Florida Journal of Educational Research, 1974, 16

PILES, AND NUMBER AND KIND OF CATEGORIES, IN THEME GRADING

John Follman
University of South Florida

SUMMARY

Three studies dealing with the effect of sorting essays into piles on the reliability and level of grades awarded are reported. In the first study, 40 S's were randomly assigned to one of four experimental conditions; in the second study, 30 S's were randomly assigned to one of six conditions; and in the third, 32 S's were assigned to one of eight conditions. The conditions involved combinations of piles or no piles and different grading systems. In each study, the S's graded 12 typical high school senior and/or college freshman themes. It was concluded that the use of piles neither increases group grading reliability estimates nor greatly influences the level of grades awarded.

INTRODUCTION

The main objective of the three studies summarized in this survey was to determine the effects of the use of piles on the reliability and level of grades awarded to English composition themes. A second objective was to determine the effects of the number and also of the kind of categories on the reliability and level of grades awarded to English compositions.

Piles vs No Piles. A recommendation made by many measurement text authors is that essay graders sort answers into piles, each pile representing either a priori determined or then being judged, grades, ranks, piles, qualitative levels, etc. The a priori categories have included model essays. Authorities have often encouraged the use of piles, but there has been little empirical underpinning. Despite this lack of research there has been considerable variation in the number of piles recommended, presumably used, and used. Bracht and Hopkins (1970) used five piles. Stalnaker (1951) advocated the use of six piles. Peaker (1953) employed seven categories, and French and Diederich (1968) used nine. Diederich (1950) used ten piles. Ebel (1965) noted the statistical nicety that the more piles used the greater the precision in scoring. No empirical evidence is cited in any of these essay grading articles documenting the number of piles recommended or used. However in a different type of task, rank order judgments of size, Madden, Hazel, and Bourdon (1965) compared the following five sorts: two piles; large piles; three piles; large-small; and free. The free sort, using any procedure the S desired, produced significantly less errors than the other sorts.

Therefore the main objective of these three studies was to determine empirically the effects of piles on the reliability and level of grades awarded to themes.

Number of Categories. The number of categories, intervals, discriminations, points, grades in essay grading, particularly grading of themes, has received relatively little empirical examination. While there has been little research of the effects of the number of grading categories there has been considerable variety in the number actually used, two through 300 or more categories. One two-category system frequently used is satisfactory or unsatisfactory. Three categories are also used, such as above average, average, below average, or 1, 2, 3. More common, especially in the United States, are five categories A, B, C, D, F, or 1, 2, 3, 4, 5. A refinement of this is the addition of plusses and minuses, A+ through F-, for a total of 15 categories. This scale also has been commonly used in the United States as well as in England (Vernon and Millican, 1954). (The Vernon and Millican article is an excellent overview of essay grading of compositions in England and is strongly recommended to Americans interested in theme grading.) Kincaid (1953) obtained moderately high and high rater reliability estimates with a 10 category system. Wiseman (1949) used 20 categories and other refinements and obtained reliability estimates as high as those of most standardized objective achievement tests. (Examination of his approach is also recommended to interested Americans.) Lamb (1953) reported high reliability estimates using 30 categories. Wood (1928) used a 100 point scale. Buxton (1958) obtained high reliability with a 300 point system.

There has been little research of the effects of the number of grading categories on reliability and particularly level of grades awarded as McColly (1970) noted. McColly (1970), in a recent overview of theme grading, states that there is little knowledge of the effects of the number of points or intervals and that we are in need of trustworthy experimental findings. McColly and Remstad (1965) in a major theme study compared a 4-point scale with a 6-point scale and found little or no differences in reliability estimates or distributions. Wiseman (1949) stated that a 15 point A+ through F- scale gave less spread than the numerical scale, apparently 20 categories, which it superseded. Conversely, Peaker (1953) recommended using a few classes, Very Good, Good, Middling, Fair, Poor.

Peters and Van Voorhis (1940) estimated reliability coefficients associated with different numbers of categories. For material with a reliability of .90, 15 categories reduced the coefficient to .89, 10 reduced it to .87, five to .80, and two to .60.

Consequently, another objective was to determine empirically the effects of number of grade categories on the reliability and level of marks awarded.

Kinds of Grade Categories. There has been little research reported on different kinds of categories of grades used for English themes or for essay grading in general. Ebel (1965) advocated the use of numerical symbols rather than letter grades. No empirical evidence was cited however. Marshall (1968) has long advocated that traditional grades be abandoned and that descriptive words be used instead. However he did not cite any empirical evidence either.

An objective of one of the three studies was to determine empirically the effects of kinds of grades on reliability and level of marks awarded.

METHODOLOGY

The method followed in each of the three studies was to randomly assign college undergraduates (Ss) to their respective piles and/or number (kind) of grade category conditions.

In the first study (Follman, Wong, and Miller, 1971) 40 Ss were randomly assigned, 10 each, to one of four experimental conditions: Plus and Minus A, B, C, D, F, 15 category grades - Piles; Plus and Minus 15 category A, B, C, D, F, grades - No Piles; A, B, C, D, F, grades - Piles; A, B, C, D, F, grades - No Piles.

In the second study (Follman, Kleg, and Neel, 1972) 30 Ss were randomly assigned, five each, to one of six experimental conditions: Piles - Letter grades; No Piles - Letter Grades; Piles - Number Grades; No Piles - Number Grades; Piles - Word Grades; No Piles - Word Grades.

In the third study (Follman, Neel, and Miller, 1972) 32 Ss were randomly assigned, four each, to one of eight experimental conditions: Piles - Grades 3, 2, or 1; No Piles - Grades 3, 2, or 1; Piles - Grades 4, 3, 2, or 1; No Piles - Grades 4, 3, 2, or 1; Piles - Grades 5, 4, 3, 2, or 1; No Piles - Grades 5, 4, 3, 2, or 1; Piles - Grades 6, 5, 4, 3, 2, or 1; No Piles - Grades 6, 5, 4, 3, 2, or 1.

All Ss were instructed to read all themes before grading any, to grade holistically, that they could use each grade as many times as desired, and to comment on the themes as they would if the themes were to be returned to the original writers.

Piles Ss were instructed to place the themes in piles, each pile representing a currently being determined grade category, one for each letter grade.

The themes used were 12 typical high school senior and/or college freshman level themes used in many studies by the senior author. For a description see Follman and Anderson (1967).

Letter grades were converted into raw scores A = 5, B = 4, C = 3, D = 2, F = 1. Winer's (1962) ANOVA was used to obtain adjusted group reliability estimates. In addition ANOVA's were conducted to determine if significant differences were associated with the different experimental conditions.

RESULTS

In the first study reliability estimates extended from .61 to .87 for the four experimental groups. Higher reliability estimates were not associated with the 15 category grades vis-a-vis the 5 category grades. No significant differences obtained in grade levels for the piles - no piles comparisons. Higher reliability estimates were associated with the no piles conditions vis-a-vis the piles conditions.

In the second study reliability estimates extended from .55 to .92 for four of the experimental groups. For each of the other two groups the ANOVA was inappropriate because the variance within individual graders across themes was large compared with the variance between graders across individual themes. Neel's Discrepance (1970) indicated considerable consistency across raters for each theme within each of the two groups. No significant differences obtained in the piles - no piles comparisons. Mean grades were 3.45 for Number Grades, 3.23 for Letter Grades, and 2.99 for Word Grades. These differences while not significant are interesting.

In the third study reliability estimates were ca. .85 for three of the eight experimental groups. For the other five groups the ANOVA was again inappropriate because of the within-between variance anomaly. Neel's Discrepance (1970) indicated much consistency across raters for individual themes within each group.

Significant ($p < .05$) differences obtained in the piles - no piles comparisons with higher grades in three of the four comparisons in favor of no piles. Another significant ($p < .05$) finding was for number of grade categories. This finding reflected the fact that as the number of categories increased the use of the lowest grade categories decreased correlatively.

OVERVIEW

Overview of the reliability results of the three studies indicates that the use of piles apparently does not enhance the reliability of group grading of English themes. Generally there was moderately high to high grader reliability within each group with use or non-use of piles producing no discernible effects.

Overview of the level results of the three studies indicated no differences for the first two studies and surprisingly, significantly higher grades associated with the no piles condition in three of four comparisons in the third study.

In the second study, number grades were higher than letter grades which were higher than word grades.

In the third study an interesting finding was that as the number of possible grading categories available in a set increased, the use of the lowest grade in each set decreased.

Finally, since reliability estimates were not higher in groups with many grade categories as opposed to groups with few grade categories as was expected, and since the use of the lowest grade levels declined as the number of grade categories increased, more research on number of grade categories is recommended.

CONCLUSIONS

Examination of the empirical evidence across the three studies suggests the following conclusions:

1. The use of piles does not increase group grading reliability estimates.
2. Level of grades awarded is little influenced by the use of piles.
3. Reliability estimates, associated with many grade categories as opposed to few categories, are not higher as would be anticipated.
4. As the number of possible grade categories increases, the use of the lowest grades decline.

This paper was presented at the National Conference on Measurement in Education, New Orleans, Feb. 27, 1973.

REFERENCES

Bracht, G., and Hopkins, K. The communality of essay and objective tests of academic achievement. Educational and Psychological Measurement, 1970, 30, 359-364.

Buxton, E. An experiment to test the effects of writing frequency and guided practice upon students' skill in written expression. Unpublished Ph. D. dissertation, Stanford University, 1958.

Diederich, P. The 1950 College Board English Validity Study. Research Bulletin, RB-50-58, Educational Testing Service, Princeton, New Jersey, 1950.

Ebel, R. Measuring Educational Achievement. Englewood Cliffs, New Jersey, Prentice-Hall, Inc., 1965.

Follman, J., and Anderson, J. An investigation of the reliability of five procedures for grading english themes. Research in the Teaching of English, 1967, 1, 190-200.

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., Wong, M., and Miller, W. Piles, number of grade categories, and theme grading. University of South Florida, Unpublished study, 1971.

Follman, J., Kleg, M., and Neel, J. Piles vs no piles, and letter vs number vs word grades in theme grading. Journal of English Language Teaching, 1972, 7, 24-26.

Follman, J., Neel, J., and Miller, W. Piles vs no piles, and 3 vs 4 vs 5 vs 6 categories in theme grading. University of South Florida, Unpublished study, 1972.

French, J., and Diederich, P. Wanted: Papers to be graded by sixty readers. Educational Testing Service, Princeton, New Jersey, 1968.

Kincaid, G. Some factors affecting variations in the quality of students' writing. Unpublished Ed. D. dissertation, Michigan State University, 1953.

Lamb, H. The english essay in secondary selection examinations: A comparison of two methods of marking. British Journal of Educational Psychology, 1953, 23, 131-133.

Madden, J., Hazel, J. T., and Bourdon, R. D. A comparison of error in five sorting procedures for ordinal ranking. Journal of Applied Psychology, 1965, 49, 170-171.

Marshall, M. S. Teaching without grades. Corvallis, Oregon State University Press, 1968.

McColly, W. What does educational research say about the judging of writing ability? The Journal of Educational Research, 1970, 64, 148-156.

McColly, W., and Remstad, R. Composition rating scales for general merit: An experimental evaluation. Journal of Educational Research, 1965, 59, 55-56.

Neel, J. H. The Discrepance, a Measure of Inter-Judge Reliability. Paper given at American Educational Research Association, Minneapolis, 1970.

Peaker, G. A sampling design used by the ministry of education. Journal of the Royal Statistical Society, 1953, 116, 140-165.

Peters, C., and Van Voorhis, W. Statistical Procedures and their Mathematical Bases. New York, McGraw-Hill Book Company, 1940.

Stalnaker, J. The essay type of examination. Educational Measurement E. F. Lindquist, Editor) American Council on Education, Washington, D. C., 1951.

Vernon, P., and Millican, G. A further study of the reliability of english essays. The British Journal of Statistical Psychology, 1954, 7, 65-74.

Winer, B. J. Statistical Principles in Experimental Design, . New York, McGraw-Hill Book Company, 1962, 124-132.

Wiseman, S. The marking of english composition in grammar school selection. British Journal of Educational Psychology, 1949, 19, 200-209.

Wood, E. P. The scoring reliability of test material in the free answer and short answer forms. Public Personnel Studies, 1928, 6, 98-108.