Introduction

After draft items have been reviewed and field tested, and item analysis statistics have been obtained on the items, operational test forms can be assembled. Each new test form for the exam program ought to be assembled to satisfy all of the elements in the test specifications, and particularly in the test blueprint. Items are selected from the bank to satisfy the content and cognitive proportions specified in the test blueprint, as well as to meet certain statistical criteria. This process helps ensure that each resulting test form satisfies the intended purposes of the test. Most exam programs follow this process to develop not just a single test form, but multiple forms. These multiple test forms should be assembled so that they are as similar to one another as possible, or parallel. The use of multiple, parallel test forms provides improved test security for the exam program. After test administration, statistical equating of the parallel forms is often needed.

Aspects of Test Assembly

Statistical Specifications

The test blueprint for an exam program is used to guide the assembly of one or more test forms. In addition to selecting items to satisfy content and cognitive specifications, statistical criteria are often also used when assembling test forms. For example, an exam program may have a test assembly criterion that no item on the test should have a discrimination index less than .2. Norm-referenced tests are frequently assembled to have a broad range of item difficulty indexes in order to help spread out the examinees' test scores. On the other hand, criterion-referenced tests, with their emphasis on mastery testing and their strong focus on linking the items to job performance, are more likely to have a narrow range of high item p-values.

Parallel Forms

When an exam program has multiple test forms it is critical that they be assembled to be parallel to one another. Two or more forms of an exam are considered parallel when they have been developed to be as similar to one another as possible in terms of the test specifications and statistical criteria.

The primary reason for having multiple exam forms is to improve test security. In the simplest case, an exam program may have two test forms. One form may be currently

used in regular administrations and the second form may be available for examinees who are restesting. Alternatively, the second form may be held in reserve, to be available if the security of the first form is breached. High-stakes exam programs with greater security concerns may have multiple forms in use at every test administration. To obtain even greater test security, some exam programs use new test forms at every administration.

Alternate forms of a test may also be developed over the life of an exam program to address content changes in the field. As new information is discovered, or laws regarding occupational practice and regulation are changed, the text of some items may become unclear or even incorrect. To stay current and accurate, new items, and new test forms, may be needed.

Equating Methods

Even when every effort is made to develop parallel forms, some differences in the statistical characteristics between the test forms can still be expected. The statistical method used to resolve these test form differences is called equating. A test form is statistically equated to another test form to make the resulting test scores directly comparable.

In order to conduct an equating, data must be collected about how the test forms differ statistically. That is, information is needed to determine whether differences in the two groups of test scores are caused by a difference in the proficiency of the two examinee groups or by a difference in the average difficulty in the two tests. Two of the most common data collection designs that are used for equating are the *random groups design* and the *common-item nonequivalent groups design*.

In the *random groups design*, two (or more) test forms are given at a single test administration: the test forms are distributed across examinees through a spiraled process. For example, Form A may be given to the first examinee, Form B to the second examinee, Form A to the third examinee, and so on. When the random assignment of test forms to examinees is used, the two examinee groups can be considered equivalent in proficiency. Any statistical differences across the two groups on the two test forms can be interpreted as a difference in the test forms. For example, if the group of examinees who took Form A overall performed better than the group of examinees who took Form B, you can probably assume that Form A is easier than Form B.

Sometimes, for pragmatic reasons, only a single test form can be administered on a given date. In these instances, the *common-item nonequivalent groups design* can be used. Since in this case the test forms are not randomly assigned to the two examinee groups, you cannot assume that the two groups have the same average proficiency. In this data collection design, therefore, it is necessary to identify any difference in proficiency between the two examinee groups. To do this, a subset of identical test items is placed on both test forms. This set of common items is referred to as the anchor test. Because these common items are administered to all examinees, they can be used to estimate differences in proficiency across the two examinee groups. Once the examinee group differences have been identified, any remaining score differences can be interpreted as a difference in the difficulty of the two test forms. Equating can then properly adjust for these test form differences.

Once the data have been collected across the two (or more) test forms, the equating can be conducted. There are several statistical methods for equating. In one commonly used method, *linear equating*, the mean and standard deviation of Form B are placed on the scale of Form A. This method can be used with small samples and is most accurate near the mean of the score distribution. Another equating method, *equipercentile equating*, determines the appropriate percentile rank on Form A for scores from Form B. The equipercentile method requires larger sample sizes; however, it results in greater accuracy along the entire score scale.

If they are conducted properly, both linear and equipercentile equating enable you to directly compare examinees' performances, even when those examinees took different test forms and even if those test forms differed in difficulty.

Summary

Each test form for an exam program should be assembled according to the test blueprint that was developed. When multiple forms of an exam are needed they should also be assembled so that they are highly similar to one another in terms of statistical criteria such as average difficulty. After the forms have been administered they should be statistically equated to overcome any remaining test form differences.