Bulletin of the Psychonomic Society
1984, 22 (6), 566-567

A simplified method of test construction from traditional methods of item analysis

ALLEN J. SCHUH
California State University, Hayward, California

Traditional methods of test construction have worked rather well for the development and validation of tests. But, in recent years, tests and test-construction procedures have been examined more carefully than at any other time. The author has noted that when one has available knowledge of item-difficulty levels and of correlations of items with the total test score and with the criterion, it is possible to refine tests even better and simpler than we have in the past. A classic textbook example set of data in Gulliksen (borrowed from Mollenkopf) is analyzed to demonstrate the usefulness of the simplified method. The simplified method can greatly shorten a test while maintaining the full range of item difficulties, maintaining test validity, and even possibly raising test reliability.

The research on item analysis and test construction that occurred up to 1950 defined extremely well the procedures that others were to use since then. The literature cited by two popular books (Guilford, 1954; Gulliksen, 1950) remains a definitive expression of what was known and unknown about test construction. Many professionals still find the material in these early works to be almost a current statement of where we are in preparing a refined test from a large sample of items presented earlier for tryout on an experimental basis. Early work by Horst (1934) and Flanagan (1936) described procedures that later were refined by Wherry and Gaylord (1946). The procedures that Wherry and Gaylord presented were quantitatively sophisticated and yet easy for the unsophisticated to use and appreciate; they remain the method of choice in item selection for test-construction professionals.

Undergraduate students in courses in educational and psychological measurement frequently have difficulty in understanding advanced issues, even though the fundamentals have been covered, especially when the fundamentals form a basis of so much other material the students need to learn. Indeed, the fundamentals of test construction form the model for all evaluation in education and psychology, including employee screening, employment interviewing, training evaluation, performance appraisal, and even program evaluation.

The author reexamined much of the early literature and found a data set in Gulliksen's (1950) *The Theory of Mental Tests* that had everything necessary as a starting point to discuss test construction and validation, including peripheral issues such as suppressor variables (Gulliksen's table has one item with a low negative item-criterion correlation). The author believes that the simplified method of test construction presented in the present paper is useful for explaining such material to undergraduates who will need to comprehend test construction prior to advanced training and eventual professional employment. Of course, the Horst, Flanagan, and Wherry-Gaylord methods may still be the choice of professionals.

Requests for reprints should be sent to Allen J. Schuh, Department of Management Sciences, School of Business and Economics, California State University, Hayward, CA 94542.

METHOD

The table on page 395 of Gulliksen (1950) forms the starting point for this analysis. The data were provided to Gulliksen by W. G. Mollenkopf of the Educational Testing Service. The data for 35 items include the item number, proportion of subjects answering the item correctly (referred to in this article as the item-difficulty level), the point biserial correlation of the item with the total test score (referred to in this article as r_{it}), the point biserial correlation of the item with a criterion score (referred to in this article as r_{ic}), and three additional bits of information that are not necessary for the analysis reported here (the standard deviation of each item, a reliability index, and a validity index for each item).

Traditional methods of reducing a large pool of experimental items to a smaller test of known size would consist of using the item difficulty levels (all items must have some variance), a correlation of the item with the total test score (the correlation, which should be positive but not too high, will change, of course, as some items are dropped from the test, thus changing the total test score), and the correlation of the item with a criterion measure (all such correlations should be positive). The traditional methods also need the intercorrelation of the items, but these were not shown in Gulliksen's example.

The traditional methods do not do three things that the author believes are necessary to build a better test and to stress certain points for students:

(1) The sample size for the experimental group must be known so that only items correlating significantly with

the total test score are retained. A significant correlation would indicate that the item belongs in that test.

(2) The correlations of the items with the criterion must also be tested for significance. No item should be kept that does not measure significantly what it should measure.

(3) The frequency distribution of the item difficulties must be known so that the full range of item difficulty is represented by the final test. Items will be selected in equal numbers from intervals across the full range. Guilford (1954) suggested that test reliability would be better with such a procedure than it would be if all items had exactly a .5 item-difficulty level. That is, the mean item difficulty should be .5, but the full range of values should be present to increase the standard deviation of the test scores.

PROCEDURE

The algorithm consists of the following steps:

(1) Calculate and list by item number the item-difficulty level, the correlation of the item with the total test score (r_{it}), and the correlation of the item with the criterion (r_{ic}). In the example here, we will use Gulliksen's (1950) table on page 395 of his text because the text is a classic well known to professionals in measurement.

(2) Rank within class intervals ($>.85$, $>.75$, $>.65$, $>.55$, $>.45$, $>.35$, $>.25$, $>.15$, $>.01$) the item difficulties identified by original item number with their respective r_{it}s and r_{ic}s.

(3) Find the lowest frequency by class interval and prepare to select that number of items from each class interval. With Gulliksen's (1950) table, the lowest class interval was $>.15$ at 2. Thus, to maintain a rectangular item-difficulty distribution, two items will be taken from each other class interval as well. Because the lowest item frequency within the intervals was two within all class intervals, a budget of only two items will be selected by the procedure of eliminating items that do not correlate significantly with both the total test score and the criterion.

(4) Of the remaining items, select those with the highest r_{ic} Gulliksen (1950) did not describe his sample size. The author arbitrarily assumed it was 100, and the correlation of .166 or better would be needed for one-tailed statistical significance at the conventional .05 level. For Gulliksen's data, there were four items in the $>.75$ interval; Item 2 was excluded for a nonsignificant correlation with the criterion. For the three remaining items (Items 4, 1, and 8), Items 4 and 1 were selected for having the highest r_{ic}.

Consistently applying this algorithm reduced the test from Gulliksen's (1950) table to Items 1, 4, 5, 7, 11, 12, 13, 14, 18, 19, 20, 21, 23, 25, 27, and 29. The average r_{ic} was .209 for the original item test. Gulliksen (1950, p. 467) deleted five items (Items 2, 5, 22, 31, and 32), for an average r_{ic} of .222; the simplified method presented here raised the average r_{ic} still further to .254. The nine highest r_{ic}s were the same by Gulliksen's method and the simplified method. The item difficulties for these nine vary considerably, and therefore they probably do not have high intercorrelations and probably make at least some unique contribution to common variance with the criterion.

The resulting 16-item test is at least shorter than Gulliksen's (1950) 30-item test, is more refined statistically with the significance level check, and is probably more reliable because of the item-difficulty rectangular distribution. If it were possible to calculate new r_{it}s, the ratio of the r_{ic} to r_{it} would indicate the new validity for the shorter test (Guilford, 1954).

The new method is considerably easier to grasp by undergraduates, easier to present by professors, and yields sound results psychometrically. It is hoped that professionals in other settings will also try the simplified procedure and use it not just to work with textbook examples but also to work with real data being evaluated in practical educational and psychological settings. A better understanding of good test construction and better, shorter tests should result.

REFERENCES

FLANAGAN, J. C. (1936). A short method for selecting the best combination of test items for a particular purpose. *Psychological Bulletin*, **33**, 603-604.

GUILFORD, J. P. (1954). *Psychometric methods*. New York: McGraw-Hill.

GULLIKSEN, H. (1950). *Theory of mental tests*. New York: Wiley.

HORST, A. P. (1934). Item analysis by the method of successive residuals. *Journal of Experimental Education*, **2**, 254-263.

WHERRY, R. J., & GAYLORD, R. H. (1946). Test selection with integral gross score weights. *Psychometrika*, **11**, 173-183.

(Manuscript received for publication August 13, 1984.)