DOCUMENT RESUME

ED 041 940 TM 000 032

AUTHOR	Garvin, Alfred D.; Ralston, Nancy C.
TITLE	Improving the Reliability of Course Pretests.
PUB DATE	Mar 70
NOTE	12p.; Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, Minn., March 1970
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.70
DESCRIPTORS	College Students, Multiple Choice Tests, *Objective Tests, *Pretests, *Student Testing, *Testing Problems, *Test Reliability

ABSTRACT

Confidence Weighting (CW), after Ebel, and Multiple Responding (MR), after Coombs, are compared empirically to determine which improved test reliability more in the case of a course pretest derived from the final examination. It was hypothesized that MR, which purportedly measures partial knowledge, would be more effective than CW, which measures extra knowledge. The subjects were 58 students enrolled in a graduate course in Adolescent Psychology. The pretest comprised 20 four-choice items drawn from the final exam. One randomly selected half of the class took the pretest under CW instructions; the other, under MR instructions modified to yield a "raw" score also. The reliability of raw scores was zero in both groups. MR-score reliability improved to .38 while CW-score reliability remained zero. (Author)

IMPROVING THE RELIABILITY OF COURSE PRETESTS [1]

Alfred D. Garvin and Nancy C. Ralston

University of Cincinnati

The potential uses of the course pretest range from a preview of future learnings to a review of past learnings. More specifically, pretests are commonly used for purposes as disparate as the following:

1.) Providing the class with an explicit preview of the kinds of questions they should be able to answer by the end of the course.

2.) Disclosing any points in the prospective syllabus that most of the class have, somehow, already learned.

3.) Disclosing any points in the prospective syllabus regarding which many of the class hold the same misconception.

4.) Providing the instructor with an estimate of the distribution of prerequisite learnings within the class.

The first three of these purposes require that the items comprising the pretest be representative of previous final examinations for that course. The last purpose requires that such a test be usefully reliable. In the rare case where the course involved introduces some wholly unfamiliar content area, it is practically impossible to accomplish both the preview and review purposes of pretesting through a single test. However, in the more common case of a second or subsequent course in any more or less cumulative and coherent content area, e.g., the behavioral or social sciences, it is reasonable to assume that at least some of the students beginning any such course would have at least some relevant information (and some misinformation) that could be assessed through items drawn from previous final examinations. Further, since prerequisite learnings would be manifested in such an assessment, a single pretesting session--which, as a practical matter, is all there is ever time for--could accomplish all of the purposes of pretesting if this test was usefully reliable.

We will assume here that the final examinations from which a pretest might be drawn were objective tests. The objective format is used for final examinations primarily because of its convenience in scoring, not because it is inherently reliable. Indeed, the reliability of such a test is commonly quite low even when given in its full length and after intensive instruction on its content. Thus, the general problem here is how best to attain a useful level of reliability in a shortened version of such a test given before instruction on its content has even begun.

There are two well-known special testing procedures designed to increase the reliability of objective tests. Each has been found to "work" in certain instances but they have rarely been compared with each other in a particular instance. The purpose of the study reported here was to determine empirically which of these two promising special testing procedures worked better in the specific instance of the course pretest.

BACKGROUND

The most perennial of the special testing procedures intended to improve the reliability of objective tests is generically termed Confidence Weighting (CW). Under CW procedure, a respondent first selects the one best alternative of those offered for a given item as he would under conventional or Rights-Only (RO) scoring procedure. Then, if he is sufficiently confident that his selection is right, he has the option, exercised independently on each item, of indicating such confidence by a special response symbol under the contingencies of a specified point bonus if he is, indeed, right against a specified penalty if he is wrong. If he does not elect the CW option for a given item, the RO contingencies of one point if right against zero if wrong apply.

Studies on the efficacy of CW procedure have been reported since the middle thirties (Hevner, 1932; Soderquist, 1936; Swineford, 1939). Oddly enough, the earliest studies embodied the most elaborate procedures. Typically, four levels of confidence were offered with an appropriate range of bonuses and penalties. More recently, Ebel (1965a, 1965b) and Garvin (1969) have concentrated on a much simpler form of CW involving only two confidence levels--some or none. Under this scheme, the score contingencies are +2 if right *vs.* -2 if wrong where a response is "weighted," *i.e.*, coded as confident, and +1 if right *vs.* 0 if wrong if not weighted.

CW procedure yields two score distributions: an RO distribution and a CW distribution reflecting bonuses and penalties. The reliability of each distribution may be estimated by any of the standard reliability algorithms.

These may be compared through a re-arrangement of the Spearman-Brown Prophecy Formula to yield an index termed Improvement Factor (IF) by Ebel (1965), indicating the factor by which a given test, administered under RO procedure, would have to be lengthened to attain the reliability it displayed under CW procedure. The IF's reported in the studies cited above range from almost no improvement to about 85 percent improvement.

The second special testing procedure designed to improve objective test reliability may be called Multiple Responding (MR). Under MR procedure, the respondent eliminates as many wrong alternatives as his knowledge (and confidence therein) permits. He receives one point for each appropriate elimination up to a maximum item score of A-1 points, where A is the number of alternatives; he loses A-1 points if he eliminates the right alternative. Thus, item scores may range from +(A-1) to -(A-1). Ordinarily, MR procedure does not yield a concurrent RO score; however, it can do so if a respondent is required to indicate additionally what his first choice would have been under RO procedure.

The definitive study on the efficacy of MR procedure is that by Coombs, Milholland, and Womer (1956). The IF's attained in that study were modest, averaging about 20 percent. Only one study has attempted to compare the relative efficacy of CW and MR (Dressel and Schmidt, 1953). The testing situation was a patently contrived, non-academic one and the results were equivocal. CW and MR have never been compared in the specific instance of the course pretest.

THEORY

There are two sources of error inherent in objective tests administered under RO scoring procedure. The most obvious of these is chance error, i.e., error due to guessing. This is also the least reducible for any given item format. Gulliksen (1950) has concluded that in a power test situation--the typical classroom testing situation--the commonly used corrections-for-guessing are wholly ineffectual. The other source of error may be termed "truncation error." Holding chance error at zero, we may posit for any given item a level of relevant knowledge that is "just enough" to enable the respondent to isolate the right alternative. Under RO procedure, partial knowledge less than this level is truncated down to zero, indicating no knowledge at all. Extra knowledge beyond this level is also truncated down to a score of one point, indicating "just enough" knowledge. Chance error is always positive; truncation error is always negative. The resultant of these two components of error yields the more or less symmetrical "error of measurement" assumed in classical measurement theory.

MR procedure is presumed to "work" by recovering partial knowledge otherwise lost to truncation, thereby reducing that component of error. Indeed, the title of the article by Coombs, et al., cited above, is "The Assessment of Partial Knowledge." As the operational complement of MR, CW may be thought to work by recovering extra knowledge otherwise lost to truncation. Thus, MR ought to work better in those testing situations characterized by a preponderance of partial knowledge while CW ought to work better in those testing situations characterized by a preponderance of extra knowledge. A pretest comprising items drawn from previous final examinations for that course is clearly a partial knowledge situation. With respect to such a test, we might expect the levels of relevant knowledge in the class to range from being partially informed, through uninformed, to misinformed.

HYPOTHESIS

In accordance with the theory outlined above, it was hypothesised that the reliability of such pretests would be increased more by the MR procedure than by the CW procedure.

PROCEDURE

The subjects were 58 typical students enrolled in the junior author's graduate-level course in Adolescent Psychology. About 80 percent of these were Education majors; most of the rest were Psychology majors. All Ss had completed at least one course in general psychology, 63 percent had teaching experience, and 25 percent had children of their own.

The pretest used comprised 20 four-choice items selected from a previous final course examination so as to be minimally dependent upon specific course content, whether given by lecture or learned from the text.

At the beginning of the first class session it was announced that a course pretest would be given to ". . . let you know what the course is to be about . . ." and to ". . . indicate to [the instructor] how much each student already knows about the course content . . ." as a basis for making differential assignments, if indicated. It was explained that while this test wouldn't "count" in determining final grades, it was in each student's interest to do his best on this pretest. Next, it was explained that there are two different ways to try to improve the reliability of such a test, i.e., ". . . to get more information out of it . . ." and that half of the class would use each way ". . . to see which method works better in cases like this." The response options and score contingencies of the Ebel-type CW procedure and those of the Coombs-type MR procedure were explained and the consequences of various response strategies in each procedure were illustrated. Finally, the class was randomly partitioned into two subgroups of equal size. One subgroup was selected arbitrarily and was told they were to use CW procedure; the other was told they were to use MR procedure.

The pretest was then distributed, together with special answer sheets appropriate to the special response procedure to be used in each group. For the CW group, the best alternative was to be recorded as in RO procedure; if S wished to "weight" a given response, he simply circled the alternative symbol selected. The answer sheet for the MR group permitted S to record from one to three "eliminations" per item. In addition, the MR S was to record for each item what his first choice would have been under RO procedure. The pretest was then conducted under power-test time limits.

RESULTS

The response patterns in each group indicated that all Ss understood their respective special response instructions and exercised their options appropriately. Several *Ss* from each group requested that future tests be conducted "this way." No one from either group voiced a preference for "the other" procedure.

- - - - - - - - - - - - - - - - - - - -

Insert Table 1 about here

- - - - - - - - - - - - - - - - - - - -

On the basis of RO scores, the test proved moderately difficult, with little score dispersion; $\bar{X}$ was about 14 (out of 20 four-choice items), with an SD of about 2, in each group. It was also completely unreliable; the K-R (20) reliability of RO scores was < 0 in each group. On the basis of special-procedure scores, the K-R (20) *r* in the CW group was still < 0; in the MR group, it was +.38. All *rs* were recomputed by the odd-even split-halves method and the results were substantially the same.

DISCUSSION

The RO data in Table 1 indicate that the two groups were substantially equivalent in their distribution of relevant ability. The reactions of the *Ss* during and after the test indicated that the two procedures were equally acceptable to them, in a psychological sense. The vulnerability of the K-R (20) *r* to depression by factorial heterogeneity in relatively short tests was "covered" by corroborating data from split-halves *rs*. Thus, the results observed would seem to be relatively free of the more obvious methodological artifacts.

The experimental hypothesis is clearly supported by these data and the theory from which it was derived is seen as a credible one. Of course, this same theory suggests that the reliability of a mastery-type final course examination, where a preponderance of extra knowledge might be expected, would be improved more by CW- than by MR procedure. This much is clear: future research on CW, MR, or on comparisons of these procedures should take account of the extra- vs. partial-knowledge factor in their designs.

It must be conceded that a test reliability of .38 is hardly impressive, *per se*; nevertheless, in the Spearman-Brown sense, it represents an infinite improvement over an *r* of zero. The utility of this observation lies in the fact that relatively short pretests, drawn from previous final examinations, are, for reasons given earlier, about the only practical kind--and these are very likely to yield RO reliabilities of very close to zero. Whatever is to improve the reliability of pretests must improve it in this kind of pretest.

REFERENCES

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, Spring, 1956, 16, 13-37.

Dressel, P. L., & Schmidt, J. Some modifications of the multiple-choice item. Educational and Psychological Measurement, Winter, 1953, 13, 574-595.

Ebel, R. L. Confidence weighting and test reliability. Journal of Educational Measurement, 1965a, 2, 49-57.

Ebel, R. L. Measuring educational achievement. Englewood Cliffs, New Jersey: Prentice-Hall, 1965b.

Garvin, A. D. The effect of confidence weighting on variation of the error of measurement. (Doctoral dissertation, University of Maryland) Ann Arbor, Mich.: University Microfilms, 1969. No. 69-7621.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Hevner, K. A method of correcting for guessing in true-false tests and empirical evidence in support of it. Journal of Social Psychology, 1932, 3, 359-362.

Soderquist, H. O. A new method of weighting scores in a true-false test. Journal of Educational Research, 1936, 30, 290-292.

Swineford, F. The measurement of a personality trait. Journal of Educational Psychology, 1939, 29, 289-292.

FOOTNOTE

[1] Paper (to be) presented at the 1970 NCME Annual Meeting in Minneapolis.

TABLE 1

Test Statistics by Scoring Procedure and Group

	RO Scores				Special Scores			
Group	$\bar{X}$	SD	$\underline{r}_{(20)}$[a]	$\underline{r}_{s-h}$[b]	$\bar{X}$	SD	$\underline{r}_{(20)}$[a]	$\underline{r}_{s-h}$[b]
CW	14.2	1.64	<0	<0	18.4	4.05	<0	<0
MR	13.5	2.09	<0	<0	33.9	8.30	.38	.32

[a] K-R (20) reliability coefficient.

[b] Split-halves reliability coefficient.