or double-keyed. A distractor analysis can also reveal an implausible distractor. In CRTs, where the item p-values are typically high, the proportions of examinees selecting all the distractors are, as a result, low. Nevertheless, if examinees consistently fail to select a given distractor, this may be evidence that the distractor is implausible or simply too easy.

### Item Review

Once the item analysis data are available, it is useful to hold a meeting of test developers, psychometricians, and subject matter experts. During this meeting the items can be reviewed using the information provided by the item analysis statistics. Decisions can then be made about item changes that are needed or even items that ought to be dropped from the exam. Any item that has been substantially changed should be returned to the bank for pretesting before it is again used operationally. Once these decisions have been made, the exams should be rescored, leaving out any items that were dropped and using the correct key for any items that were found to have been mis-keyed. This corrected scoring will be used for the examinees' score reports.

### Summary

In the item analysis phase of test development, statistical methods are used to identify potential item problems. The statistical results should be used along with substantive attention to the item content to determine if a problem exists and what should be done to correct it. Items that are functioning very poorly should usually be removed from consideration and the exams rescored before the test results are released. In other cases, items may still be usable, after modest changes are made to improve their performance on future exams.