Item Calibration and Pretesting

Question: *What is item calibration, and what role does it play in testing?*

Answer: Item calibration is a part of the larger topic of item response theory (IRT). Crocker and Algina describe *person-free item calibration* as the process by which "the parameters of large numbers of items can be estimated even though each item is not answered by every examinee." (p. 363) They also provide an example in which different forms of a test have a set of linking items that enable the creation of a common scale on which all item parameter estimates can be expressed. Item parameters differ, depending on the IRT model used, but all include item difficulty level. A two parameter model adds item discrimination, and a three parameter model in addition has a "pseudo-chance" or guessing parameter.

The goal of item calibration is to develop a pool or bank of items which are on the same scale. Although IRT is widely used in paper-and-pencil testing, the advantages of this approach are most evident in computer-based testing. There are two major applications of use of a calibrated item pool. When fixed-form tests are offered continuously on computer (rather than being offered in a "window" of limited time), test assembly using calibrated items enables the calculation of the passing score for the test form to be completed before the first candidate tests. In contrast, traditional common-item equating requires collection of a number of candidate responses over a period of time in order to use the statistics from the administrations to determine the passing score for the test.

The second major use of a pool of calibrated items is for adaptive testing. In adaptive testing, a candidate ability estimate is obtained after administration of a small number of initial items, and subsequent item selections are targeted to the candidate's estimated ability. If the test pool is large, overlap of items between individual tests is minimal.

In addition to these two applications, another possibility is linear testing "on the fly." Each candidate gets a somewhat different test, but the items are selected to meet a target test characteristic curve instead of being selected based on the candidate's estimated ability.

All of the approaches above using IRT for computer-based testing require that all scored items have been previously administered and calibrated. New items get into the mix by presenting them as unscored pretest items. Once sufficient candidate responses have been obtained, the pretest items are taken offline and calibrated. Then they can be introduced into the used item pool as scored items.

Crocker, L. and Algina, J. (1986) *Introduction to Classical and Modern Test Theory*. Harcourt Brace Jovanovich College Publishers: Fort Worth, 527 pp.