DOCUMENT RESUME

ED 352 926 IR 015 857

AUTHOR Brown, William L.; Stevens, Betty L.
TITLE Using the Microcomputer To Equate Ratings of Student Writing Samples.
PUB DATE 21 Apr 92
NOTE 11p.; Paper presented at the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Education; Evaluation Methods; *Interrater Reliability; Judges; Junior High Schools; *Microcomputers; Models; Statistics; *Writing (Composition); *Writing Evaluation; Writing Tests
IDENTIFIERS *Rasch Model

ABSTRACT

The objectives of this study were to determine whether student writing portfolios could be rated reliably by trained judges; study the effects on student ratings of the differential leniency of the judges; and ascertain the effects of writing-prompt difficulty and its interactions with rater leniency. Writing samples from 127 students in grades 3, 6, and 9, were randomly selected and analyzed by experienced volunteer judges working under the direction of a consultant experienced in both training judges and reading student writing. The Rasch Multi-Faceted Model using a microcomputer was applied to the rating and a rankings of the writing samples. The most important conclusion reached from this study is that the microcomputer can be used as a tool for equating the ratings of student writing at an economical cost; the Rasch Multi-Faceted Model can be used to compensate for variation in judge leniency and produce consistent ratings; and although the judges appeared to be able to rate writing pieces effectively, they were unable to come up with any consistent approach to ranking pieces within a portfolio. Eight tables illustrate the ratings. (ALF)

Using the Microcomputer to Equate Ratings of Student Writing Samples

by
William L. Brown, Consultant
and
Betty L. Stevens, Consultant
Michigan Educational Assessment Program

Presented to the
American Educational Research Association
San Francisco, California
April 21, 1992

Running Head: Equating Ratings of Student Writing Samples

OBJECTIVES

Writing is the process of selecting, developing, arranging and communicating ideas. The process requires students to write in a variety of forms (letters, stories, journals, essays, etc.) for a variety of purposes (to inform, to persuade, to describe, etc.), and for a variety of audiences (peers, teachers, self, etc.). Each form, purpose and audience demands differences of style, and approach. A wide variety of writing experience, therefore, is critical in order to becoming an effective writer.

The objectives of the reported study were to (1) determine whether student writing portfolios could be rated reliably by trained judges, (2) study the effects on student ratings of the differential leniency of the judges, and (3) ascertain the effects of writing-prompt difficulty, and its interactions with rater leniency.

PERSPECTIVES

Traditionally, there have been difficulties in using judges to rate writing samples because of variations in the leniency of the judges and variations in the difficulties of the dimensions being rated. These exist even when the judges are highly trained and professional, and have led to a reluctance to use judges to score materials which will result in comparisons among writers.

The measurement model developed by Rasch and later refined by researchers at the University of Chicago (Wright & Masters, 1982; Linacre, 1989) has brought about the possibility of accounting for the stochastic nature of ratings by using statistical information to "correct" for differences among raters and among dimensions being measured. As long as there are "crossings" of raters across common dimensions and items, the microcomputer is capable of applying corrections for rater leniency and dimension difficulty.

DATA SOURCE

As a part of the Michigan Educational Assessment Program's mission to develop a state assessment that is consistent with the process writing instructional model advocated by Michigan educators, a sample of student writing was measured in 1990. This research pilot studied instructional delivery and assessment as variables that impact student writing. In all, writing samples from 127 students were randomly selected and analyzed.

MAY 1990 WRITING STUDY

The voluntary pilot was conducted in four suburban and rural Michigan school districts in May of 1990. One suburban district (for purposes of this report referred to as District A, 5,424 students) has an articulated K-12 writing program supported by a yearly writing assessment; one rural district (District B,779 students) has a K-12 writing program; while another (District C, 2,815 students) is considering initiating a writing program in the near future. In the fourth district (District D, 16,628 students), there are some classes with excellent writing programs while, at the same time, writing opportunities are nonexistent in other classes.

Each district was asked to select at least one class each at grades three, six, and nine to participate in the pilot. A district coordinator was designated the responsibilities of disseminating and collecting materials and providing information to classroom teachers. Each student was given a writing folder and asked to do the following:

1. Produce a finished writing piece over a period of time of up to five days, keep all materials related to the writing piece in an individual student folder to be turned in at the end of the five days, and complete a survey related to the process employed in producing the finished paper. Students were asked to respond to the following prompt: "Write about an environmental issue or topic of your choice. You may write for the school newsletter, a newspaper or a magazine."

2. Produce a finished writing piece in a single sitting consisting of no more than a 40-minute time period. Students were asked to respond to the following prompt: "Choose an interesting character or person to write about. Tell why that person is interesting. You may choose anyone, including a family member, a personal friend, a character from history or someone from a story, a T.V. show, or movie."

3. The student may have chosen to include a self-selected, finished writing piece. This paper could have been written for this or another class; it could have been a letter written for a specific purpose, thoughts collected in a journal, or any other type of writing. It was stressed that the student must select the writing piece. The student had the option to respond to this request by stating "None is available."

All student writing and related work was placed in the individual student folder and returned to the district coordinator. All student folders were returned to the project coordinator who sorted, coded, and prepared them for reading.

PROCEDURE FOR EVALUATION

Experienced volunteer judges were divided into three grade level teams. Working under the direction of a consultant experienced in both training judges and reading student writing, each team randomly selected sample student folders and proceeded as follows.

1. Judges attempted to score the folder holistically using a four point scale (four being the highest score). At the same time, they ranked from one to three, the individual pieces of student writing within the folder.
2. Having read the entire collection of folders, each team agreed upon a team score for each student folder.
3. After reading a number of five-day selections, the teams selected anchor papers for the purpose of establishing scoring criteria.
4. Samples of five-day writing pieces were read and scored holistically using the scoring criteria.

Keeping in mind that the reading of student folders and individual papers was investigative as well evaluative, the reading teams continually communicated in an attempt to provide consistency with regard to procedure.

METHODS

The Rasch Multi-Faceted Model was applied to the ratings and rankings of the pilot writing samples. The microcomputer time needed to analyze the ratings for 127 students was less than five minutes, or about 2.5 seconds per student. It was found that rankings of portfolio selections are very unreliable (reported reliability = 0.0). Upon inspection of the data, it appears that some raters gave very high ratings to the 5-day writing sample, for example, yet ranked the 5-day sample very low in the portfolio. This situation was verified by the judges, who reported that they were very uncomfortable about ranking items in the folders. Whatever process was operating in this procedure, it remains that statistically it appears to be inconsistent and therefore leads to unreliability.

In total, writing samples from 127 students were evaluated. The number of students whose papers were rated is shown in Table A.

TABLE A: Number of Students

	Grade 3	Grade 6	Grade 9
School A	21.0	20.0	16.0
School B	20.0	10.0	21.0
School C	18.0	12.0	14.0
School D	21.0	0.0	14.0

For this application, the items being rated were a (individual judge) portfolio score, an overall (group agreed-upon) portfolio score, and an individual judge rating of the 5-day writing sample. The measurements obtained included the student writing performance (as determined from the three ratings) and the rater leniency. In terms of the scores given to the writing samples, the reliability was very high, ranging from a low of 0.81 to a high of 0.92, as shown in Table B:

TABLE B: Reliabilities (Ratings)

	Student Meas't Reliability	Judge Meas't Reliability	Item Meas't Reliability
Grade 3	0.92	0.28	0.00*
Grade 6	0.87	0.92	0.68
Grade 9	0.81	0.00*	0.87

(*Note: Reliabilities of 0.00 occur when all of the variation in the data can be accounted for outside of the characteristic being estimated. For example, a judge measurement reliability of 0.00 indicates that all of the variation could be accounted for by error of measurement. This means that, in this model, the microcomputer could not tell the judges apart from each other. This may or may not be desirable.)

Table C shows the average ratings for each of the three grades in each of the four schools in the pilot sample.

TABLE C: Average Ratings

	Grade 3	Grade 6	Grade 9
School A	3.08	3.21	2.93
School B	2.11	2.52	2.98
School C	1.57	2.48	2.44
School D	2.80	2.74	1.98

Although these ratings appear to be very close together, they actually represent a significant amount of difference when item difficulty and judge leniency are taken into consideration. The measurements described in Table D are based on the multi-faceted model described above, where -10 is the lowest rating and +10 is the highest, with an average being 0.

TABLE D: Average Abilities

	Grade 3	Grade 6	Grade 9
School A	3.88	3.33	4.08
School B	2.59	0.10	4.16
School C	6.94	-0.10	0.58
School D	2.12	1.05	-2.64

Since students with an average rating of 1 (the lowest possible) or 4 (the highest possible) are considered to not fit the measurement model (i.e., their true scores cannot be determined because they may really be lower than a 1 or higher than a 4 but the scale cannot show these extreme levels), it is important to know how many students fell into those categories. Table E shows the percentage of students from each group which scored an average rating of 1. In School A, 14% of Third Graders and 10% of Sixth received perfect 4 ratings.

TABLE E: Percent of Students with all Ratings = 1

	Grade 3	Grade 6	Grade 9
School A	0.0	0.0	0.0
School B	5.0	0.0	0.0
School C	39.0	0.0	7.0
School D	0.0	0.0	0.0

TABLE F: Reliabilities (Rankings)

	Student Meas't Reliability	Judge Meas't Reliability	Item Meas't Reliability
Grade 3	0.00	0.00	0.66
Grade 6	0.01	0.00	0.81
Grade 9	0.00	0.00	0.00

Table F shows that practically none of the variability in student scores could be accounted for by variations in student writing performance, judge performance, or type of writing in the folders (with the exception of a slight variation in writing difficulties at grades 3 and 6). This makes sense, since the rating scores shown earlier indicate that good writers were consistently rated high by all judges on all items (writing types), and poor writers were consistently rated low by all judges on all items. All items (writing types) appeared to have equal difficulty at third grade (mean rating = 2.5), whereas at grades six and nine the mean ratings varied as shown in Table G:

TABLE G: Student Average Ratings

	Grade 3	Grade 6	Grade 9
Combined Folder	2.50	3.10	2.90
Folder	2.50	2.50	2.80
5-day Paper	2.50	2.90	2.60

The judges produced average ratings shown in Table H:

TABLE H: Judge Average Ratings

	Grade 3	Grade 6	Grade 9
Judge A	2.50	2.80	3.00
Judge B	2.60	3.00	2.60
Judge C	2.60	2.50	2.70
Judge D	2.40	2.70	2.80
Judge E	-	-	2.70
Judge F	-	-	2.80

These results show a high degree of consistency, although there is a significant variation in rater leniency at each grade level (which the microcomputer model takes into account).

CONCLUSIONS

The most important conclusion which can be reached from this study is that the microcomputer can be used as a tool for equating the ratings of student writing at an economical cost. As long as at least two judges rate each paper, and there are "crossings" of judges (i.e., judges are not always paired the same), it appears that the many-faceted Rasch model can be used to compensate for variations in judge leniency and item difficulty to produce consistent ratings. Since the microcomputer was not part of the original design of the study, the ratings were available but the papers were not, so they could not be used to check cases where the microcomputer results differed from the judges' ratings. Validation of the validity of this assessment must, therefore, await external verification, since we have no clear standard for comparison on the sample used in this study.

In general, the judges who rated the writing samples performed consistently across all students and all items (writing types). There were no judges or items which did not fit the multi-faceted measurement model. There was a wide range of performance evidenced by the writers at all grade levels, yielding reliabilities of .81 at grade 9, .87 at grade 6, and .92 at grade 3. Item (writing types) difficulties were the same for all three dimensions at grade 3, but varied significantly at grades 6 and 9. The judges showed a remarkable degree of consistency, yet each demonstrated a different degree of leniency as reflected in average ratings.

When the judges ranked the writing pieces in the portfolio folders, instead of rating them, practically none of the variability in scores could be accounted for by variations in student writing performance, judge performance, or type of writing in the folders. On the other hand, good writers were consistently rated high by all judges, whereas poor writers were consistently rated low, across all writing types. Variability in difficulty across different types of writing occurred only at grades 6 and 9.

Although all judges appeared to be able to rate writing pieces effectively, they were unable to come up with any consistent approach to ranking pieces within a portfolio.

REFERENCES

Linacre, J. M. (1989). FACETS: Many-faceted Rasch analysis with FACFORM data formatter. (Program manual.) Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). **Rating Scale Analysis.** Chicago: MESA Press.