DOCUMENT RESUME

ED 104 915 TM 004 370

AUTHOR Andrulis, Richard S.; And Others
TITLE The Effects of Repeaters on Test Equating.
INSTITUTION American Coll. of Life Underwriters, Bryn Mawr, Pa.
PUB DATE [74]
NOTE 17p.

EDRS PRICE MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS Cutting Scores; *Equated Scores; *Recall (Psychological); Retention; Scoring; Statistical Analysis; Test Bias; *Testing; *Testing Problems; *Test Reliability
IDENTIFIERS Test Repeaters

ABSTRACT

The purpose of this investigation was to establish the effects of repeaters on test equating. Since consideration was not given to repeaters in test equating, such as in the derivation of equations by Angoff (1971), the hypothetical effect needed to be established. A case study was examined which showed results on a test as expected; overall mean was lower for repeaters. Applying these data to the available equating equations, it was shown that an additional 3 percent of the examinees was categorized as having "passed" than should if repeaters were taken into account. The practical solution offered is to hold separate the score of repeaters, execute the equating on the others, and then apply the conversion to all the examinees. (Author)

THE EFFECTS OF REPEATERS ON TEST EQUATING

Research Report

THE AMERICAN COLLEGE

THE EFFECTS OF REPEATERS ON TEST EQUATING

Richard S. Andrulis — American College

Larry M. Starr — University of Windsor

Lawrence M. Furst — Villanova University

Introduction

Test fairness implies that successive forms of a test are equivalent in all important respects. "However, since the forms cannot be precisely equivalent, it becomes necessary to equate the forms -- to convert the system of units of one form to the system of units of the other -- so that scores derived from the two forms after conversion will be directly equivalent."[1]

A classification of the various methods for equating test forms, particularly considering the kind of group used and test reliability, is presented by Angoff (1971). In this general presentation, while the equating methods are categorized by either assuming random groups or nonrandom groups (e.g., groups widely different in ability), no attention is given to the circumstance where individuals in a group are taking the test for a second time.

This paper is concerned with the effects of repeaters -- those who take a test for a second time -- on the conversion scores that result after test equating. These effects are examined, in particular, for the design assuming two equally reliable tests are administered to random groups with each test including a set of common equating items (Cureton and Tukey, 1951; Levine, 1955; Angoff, 1961; Lennon, 1964; Angoff, 1971, pp. 576 - 579).

Method

A practical method of equating and calibrating test scores involves the use of those items common to both forms of the tests (Angoff, 1971). This score, U, reflects an individual's performance on those items common to Form X (administered to Group α) and to Form Y (administered to Group β).

Lord (1955) has developed equations in which he makes maximum likelihood estimates of the population means and variances on X and Y. By definition, linear equating states that scores on two tests are equivalent if they correspond to equal standard-score deviates,

(1) $$\frac{Y - My}{Sy} = \frac{X - Mx}{Sx}$$

When the terms are appropriately rearranged, equation 1 takes the Form $Y = Ax + B$ where $A = Sy/Sx$ and $B = My - AMx$, A being the slope of the conversion line, and B the intercept (the point on the Y axis where it is intersected by the conversion line).

The test equations presented in this paper correspond to the type of conversion expressed in the form of a straight line. That is, it is reasonable to assume that, by definition, successive forms of a test are constructed to be nearly equivalent in all the important respects and the conversion of X scores to Y scores can be accomplished simply by changing the origin and unit of measurement.

The equations appropriate to a random administration of X and Y with U administered to all examinees are as follows:

(2) $$\hat{\mu}_x = Mx_\alpha + b_{xu_\alpha} (\hat{\mu}_u - Mu_\alpha),$$

(3) $$\hat{\mu}_y = My_\beta + b_{yu_\beta} (\hat{\mu}_u - Mu_\beta),$$

(4) $$\hat{\sigma}_x^2 = Sx_\alpha^2 + b_{xu_\alpha}^2 (\hat{\sigma}_u^2 - Su_\alpha^2),$$

(5) $$\hat{\sigma}_y^2 = Sy_\beta^2 + b_{yu_\beta}^2 (\hat{\sigma}_u^2 - Su_\beta^2),$$

where $\hat{\mu}_u = Mu_t$ and $\hat{\sigma}_u^2 = Su_t^2$, and $t = \alpha + \beta$. These estimates are applied to equation 1, to form the conversion equation $Y = AX + B$ where $A = \hat{\sigma}_y/\hat{\sigma}_x$ and $B = \hat{\mu}_y - A\hat{\mu}x$.

With reference to the derived equations, if Groups α and β are identical in their mean performance on test score U, then the values of the parenthetical terms in equations 2 and 3 are found to be zero. In other words, the best population estimate of mean scores on Forms X and Y is the mean that was actually observed for Groups α and β, making group adjustments unnecessary. Similarly, this holds true for equations 4 and 5.

On the other hand, if Groups α and β are not identical in their mean performance on test score U, then the equating procedure does make group adjustments necessary. Ordinarily the adjustments simply reflect sampling differences in the groups which are chosen at random. While repeaters in the testing may be noted, the adjustments are not meant to reflect their presence.

Memory effects from the first administration of a test (item) will affect the result of the second if the same test (item) is administered on two successive occasions. The individuals need only remember the response given on the first occasion and make the same response on the

second, in order to obtain complete agreement between the results of the two measurements. That is, an agreement is obtained which affects the correlation between repeated measurements but which is not an expression of the method's reliability. That component of the score obtained on the first occasion which reappears on the second occasion will in part do so, not because the tests measure the same true score, but as the result of memory.

It is clear that where a test is being administered for the first time to a group of examinees the problem of repeaters does not exist. However, in a second testing the examinees typically include repeaters who have scored relatively low the first time. Since the repeaters are not randomly distributed score-wise, as in the original sample, systematic bias is introduced. The effect is probably more acute where there was a cut-off score separating those who had "failed" from those who had "passed."

With reference to the derived equation, if at the first administration (Form Y), Group β has no repeaters, MU_β is found. At the second administration (Form X), Group α having repeaters, Mu_α would tend to be depressed in value. Consequently, the group adjustment for Group α will be upward, unfairly favoring the repeaters along with those taking that test concurrently.

It should be obvious that the effect is stronger with an increase in the number of repeaters. Furthermore, in practice, the number of repeaters in a given test administration is ignored so that the strength of the effect is an unknown. Besides this, the score distribution of the total group

would tend to be skewed positively due to the presence of repeaters and hence make the linear approximation for conversion questionable.

In the above discussion it was suggested that the mean of repeaters would tend to be lower than the rest of the group. On the other hand, if other influences are operating, such as practice effects and recall of U items, the mean of repeaters might possibly tend to be higher than the rest of the group. In order to establish the resultant direction of the "repeater" effect, if any, the following study is presented.

Empirical Results

Two examinations were administered to groups of adults studying for the Chartered Life Underwriter (CLU) designation at The American College of Life Underwriters in 1973. The first examination was administered in January of 1973 and a second, parallel examination was administered in June, 1973.[2] Each examination consisted of 100 items, 20 of which (Form U) were common to both. The descriptive statistics for the non-repeaters (NR) taking the examination in January, the repeaters (R), taking the examination in June, and the combined (C) groups are presented in Table 1.

INSERT TABLE 1 HERE

Assuming that the groups taking the examinations in January and June are basically equivalent, the difference in the mean values of the common items, $\overline{U}y$ vs. $\overline{U}x$, should be merely due to sampling error. A t-test for uncorrelated means for unequal sample size was performed which did not

support this assumption ($\overline{U}y = 43.99$ and $\overline{U}x = 42.69$, $t = 4.66$, $df = 433$, $p < .01$). Based on this finding, it would appear that the groups taking the examinations in January and June are not random sample distributions from an underlying population distribution.

However, the nonrandom effect under investigation here has to do with repeaters present in the second administration of the examination. First of all, a t-test was performed with respect to the January examinees (Form Y, initially all are nonrepeaters), and those nonrepeaters of the June examinees (Form X); i.e., $\overline{U}_\beta$ vs. $\overline{U}_\alpha$ (NR). A statistically significant result was found ($\overline{U}_\beta$ (NR) = 43.99 and $\overline{U}_\alpha$ (NR) - 42.87, $t = 2.57$, $df - 423$, $p < .05$). Although restricting the comparison of the groups to the nonrepeaters, it appears that there is a significant difference albeit the level of significance moves from $p < .01$ to $p < .05$.

However, of direct interest in this investigation is the result of the t-test performed with respect to the repeaters (R) and nonrepeaters (NR) present in the June administration. Interestingly enough, while the greatest mean difference is observed for these groups ($\overline{U}x$ (NR) - $\overline{U}x$ (R) = 1.57 > $\overline{U}y$ - $\overline{U}x$ (combined) = 1.30 > $\overline{U}y$ - $\overline{U}x$ (NR) = 1.12), no statistical significance was found ($t = 1.57$, $df = 170$, $p > .05$). This finding makes sense if one notes the small sample size of the repeaters (N = 20) along with the low observed mean $\overline{U}x$ (R) = 41.30.

In other words, while the repeaters tend to introduce a systematic bias in the equating process, the small sample size masks the effect and traps those who attempt equating examinations into ignoring the influence. Nevertheless, while statistical significance was not found

with respect to repeaters and nonrepeaters in the June administration, the effect is now studied in terms of the conversion scores.

Applying the combined data to the equating equations (2) and (5), the conversion values A_1 and B_1 obtained are $A_1 = 1.0728$ and $B_1 = -9.9736$ for intact groups of January and June. The mean value for the June group ($\bar{X} = 90$), becomes 100 after conversion. The conversion was calculated by the linear addition of a constant to the January and June scores for the total test, t_α and t_β, equating subtest, U_α and U_β, and the remaining test items, Y_α and Y_β.

Applying only the data from nonrepeaters in similar fashion, the conversion values $A_2 = 1.0771$ and $B_2 = -9.0419$ are obtained. On this basis, the mean value for the June group of nonrepeaters ($\bar{X} = 89$) equates to 100 after conversion. If one were to use the equated mean values obtained from the combined data as the cutting score for "pass" and "fail", the difference of one point translates into "passing" an additional three percent of the examinees (r subjects) of the 172 total.

Extrapolating the extreme instance, where solely the repeater data is applied in the equating procedure, the conversion values $A_3 = 2.0951$ and $B_3 = -50.2374$ are obtained. The mean value for the June group of 80.00 then becomes 100 after conversion. The difference of 9 points now translates in "passing" an additional 29 examinees of the 172 total.

It is now clear that the performance of repeaters tends to move the cutting score downward, a greater number of repeaters having a greater influence. It follows that it is to the advantage of those who are going to repeat an examination to do so at a time when their numbers are great.

Discussion of Repeater Results

In the January 1973 examination, 63 individuals received a score less than or equal to the passing score of 115. The mean of the 63 failing scores is 105.46. Of the 63 individuals who scored less than 115 in the January 1973 examination, 20 chose to repeat the examination in June. The mean of these 20 repeaters for the January 1973 examination is 106.00. It is obvious that these individuals were no different in their average score than the entire sample of individuals who had failed the examination in January. One cannot, therefore, hypothesize that these individuals would repeat because their scores were significantly close to the passing point of 115 in January than the remaining group of individuals. The 20 individuals who failed the examination in January and chose to repeat the examination achieved a June score of 115.78, after the equating parameters had been applied. This June 1973 examination score is approximately 10 points higher than the January 1973 examination score achieved by the same group. Of the 20 repeaters that decided to take the June 1973 examination, 13 passed and 7 failed. A chi-square test of significance between CLUs and non-CLUs who repeated the examination in June, along the dimensions of those who received a passing or failing score was not significant ($p > .05$, $df = 3$).

A continuing analysis of repeaters and nonrepeaters on the equating items and a random set of items selected for comparison between the January and June scores was also carried out. Results indicated that repeaters' performance on equating items, who were categorized according to whether or not they were a CLU or non-CLU ($N = 11$ and 9 respectively),

was not significant at the $p < .05$ level. The chi-square value for this 2 x 4 analysis indicates that CLUs and non-CLUs did not significantly change their responses on the equating items from the January to the June examination ($\chi^2 = 2.79$, df = 3). This leads to the tentative conclusion that the equating items were reliable.

To determine whether this result was occurring by chance, a random set of 20 items was selected from both the January and June examinations. The criteria for selection was that these items may not occur in the equating subset. The 20 repeaters' responses and changes from January to June for these 40 items were recorded. By determining a frequency count of stability to these 40 items, a chi-square value was obtainable. The chi-square value of 9.101 was significant at $p < .05$ (df = 3). For this 2 x 4 analysis the results clearly indicate the high degree of change in responses for the repeaters from January to June on the random set of items that were selected.

Similar chi-square tests were carried out using nonrepeaters, 20 selected from the January examination and 20 from the June examination on the 20 equating items. This analysis, as with similar analyses, was matched for the 11 CLUs and 9 non-CLU students. The 2 x 4 chi-square was significant at $p < .01$ ($\chi^2 = 20.14$, df = 3). This result illustrates the change in score patterns from January to June for the 40 nonrepeaters on the 20 equating items. The final comparison was made for the same 40 nonrepeaters on 20 items randomly selected from both the June and January tests. Again, the chi-square was significant at the $p < .01$ level for the 2 x 4 analysis ($\chi^2 = 14.60$, df = 3).

An additional analysis was also carried out for the repeaters and nonrepeaters on the equating and randomly selected items. This information, presented in Table 2, shows the percent change for CLUs and non-CLUs, first from a right response to a wrong response, given the total of correct responses in the January administration; and secondly, from a wrong response to a right response given the total number of wrongs on the same test. This table illustrates these findings for the four groups mentioned above, that is, repeaters on the equating items and on the randomly selected items and nonrepeaters on both the equating items and randomly selected items.

INSERT TABLE 2 HERE

It is obvious that in comparing CLUs and non-CLUs on the percent switching from right to wrong response and wrong to right response on equating and randomly selected items, the greatest stability is achieved for the repeater group on the equating items. The average degree of switching for repeaters on the equating set of approximately 37 percent is less than any of the other three groups.

Suggested Solution

The common practice is to ignore the presence of repeaters and somehow vaguely assume a form of randomness has taken care of the problem. Giving some thought to the problem, a possible solution seems to lie in deriving equations which do not assume random groups. However, this sidessteps the problem rather than dealing with the presence of known repeaters.

A practical solution of how to deal with repeaters is simply not to include their scores in the calculations. Thus, the assumption of

randomness with respect to groups is not violated for this reason, and the conversion equations better reflect differences due to random sampling of groups. Subsequently, once the conversion equations are determined, the scores of the repeaters are subjected to adjustment in the same manner as those of the others.

Summary

The purpose of this investigation was to establish the effects of repeaters on test equating. Since consideration was not given to repeaters in test equating, such as in the derivation of equations by Angoff (1971), the hypothetical effect needed to be established. A case study was examined which showed results on a test as expected, overall mean was lower for repeaters. Applying these data to the available equating equations, it was shown that an additional three percent of the examinees was categorized as having "passed" than should if repeaters were taken into account. The practical solution offered is to hold separate the score of repeaters, execute the equating on the others and then apply the conversion to all the examinees.

TABLE 1

Means and Standard Deviations of Repeaters and Nonrepeaters Taking CLU Examinations in January and June, 1974

MONTH	SCALE (FORM)	CLASSIFICATION MEASURES (a)					
		NONREPEATERS (NR)		REPEATERS (R)		TOTAL (T)	
		$\bar{X}$	S.D.	$\bar{X}$	S.D.	$\bar{X}$	S.D.
JANUARY	Y	79.46	11.20				
	U	43.99	3.66				
	TOTAL (C)	183.45	14.32				
JUNE	X	81.71	11.64	80.00	7.04	81.52	11.20
	U	42.87	4.32	41.30	3.28	42.69	4.20
	TOTAL (C)	124.59	15.53	121.30	8.71	124.20	14.93

(a) JANUARY - (NR): N = 273; JUNE - (NR) = 152; JUNE - (R): N = 20; JUNE - (T): N = 172

TABLE 2

Percent of Change for CLUs and NonCLUs for Four Groups Based on Switching Responses From Correct to Incorrect and From Incorrect to Correct

Response Choice	Repeaters With Equating Items		Repeaters With Non-Equating Items Random Choice		Nonrepeaters With Equating Items		Nonrepeaters With Nonequating Items Random Choice	
	CLU	NonCLU	CLU	NonCLU	CLU	NonCLU	CLU	NonCLU
Correct To Incorrect	29	39	18	36	26	68	22	77
Incorrect To Correct	37	42	69	75	43	39	43	64
	$\bar{X}$ = 37%		$\bar{X}$ = 50%		$\bar{X}$ = 44%		$\bar{X}$ = 51%	

Footnotes

[1] Angoff, W.H., "Scales, Norms, and Equivalent Scores," Educational Measurement (2nd Ed.), 1971, page 562.

[2] All raw scores have been transformed (by addition of a constant).

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.), Washington, D. C.: American Council on Education, 1971.

Angoff, W. H. Basic equations in scaling and equating. In S. S. Wiles (Ed.), Scaling and Equating College Board Tests, Princeton, N.J., Educational Testing Service, 1961.

Cureton, E. E. and Tukey, J. W. Smoothing frequency distributions, equating tests, and preparing norms. American Psychologist, 1951, 6, 404.

Lennon, R. T. Equating nonparellel tests. Journal of Educational Measurement, 1964, 1, 15-18.

Levine, R. S. Equating the score scales of alternate forms administered to samples of different ability. Educational Testing Service Research Bulletin, 1955, No. 23.

Lord, F. M. Equating test scores -- a maximum likelihood solution. Psychometrica, 1955, 20, 193-200.