Florida Journal of Educational Research, 1973, 15

GENEROSITY IN ESSAY GRADING

John Follman
University of South Florida

Robert Reilly
University of Arkansas

SUMMARY

The generosity error in grading essay test responses is the bias of assigning too high a grade category. This bias should be reduced by manipulating the nature of the grading categories used.

Fifteen teachers graded 12 themes. The 15 teachers were randomly assigned, five each, to one of three types of rating category procedures. The three procedures were: Conventional (two positive, one neutral, two negative categories); Generosity (three positive, one neutral, one negative category); and Number (5, 4, 3, 2, 1). As expected, the conventional five categories produced a higher score than did the unbalanced five categories recommended by Guilford for offsetting the generosity error.

INTRODUCTION

The generosity error is the tendency of raters to assign ratings which are too high. Gronlund (1965) suggests that there are two consequences, both psychometrically undesirable, of the generosity effect. Initially, high ratings (generosity effect) may reflect more of the characteristics of the rater than of the ratee. Secondly, high ratings (generosity effect) may limit the range of ratees' ratings because of a pile up in the top category thus precluding reliable discriminations among those at the top. However, there is little empirical information about the generosity effect. Guilford (1954) suggested one way to counteract the leniency, generosity error. He suggested use of three positive, one neutral, and one negative category in the rating scale, rather than a balanced number of positive and negative terms.

The purpose of this study was to investigate the effects of different rating formats on level of grades awarded so that inferences could be made about the size of the generosity effect and also about the value of one way to reduce it. Specifically investigated were the effects of the Guilford (1954) anti-generosity format, a conventional format, and a numbers format, on the level of grades awarded.

A Conventional, a Generosity (anti-generosity), and a Number format were used. The first two formats were anchored to verbal descriptions while the Number format was not. The Conventional categories were Superior, Above Average, Average, Below Average, Inferior. Generosity categories were Superb, Superior, Above Average, Average, Below Average. Thus the Generosity format has one additional positive category compared with the Conventional format. Number Categories were 5, 4, 3, 2, 1.

PROCEDURE

Fifteen teachers enrolled in a masters level educational psychology course at the University of South Florida in July, 1971 were randomly assigned, five each, to one of the three grading groups, Conventional, Generosity, or Number. The Ss graded 12 themes considered to be fairly typical high school and college freshman level themes. The themes represented a range of quality. The Ss were instructed to read all 12 themes before grading any and to grade holistically. The Ss were instructed to write comments on the themes as if the themes were to be returned to the writers. The Ss in the Conventional group, the Generosity group, and the Number group, respectively, were told to use the appropriate respective grade format and that they could use each grade category as many or as few times as desired. For additional information concerning the themes and procedure see Follman, Miller, Lowe, and Stefurak (1970) or Follman, Lowe, and Miller (1971).

The categories for Conventional, Generosity, and Number, respectively, were converted into raw scores for the statistical analyses thus: Superior (Superb) (5) as 5; Above Average (Superior) (4) as 4; Average (Above Average) (3) as 3; Below Average (Average) (2) as 2; and Inferior (Below Average) (1) as 1.

RESULTS

Means and standard deviations of scores were 2.65 (s=.10) for Conventional, 2.02 (s=.21) for Generosity, and 2.93 (s=.39) for Number. A type 1 ANOVA indicated significant ($p = .01$, or less) F's of 11.44 for formats, 29.98 for themes, and 2.10 for the formats X themes

interaction. These findings are compelling evidence of the effects of the different grading formats. The Generosity format produced the lowest mean, evidence of the reduced generosity effect. The lower Generosity format mean compared with the Conventional format is not inconsistent with the findings of Hill (1953). Hill (1953) compared a nine category favorable-unfavorable attitude continuum with a seven category "favorable only" attitude continuum. The category statements, had been scaled previously by the method of equal-appearing intervals. The seven-category "favorable only" statements had both a lower range of scale values and also a lower average scale distance between all possible pairs of values.

It is interesting to note that the Number format produced the highest grades, higher than those of the Conventional format. This is consistent with the findings of Follman, Kleg, and Neel (1971) where means were 3.45 for Number grades, 3.23 for Letter grades, and 2.99 for Word grades.

Within ANOVAs were run for each theme independently to determine differences in levels of scores among the three formats. There were significant differences among the groups for five of the individual themes. Mean level of grade for all five themes showed Generosity lowest, Number highest, with Conventional intermediate. This is viewed as additional evidence of the differential influences of the three formats.

CONCLUSIONS

1. It is concluded that grading (or rating) formats affect the the level of grades awarded.

2. It is concluded that the anti-generosity format suggested by Guilford (1954) reduces the generosity effect.

REFERENCES

Follman, J. C., Miller, W. G., Lowe, A. J., and Stefurak, D. W. Effects of time and typeface on level and reliability of theme grades. Research in the Teaching of English, 1970, 4, 51-58.

Follman, J., Lowe, A. J., and Miller, W. Graphics variables and reliability and level of essay grades. American Educational Research Journal, 1971, 8, 365-373.

Follman, J., Kleg, M., and Neel, J. Piles vs. no piles, and letter vs. number vs. word grades in theme grading. Unpublished study, University of South Florida, 1971.

Gronlund, N. E. Measurement and evaluation in teaching. New York: The Macmillan Company, 1965.

Guilford, J. P. Psychometric methods. New York: McGraw-Hill Book Company, 1954.

Hill, R. J. A note on inconsistency in paired comparison judgments. American Sociological Review, 1953, 18, 564-566.