INTRODUCTION

After a set of drafted items have been reviewed by test development and subject matter staff, the items should be ready for field testing on examinees. When items are field tested, they are not scored and they are not used to measure examinees; instead, the items themselves are evaluated. An item field test is conducted to collect information about how well the items are likely to function operationally, before they are actually used to contribute to an examinee's score. For that information to be accurate, it is important that the field test includes appropriate examinees. The field test examinees should be representative of the test population; that is, they should be as similar to future test-takers as possible. The examinees should also be motivated; that is, they should be attempting to do as well as possible when they respond to the items. Items may be evaluated at different test development phases, under pilot test, field test, and pretest conditions. In all these instances, data are collected to ensure the high quality of the items and tests. The information is then used to conduct an item analysis and review of the items, prior to allowing items to be included on operational test forms.

TYPES OF FIELD TESTS

Item Evaluation Phases

Items may be tested out on actual examinees at several points in the development of an exam program. These various item evaluations are conducted in somewhat different ways and specific terms are used to refer to each of these types of item evaluations. (However, while these item evaluation phases are recognized as distinct, it should be noted that the terms are not used with complete consistency across test development efforts.) A preliminary, small sample evaluation of items is often referred to as a *pilot test*. In a larger sample *field test*, a greater number of examinees is used in order to obtain stable statistics on the items being evaluated. For an established exam program, the term *pretest* is used to refer to the process of collecting data to evaluate the performance of new, additional items.

Pilot Testing

Pilot testing is a phase of item evaluation that is sometimes conducted prior to large scale field testing. The primary advantage of conducting a pilot test is that it may enable you to identify item flaws relatively quickly and inexpensively. These item flaws can then be corrected before the field test phase, when the items will be further evaluated. This two-stage evaluation process is likely to result in a higher percentage of items ready for operational use in a short amount of time.

A typical pilot test might be conducted on a sample of as few as 15 to 30 examinees. These examinees should be similar to the intended population of examinees for the test program and they should span a reasonable range of proficiency. While the pilot test provides you with the opportunity to evaluate the test instructions and planned administration procedures as well as the test items, it is still a less formal proceeding than the later field test. Before beginning the test administration, the examinees should be reminded that it is the items, and not the test-takers, that are being evaluated. During the test administration, exam time limits and other standardized procedures should be followed. The examinees should be observed as they respond to the test, to note any indications of problems or confusion about particular items. After the examinees have finished, a quick item analysis of their responses should be conducted. With this information at hand, the examinees should then be interviewed, or debriefed, in order to obtain their qualitative comments and reactions to the items.

Field Testing

The term field testing has been used to describe any of the phases where items are tried out on actual examinees prior to operational use; however, it most commonly refers to a specific type of item trial. A typical item field test is conducted when a new exam program is under development. A set of items are assembled into a test form, including examinee instructions. The test form is assembled to match the content areas and proportions specified in the test blueprint. In some cases, the overall test length is increased to allow for the possibility that, upon evaluation, some of the items will not be considered acceptable.

The test form is then administered to a large, representative sample of examinees under realistic timed and proctored conditions. Approximately 100 to 200 examinees may be used in this field testing stage, to provide large enough sample sizes for stable statistics when the item analysis is conducted. For a new exam program, there are certain challenges to obtaining field test examinees, since the examinees do not receive scores and taking the exam is usually voluntary for them, rather than mandatory. Nevertheless, it is important to obtain an appropriate sample and to ensure that the examinees who participate respond in a motivated fashion. To address these concerns, incentives are sometimes offered to the examinees to encourage their motivated participation. These incentives may include the opportunity to practice for the exam under realistic conditions and the opportunity to become familiar with the types of items that the operational test may be expected to have. Furthermore, cash or other prizes are sometimes also awarded to a percentage of the top-scoring examinees in a field test.

After the field test data have been collected and an item analysis has been conducted, it may become clear that some items need to be modified or dropped altogether. Any items that are modified substantially should be returned to the item bank for another round of field testing before they are placed on an operational test form.

Pretesting

After an exam program has been established and an initial test form has been developed, there is a continuing need for new items in order to develop additional or replacement test forms. For an ongoing exam program, an ideal approach to collecting item performance data is to place a certain number of new items on every operational test form. In this way, every time an exam is given examinee response data on the new items can be collected. Furthermore, the examinees who respond to these pretest items will clearly be representative of the test population. And, as long as the test form design does not reveal which items are operational and which are being pretested, the examinees will respond in a motivated fashion. As with other types of item evaluations, the pretest items are not used to contribute to examinees' scores. Instead, after pretesting is complete the new items undergo item analysis and review. Items that pass this review with minimal or no changes are then available to be used when new test forms are developed. Items that have been more substantially modified must undergo an additional round of pretesting before they can be used operationally.

Summary

There are a number of conditions in which data is collected to evaluate the quality of new test items. These various item evaluations may be conducted during the development of a brand new exam program or they may be part of the ongoing maintenance of an existing program. Item evaluations may be conducted early in the development of an exam program in order to collect small sample qualitative data. They may also be conducted later to collect large sample data that will support item analysis. Large sample data collection efforts may be conducted for both brand new exam programs and as part of ongoing exam maintenance. Under all of these conditions of item evaluation the goal is the same: to ensure the high quality of the items themselves and of the resulting tests.