DOCUMENT RESUME

ED 075 498	TM 002 600
AUTHOR	Schmeiser, Cynthia Board; Whitney, Douglas R.
TITLE	The Effect of Selected Poor Item-Writing Practices on Test Difficulty, Reliability and Validity: A Replication.
PUB DATE	73
NOTE	15p.; Paper presented at American Educational Research Association Meeting (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Item Analysis; *Multiple Choice Tests; Technical Reports; *Test Construction; Test Reliability; Test Results; *Test Validity

ABSTRACT

Violations of four selected principles of writing multiple-choice items were introduced into an undergraduate religion course mid-term examination. Three of the flaws significantly increased test difficulty. KR-sub-20 values were lower for all of the tests containing the flawed items than for the "good" versions of the items but significantly so in only one of four comparisons. The reductions in reliability were equivalent to those expected to result from shortening the "good" test by 28 to 71 percent. Concurrent validity (correlation of experimental test scores with the midterm test of similar content) was lower in all four cases, but significantly so in only one of four cases. The reductions in validity were equivalent to those expected to result from shortening the test by 47 to 77 percent. (Author)

THE EFFECT OF SELECTED POOR ITEM-WRITING PRACTICES ON TEST DIFFICULTY, RELIABILITY AND VALIDITY: A REPLICATION

Cynthia Board Schmeiser and Douglas R. Whitney

University of Iowa

ABSTRACT

THE EFFECT OF SELECTED POOR ITEM-WRITING PRACTICES ON TEST DIFFICULTY, RELIABILITY AND VALIDITY: A REPLICATION

Violations of four selected principles of writing multiple-choice items were introduced into an undergraduate religion course mid-term examination. Three of the flaws significantly increased test difficulty. KR_{20} values were lower for all of the tests containing the flawed items than for the "good" versions of the items but significantly so in only one of four comparisons. The reductions in reliability were equivalent to those expected to result from shortening the "good" test by 28 to 71 percent. Concurrent validity (correlation of experimental test scores with the midterm test of similar content) was lower in all four cases, but significantly so in only one of four cases. The reductions in validity were equivalent to those expected to result from shortening the test by 47 to 77 percent.

THE EFFECT OF SELECTED POOR ITEM-WRITING PRACTICES ON TEST DIFFICULTY, RELIABILITY AND VALIDITY: A REPLICATION

In a previous study (Board & Whitney, 1972), the effects of selected poor item-writing practices on an undergraduate political science test were studied. The study suggested that certain of the flaws do influence student test scores. The major purpose of this study was to reexamine, in a new content area, the effect of the same poor item-writing practices on test difficulty, reliability and validity, in order to evaluate previous conclusions.

METHOD

Item-Writing Flaws Studied

The same four poor item-writing practices used in the previous study were selected for the replication. These four flaws were:

1. Items whose stems included "window dressing" or material not necessary to answer the item.
2. Items with incomplete stems which did not express either a complete statement or question.
3. Items whose keyed responses were noticeably longer or shorter than the distractors.
4. Items whose keyed responses were the only grammatically consistent answers.

The first two flaws were chosen to represent stem-related flaws which were thought to make the items more difficult either because the question is obscured by the presence of unnecessary material (flaw 1) or because the stem contains no specific question (flaw 2). The last two flaws were chosen to represent the kinds of clues a test-wise student would presumably use to achieve higher scores on a test. Items incorporating flaws (3) and (4) should be easier than corresponding "good" versions of the same items.

Instruments

The "raw material" from which the five tests were fashioned was previous examination items for the undergraduate course Religion in Human Culture at the University of Iowa. Twenty questions were chosen for the study chiefly on the basis of their difficulty (40-70%) and discrimination indices (.3 or better) and their adaptability for the kind of item flaws selected. These 20 items were reviewed and modified by the authors and used as "well-written" items (Test I) for this study.

The stems for the same 20 items used in Test I were then systematically modified to include "window dressing" and designated as Test II. Test III contained the same 20 items as Test I except that the stems were truncated to be grossly incomplete. Changes were made only in the stems for the items in Tests II and III.

Tests IV and V consisted of the same 20 items as in Test I with only the distractors modified. For Test IV the distractors were made systematically longer or shorter than the keyed response. In Test V the distractors were made grammatically inconsistent with the stems in all 20 questions.

The second midterm in the course consisted of similar content used in the items of the experimental tests. This test consisted of 100 multiple-choice items covering course material for the middle part of the course and was administered one week after the experimental tests. The KR_{20} coefficient for the second midterm was .90. This test was used to evaluate validity for the flawed tests.

Administration

The five tests were administered to 501 students during a regular 50-minute class meeting in April, 1972. The five experimental tests were distributed systematically within each examination group so that every sixth person took the same experimental test.

Students were instructed to answer every question and informed they would have 45 minutes to complete the examination. After the students completed the examination, they were given written material which described the study, explained its purpose, and assured them that the scores on the experimental test would not be included in their course average. Out of the 501 students who took the experimental tests, 35 did not fill out the proper identification needed for the analysis by course achievement and were dropped from this portion of the study. In order to obtain proportionality of cell frequencies for the analysis, it was necessary to eliminate randomly 11 additional subjects.

Analysis

The initial analysis was a treatment X levels ANOVA (Lindquist, 1953). In this analysis the A effect was that of "Item Flaw" and the B (or levels) effect was that of "Achievement Level". In the analysis, A had 5 categories (1 = Test I, 2 = Test II, 3 = Test III, 4 = Test IV, 5 = Test V) and B had 5 levels (i.e., 1 = lowest 5th, etc.).

In order to define achievement levels for this analysis, the unweighted sum of points earned on the two-100 item midterm examinations, one 100-item final examination and a discussion score (range 0-100) were used. The composite score which consisted of these sums added together was used as a blocking variable in the analyses. Since these scores were not available prior to the administration of the experimental tests, *a priori* stratified sampling was not possible. Thus, the differences in achievement, were effectively randomized across test forms.

To test for the effects of each flaw, 4 pre-specified contrasts (Test I vs. each of the poor tests) were conducted. Dunn's (1961) procedure for limiting the overall probability of Type I error ($p < .05$) was used to determine critical values used in the tests of significance for contrasts.

RESULTS AND DISCUSSION

The results of the ANOVA treatment X levels design are shown in Table 1. The significant A (Tests) effect indicated that there was a difference in difficulty among the tests. The significant B (Achievement) effect indicates that, in general, better students answered correctly, a higher proportion of even poorly written items than did poorer students. The absence of an interaction effect (AxB) indicates that the flaws did not operate differentially across achievement levels.

Insert Table 1 about here

Difficulty: Results

According to expectations, the test consisting of items including "window dressing" was appreciably more difficult than the good test. That is, the t-statistic ($t = 3.07$) was significant. This result suggests that items which incorporate window dressing do make the test more difficult for students in general.

As expected, the items having incomplete stems were more difficult than the same items on Test I ($t = 3.01$). This test flaw seems to depress the test scores of students in general.

The expected effect of foil length was to make the test less difficult since the "different" length was always the keyed response. Surprisingly, the effect of foil length made the test significantly ($t = 3.06$) *more* difficult than the "good test".

The grammatical inconsistency flaw did not substantially alter the test's difficulty (t = 0.39). This result is in contradiction with the expectation that this flaw would make the test less difficult.

Difficulty: Discussion

This replication suggested that window dressing does make items more difficult. The previous study found no such effect. There was no interaction between tests and achievement in either study of this flaw.

A previous finding (Board and Whitney, 1972) that incomplete stems make items more difficult was confirmed in this replication. No interaction of this flaw with achievement level was found in either study.

These findings concerning the effect of distractor length were somewhat at odds with previous findings. There was a significant A effect in this study, but no interaction. In the earlier study, there was no significant A effect but there was a significant interaction effect. Dunn and Goldstein (1959) found that the distractor length flaw made the test easier, as was found in the authors' first study but not in this replication.

Previous results indicated that the grammatical inconsistency flaw had little effect on test difficulty. Neither study yielded an interaction effect with achievement. This result is in contrast with the findings of Dunn and Goldstein (1959) that the grammatical inconsistency flaw yielded higher mean test scores than corresponding "good" items.

These contrasting results between the previous study and this replication concerning the item flaw length need to be viewed in light of the following consideration. The length flaw as used in our studies has differing operational definitions making it difficult to compare the results in the two studies. The length flaw can be constructed in one of two ways: by adding information to make the keyed response longer, or by shortening the keyed response. The intended effect of the flaw was to make the item less difficult. By adding information, it is possible that the extraneous information could have made the foil more ambiguous and more difficult, thereby possibly paralleling the effect of window dressing in this replication. Deleting information in the keyed response could have eliminated important elements which would have also made the item more difficult. Unfortunately, the flawed items were not constructed using only one of these methods which would have provided a comparable basis for the results. In Dunn and Goldstein's (1959) study, the length flaw was systematically modified making the keyed answer longer. It is likely, therefore, that our ambiguous definition does not allow a reasonable comparison to Dunn and Goldstein's results.

Reliability and Validity: Results

To assess the reliability of the tests, KR_{20} coefficients were computed for each test (I through V). The resulting internal consistency values were tested for population equality by a procedure described by Feldt (1969). The results are given in Table 2.

Insert Table 2 about here

As a measure of validity of each experimental test, product-moment correlations were computed between scores on the experimental tests and the second midterm described above. Tests of hypotheses that the observed validity coefficients arose from populations with a common parameter were conducted for each of the four flaws and the results are shown in Table 2.

Window dressing resulted in a lower internal consistency than that of the "good" test, but not significantly so. The validity of the test was also reduced, but, again, not significantly.

Incomplete stems resulted in lower internal consistency than that of the "good" test, but again, the difference was not significant. The validity of the test was also not significantly changed.

The effect of distractor length significantly reduced the internal consistency of the test. The correlation between the "poor" test and the "good" test and the midterm scores was significantly lower for this flaw.

The effect of grammatical inconsistency was to lower the internal consistency, but not significantly. The validity coefficient was also lower but the difference was not significant.

Reliability and Validity: Discussion

In a similar study, Dunn and Goldstein (1959, p. 177) concluded that "It cannot be said that the Kuder-Richardson reliabilities are differentially influenced by the rules studied." In accordance with these results, McMorris, et. al., (1972, p. 287) found that "the insertion of cue, grammar, and length faults did not systematically or appreciably effect either validity or reliability coefficients." In contrast to their results, three out of eight internal consistency coefficients in our two studies were significantly lower than the "good" forms of the tests, but all eight internal consistency coefficients were reduced by the inclusion of these flaws. If the probability of the smaller alpha coefficients for the flawed tests were hypothesized as .50, the series of eight of eight reduced coefficients has a probability $p < .01$.

In the McMorris study, the internal consistency coefficient for the grammar flaw was lower than that of the good test, as it was in this study. The coefficient for the length flaw was essentially the same as the coefficient of the good test. The internal consistency coefficients in Dunn and Goldstein's study were higher in seven out of eight cases for the grammar flaw, and higher in four out of eight cases for the length flaw.

In Dunn and Goldstein's study, they found that "an inspection of the set of validities revealed no pattern of validities with respect to item construction rules." Similarly, McMorris found that there was no appreciable effect upon the validity of the tests. In our two studies, three out of eight validity coefficients were significantly lower, with eight of eight coefficients reduced by the inclusion of the four item writing flaws. Again, this is a highly unlikely result if the flaws, in fact, had no effect on validity.

McMorris, et. al., found that both of the grammar and length flaws increased the validity of the test .08 and .03 respectively. In Dunn and Goldstein's study, the grammar flaw increased the validity of the test in six out of eight cases and the length flaw increased the validity of the test in four out of eight cases.

The results of Dunn and Goldstein and those of McMorris clearly contradict those of the authors. The comparison of the validity coefficients is difficult and must be done cautiously, for, it must be considered that different criterion measures were used in the studies. In our studies, parallel content in classroom examinations were used as a criterion measure as contrasted with a statewide examination used in McMorris' study and a standardized examination used in the Dunn and Goldstein study.

SUMMARY

Prior to considering the implications of this study, further attention should be directed to the homogeneity of the good test in the previous study as opposed to that of the test used in this replication. In the previous study, the good test had an average internal consistency coefficient of .73. That is, the items on this examination formed a relatively homogeneous test. However, in this replication the KR_{20} value for the good test was only .59. Thus the items formed a less homogeneous test. It is very possible that potential effects were obscured by the less homogeneous test. Within this limitation, however, the obtained results warrant the following conclusions:

1. The effect of window dressing, or extraneous material in the stems of the items, is not clear due to the conflicting results in the previous study and this replication. There was an indication in this replication that this flaw makes the test more difficult.
2. Incomplete stems make the test items more difficult for most students.
3. The difference in the length of the distractors and the keyed response has had conflicting results in these two studies. In our first study, length tended to make the test easier for the students. In this replication, differential response length made the items more difficult. Further study is needed to assess the effect of this flaw. Presence of the length flaw does appear to reduce the internal consistency and the validity of the test.
4. Grammatical inconsistency between the stem and the keyed response does not have a major effect on test difficulty.

A final comment should be made concerning the methodology used in researching item writing principles. In the studies of McMorris, Dunn and Goldstein, and our earlier study, the flawed tests were constructed in two ways: either the good items were modified to incorporate the flaws or both the good and flawed tests were constructed at the same time. Another explanation for the inconsistent results in these studies besides those previously noted might be that when flawed items are constructed from good forms of the same items, any effects could be idiosyncratic to the studies and to the particular methods used to create the "flaws".

If one assumes that classroom examinations are constructed by writing items then reviewing these items to eliminate obvious item flaws, a different methodology is suggested. It seems more appropriate to pursue this research by locating teacher-constructed tests with specific flaws and "improving" them to eliminate the flaws. This methodological procedure is the approach future research ought to take and may serve to clarify seemingly contradictory results among the studies cited.

REFERENCES

Board, C. & Whitney, Douglas R. The effect of selected poor item writing practices on test difficulty, reliability and validity. *Journal of Educational Measurement*, 1972, *9*, 225-233.

Dunn, O. J. Confidence intervals for the means of dependent, normally distributed variables. *Journal of the American Statistical Association*, 1959, *54*, 613-621

Dunn, T. F. & Goldstein, L. G. Test difficulty, validity, and reliability as functions of selected multiple-choice item construction principles. *Educational and Psychological Measurement*, 1959, *19*, 171-179.

Feldt, L. S. A test of the hypothesis that Cronbach's alpha or Kuder-Richardson coefficient twenty is the same for two tests. *Psychometrika*, 1969, *34*, 363-373.

Gulliksen, H. *Theory of Mental Tests*. New York: Wiley, 1950.

Lindquist, E. F. *Design and Analysis of Experiments*. Boston: Houghton Mifflin, 1953.

McMorris, R. F., Brown, J. A., Snyder, G. W. & Pruzek, R. M. Effects of violating item construction principles. *Journal of Educational Measurement*, 1972, *9*, 287-295.

TABLE 1

Effect of Item Flaws: Cell Means and ANOVA Summary Table

Means

Achievement Levels

	Lowest 5th	Second 5th	Third 5th	Fourth 5th	Highest 5th	Overall Mean
I (Good Test) (n = 85)	7.88	10.00	10.18	12.29	13.88	10.85
II (Window Dressing) (n = 95)	7.47	7.74	11.16	10.58	12.00	9.79
III (Incom. Stems) (n = 90)	8.39	8.61	9.17	10.50	12.22	9.78
IV (Length) (n = 100)	8.75	8.60	9.40	10.50	11.80	9.81
V (Gram. Inc.) (n = 85)	8.35	9.59	11.18	11.94	12.47	10.71
Overall Mean	8.18	8.87	10.20	11.12	12.44	10.16

ANOVA Summary Table

Source	df	MS	F	Extreme Areas
A (Tests)	4	25.97	4.79	$p < .001$
B (Achievement)	4	266.79	49.22	$p < .0005$
A x B	16	8.83	1.63	$p < .10$
W. Cells	430	5.42		

TABLE 2

Internal Consistency and Validity Coefficients
for Items With and Without Flaws

Test	N	Internal Consistency			Validity		
		KR_{20}	W^a	k^b	r^c	z^d	k
I (Good)	89	.59			.69		
II (Window Dressing)	97	.51	1.19	28%	.57	1.33	54%
III (Incomplete Stem)	86	.50	1.22	31%	.56	1.47	56%
IV (Distractor Length)	101	.30	1.71**	71%	.45	2.53*	77%
V (Grammatical Consistency)	89	.49	1.24	34%	.59	1.13	47%

*p <.05; **p < .01

[a] $W = (1-r_j) / (1-r_1)$ is approximately distributed as F with df $N_1 - 1$ and $N_2 - 1$. (Feldt, 1969)

[b] k is the factor by which the good test could be shortened and still be as reliable and/or valid as the poor test (Gulliksen, 1950, p. 83 & p. 93).

[c] r is the product-moment correlation between the scores on the portion of the experimental test and scores from a similar, but longer, midterm course examination.

[d] $z = (z_1 - z_j) / \sqrt{1/N_1 - 3) + 1/(N_j - 3)}$ here z_1 and z_2 are the Fisher z transformations of the validity coefficients for the two sub-sets of test items.