CRITERION MEASURES:

INSTRUCTION VS. SELECTION RESEARCH[1]

by

Joseph Hammock
Bell Telephone Laboratories, Incorporated
Murray Hill, New Jersey

With the current wide interest in new media and techniques of instruction, such as television, teaching machines, and preprogrammed text materials, frequent efforts are being made to compare different courses and techniques experimentally. It is of obvious importance that the dependent variables or criterion measures used in such studies be appropriate for experimental research. That is, the criterion measures should be suited to the specific task of assessing any differential effects of two or more instructional treatments along the dimensions in the universe of criterion behaviors.

Frequently, tests are accepted as criterion measures in experimental instructional research because they are already available, having been developed as course achievement tests for grading purposes or as criterion measures for selection research, and when measures are devised specifically for use in instructional research, they are often developed according to the accepted and widely published procedures appropriate to the

---

[1]Paper read at the meetings of the American Psychological Association, September, 1960.

development of achievement grading tests or selection criterion tests. This paper reports an analysis which indicates that the best measures for these other uses may not be, and probably are not, optimally suited as criterion measures for instructional research. (Although the analysis is phrased in terms of the differences between instructional and selection criterion tests, it should be understood that what is said about selection criterion tests holds equally well for any test or measure that is constructed, through item analysis, to be maximally sensitive to individual difference variables.)

First, let us look briefly at the standard procedures used in building a selection criterion measure. The first step is to arrive at an agreed upon definition for the universe of criterion performance events, and the next is to extract a representative sample of such events. Then the events in the sample are structured into scorable test items. These steps, up to this point, yield a set of items with content validity with respect to the criterion universe which could be used equally appropriately as a criterion measure for selection or instructional research. The problem arises with the steps taken beyond this point in an item analysis that attempts to produce a measure that is maximally sensitive to differences among individuals from a specific population that has received a particular type of treatment.

The principal goal of the selection testing technology is the efficient assessment and prediction of differential effects on individuals of the same or comparable treatments. Toward this end, the builder of a selection criterion measure, realizing that test administration time is limited and being desirous of obtaining maximum discrimination or variance among individuals, usually seeks out certain information on each of the items in his sample that will permit him to identify and select out those items which yield most valid discrimination between individuals, following the particular treatment with which he is concerned. In general, there are two such classes of information that he obtains for each item: one is a difficulty index, which is the proportion of subjects who pass the item, and the other is a validity or discrimination index, an index of how highly the item correlates with some estimate of individual differences on the over-all criterion. The point will be made that the selection of items according to either of these two indices is likely to attenuate the appropriateness of the test for use as a criterion measure in experimental instructional research.

Consider the difficulty index, the proportion of subjects in the sample passing the item; call this proportion "p". Since the variance contribution of an item to the test depends on the product of "p" and "1-p", then as "p" approaches

either one or zero, the variance contribution of that item approaches zero. That is, if everyone, or if no one, passes an item then that item is not discriminating between individuals for that sample and contributes nothing to an individual difference criterion; so the item is thrown out of the selection criterion test. (The most vigorous seekers after criterion variance and short criterion tests, under many conditions, use the procedure of dropping all items differing markedly from a "p" level of .5, the value at which the variance contribution is maximum.) Now, let us ask what effect this rejection of items with very low or very high difficulty indices would have on a test as an instructional research criterion. We noted earlier that the main requirement for an instructional criterion test is that it be maximally sensitive to any differential effects of various instructional treatments and, of course, this sensitivity should be over the full range in which differential effects might occur. Conceivably, some very poor combination of instructional content, method, and practice would yield very little increase in performance, whereas the optimum combination would yield a performance level approaching that of the most experienced and competent performers. This range of variation of treatment effects would most likely be much wider than the range of the distribution between the most apt and most inapt individuals

at any one point during or after instruction, for any one particular instructional treatment. It follows then that those items dropped from a test because they were passed by either all or none of the subjects, at one particular point during that particular treatment, would be essential in a test required to discriminate the effects of other treatments that might cause either the highest scorers to go higher or the lowest scorers to drop lower. So we see that selecting criterion items on the basis of sensitivity only to the range of individual differences found after one amount of one particular treatment restricts the usefulness of the test for purposes of measuring the effects of different treatments. The appropriate use of difficulty indices in constructing an instructional criterion measure will be seen in the following discussion of discrimination indices.

We noted earlier that the discrimination which is of interest in constructing a selection criterion is that between the apt and the inapt within a group, all of whom have had the same treatment. On the other hand, for an instructional criterion, we would be interested in discriminating between groups that are similar in aptitude but have had differential instructional treatments. To produce an instructional criterion measure with maximal sensitivity over the full range of instructable performance, items should be selected for maximum discrimination between two groups that are equated for aptitude

variables but are maximally different in instructable performance. Two such groups would be (a) a group of persons who are highly experienced and competent in the area of criterion performance and (b) the same group before they received any instruction or practice in that area of performance. (An alternative for "b" would be, of course, a different group of persons who are matched with the experienced group for aptitude variables but have had no instruction in the area.) The index of instructional discriminability for an item would be the amount of change in "p" between the no instruction group and the maximum competence group. Therefore, the most discriminating items for a full range instructional criterion would be those that change from a "p" level of zero for the no instruction group to a "p" level of one for the maximum competence group. These items and those showing the largest gains in "p" level would be retained for the test. The items that show little or no change in "p" level would be assumed to be either irrelevant to criterion performance or noninstructable and would be dropped from the test. (More sensitive indices would be the amount of change in "p" level between the no instruction and the competent groups for each of several levels of aptitude or ability within the distributions.) If the range of sensitivity over which the treatments of concern might vary can be predicted to be something less than the full range of instructable performance, then the items should be selected to be maximally discriminative over this narrower range.

While the operations recommended here for building instructional criterion measures are roughly the reverse of those widely recommended for selection criterion measures, it should be noted that the general principle involved is the same. In an item analysis for a selection criterion test the variables of concern are individual difference variables, and these are permitted to vary while other variables, such as types and amounts of instruction and motivation, are held constant. This same principle applied to item analysis for instructional criterion tests has implied that instructional effects should be varied while other potentially confounding variables, such as aptitudes and motivation, are held constant. Now, there is no reason why the same principle cannot be extended to the development of criterion tests for studies of motivational variables, such as drive or stress. In this case, groups equated for individual difference variables and in-instructional treatments would be given different motivational treatments, and the items showing most sensitivity to these differential treatments would be selected for the criterion measure.

The universal criterion measure for selection, instructional, and motivational studies would be the original representative sample of items taken from the universe of criterion behaviors. This measure would have criterion content

validity and would be unbiased with respect to the effects of any independent variables. The different item analysis procedures that have been discussed here would each produce some subset of these items specifically sensitive to, and therefore economical for the study of, one of the specific classes of variables. It is an empirical question as to whether the different subsets would be, in fact, different, and if so, as to what effects the different items would have on the statistical evaluation of the research results and the decisions made as a consequence of the results.

In closing, it is noted that one of the major frustrations in instructional research to date appears to be that most experiments on alternative techniques of instruction have shown insignificant differences between techniques when compared on achievement and proficiency criteria, whereas striking differences have been found in lengths and costs of courses that yielded comparable degrees of achievement or proficiency. Much of this problem may lie in the neglect of the appropriate rationale in the selection or design of the achievement and proficiency criterion measures for these studies.