DOCUMENT RESUME

ED 287 889 TM 870 651

AUTHOR Lehmann, Rainer H.
TITLE Reliability and Generalizability of Ratings of Compositions.
PUB DATE 22 Apr 87
NOTE 18p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, DC, April 20-24, 1987). Some pages contain small, light type.
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Evaluation Problems; Foreign Countries; High Schools; International Programs; *Interrater Reliability; Multivariate Analysis; Scoring; Student Characteristics; *Test Reliability; *Writing Evaluation
IDENTIFIERS International Assn Evaluation Educ Achievement; International Study of Achievement in Written Comp; West Germany (Hamburg)

ABSTRACT

A total of 1,487 eleventh grade students from the Hamburg (West Germany) school system were asked to complete four writing assignments used in an International Association for the Evaluation of Educational Achievement (IEA) study of writing assessment. In analyzing the writing samples, the study focused on: (1) between-rater effects; (2) within-rater effects; (3) between-assignment effects; and (4) within-student effects. Two independent sets of scores on a 5-point scale were awarded to each essay in accordance with the international scoring guides. Since a study of the above four factors involves a complex array of statistical analyses, researchers did not rely on parametric test models alone. They included more intuitive statistics such as percentage of perfect agreement between two independent ratings and percentage of loose agreement defined as the percentage of differences between two ratings not greater than one percentage point. Results showed that there was not a single or simple answer to the reliability of measuring general writing achievement across the tasks used. Instead, the solution depended on the kind of assumptions about the tasks that one is prepared to make. Statistically, the answer was a function of whether within-student variation was considered as true variance or error. It was also concluded that 13 writing tasks would have been needed to obtain satisfactory generalizability. (KSA)

Rainer H. Lehmann School of Education
University of Hamburg
Von-Melle-Park 8
D-2000 Hamburg 13
West Germany

Reliability and Generalizability of Ratings of Compositions

Presentation at the Symposium 'Issues in Scoring and Score Reporting in Large-scale Writing Assessments', AERA Annual Meeting, April 22, 1987

1. Introduction

One of the aims of the IEA 'International Study of Achievement in Written Composition' is 'to make a contribution toward solving problems related to the assessment of essay-type answers' (IEA 1985, p.29). There are, of course, - as anyone seriously engaged in the field knows - many such problems, only a few of which can be dealt with in this paper. Here, only such questions are treated which focus on the following four sources of variation in the assessment of student writing:

1. between-rater factors
2. within-rater factors
3. between-assignment factors
4. within-student factors. (cf. Wesdorp et al. 1982, p. 299f)

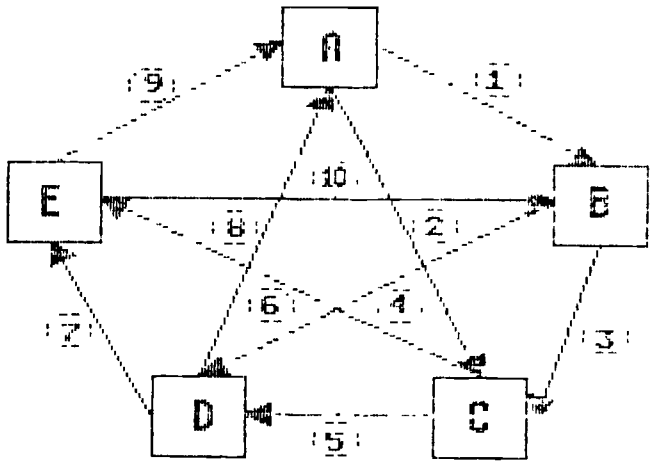
Since in the IEA Study no students were asked to work repeatedly on the same assignment, within-student and between-assignment factors are confounded, so that they can only be analyzed conjointly here. However, a suggestion will be elaborated which is based on the notion of variance components and which allows for a simultaneous evaluation of those effects.

The data used for exemplification come from the West German component of the IEA Study, involving a total of 1487 11th-grade students from 71 classrooms in eight different tracks of the school system of the City of Hamburg. Each of these students had been asked to complete four writing assignments, i.e. one more than internationally obligatory, in a partly rotated design:

1. one of the four short assignments in the format of a letter (international Tasks 1A, B, C, E);
2. one of three longer international assignments (narrative Task 5; argumentative Task 6; reflective Task 7);
3. the assignment of a letter of advice to a friend (international Task 9);
4. the assignment of paraphrasing, analyzing rhetorically, and evaluating a newspaper commentary (nationally optional task, here referred to as Task 0).

Two independent sets of scores on a five-point scale were awarded to each essay in accordance with the international Scoring Guides (IEA 1984). In each case, the raters were two from a jury of five fully trained and certified mothertongue teachers. The essays were distributed among the raters in such a way as to allocate approximately equal portions to all possible combinations of raters.

Figure 1: Scoring Scheme of the Hamburg Study of Achievement in Written Composition



Rater	1st round package no.		2nd round package no.	
A	1	2	8	9
B	3	4	10	1
C	5	6	2	3
D	7	8	4	5
E	9	10	6	7

2. Methodological Considerations

The study of between-rater effects, within-rate effects, and between-assignment/within-student effects entails, for a study of this size and complex design, an impressive array of very specific statistical analyses, far beyond of what can be reported here. To give just an indication of what is involved, it should be mentioned that the international reporting requirements for scoring included, for the case of the Hamburg Study, the completion of 144 form sheets, containing a total of several thousand statistics referring to the 9 different tasks, 5 different raters, and 10 possible combinations of raters. When these forms were conceptualized, little was known with respect to the actual rater performance. Thus it had seemed appropriate not to rely on parametric test models alone, but also to include more 'intuitive' statistics.

Two such intuitive concepts are 'percentage of perfect agreement' between two independent ratings and 'percentage of loose agreement', defined as the percentage of differences between two ratings not greater than one scale point. While these concepts have the advantage of not being based on assumptions as to the homogeneity of means and variances, their major draw-back is that they cannot be converted into well-defined reliability coefficients. Moreover, 'percentage of loose agreement' does not discriminate well between quality levels, if the agreement between two independent ratings is generally high: in the Hamburg data, there is more than 97 percent loose agreement between independent ratings. Perfect agreement was achieved on 73.2 percent of all 5362 rated compositions.

If, however, the hypotheses of homogeneous variances (and means) between independent ratings can be maintained, correlations between ratings and associated measures such as regression coefficients, correlation ratios, and variance components can be used. They are clearly superior in that they can be related to the classical reliability coefficient which is defined as the (estimated) true variance divided by the observed variance:

$$\text{Reliability} = \frac{\sigma^2_t}{\sigma^2_o} = \frac{\sigma^2_t}{\sigma^2_t + \sigma^2_e}$$

with σ^2_t = true variance
σ^2_o = observed variance
σ^2_e = error variance (cf. Thorndike 1982)

It is assumed that an appropriate treatment of problems involving the reliability and generalizability of essay ratings should be based on this concept.

3. Testing the assumptions - homogeneity of score variances and means

There are a number of ways to ascertain that scores from different raters do, indeed, display variances sufficiently similar to be compatible with the hypothesis of homogeneity. One of these is to look at the extremes within tasks, i.e. that pair of raters for which in a given task the observed difference in standard deviations is largest. A statistical problem lies in the partial overlap of the sets of compositions scored by these rater pairs. So, it is necessary to apply two different tests: (1) the conventional F-test for comparing the sub-sets which were unique to either one of the raters in the pair, (2) a t-test for paired observations (Ferguson 1966), applying only to that portion which was scored conjointly (but independently!) by the two. The results from the Hamburg data clearly suggest to retain the hypothesis of homogeneous variances: while for five of the nine tasks, not even the largest observed differences were significant on the first criterion, none of the pairs investigated showed significant differences on the second. Conversely, the t-test identified only four overlapping sets for which there were significant differences in <u>any</u> of the ten possible combinations of raters, but these findings could not be reproduced on the basis of the F-criterion for independent samples. Thus it appeared reasonable to proceed to check for possible mean differences between raters on the assumption of homogeneous variances.

The rationale guiding this investigation was basically identical to that used in the previous tests: while some of the selected extreme mean differences were significant in the independent sample portion, none of these findings could be confirmed on the basis of the respective sub-set with paired observations.

In terms of measurement theory, then, the obtained ratings got very close to the ideal of 'equivalent forms', which can legitimately be summed or averaged with a corresponding increase in 'true variance'.

4. Estimates of inter-rater reliability

Insofar as the independent ratings can be regarded as equivalent, it is justified to employ Cronbach's Alpha as an estimate of the achieved inter-rater reliability. For the special case of two such ratings, the well-known Spearman-Brown-formula may be used:

$$\text{Cronbach's Alpha}_{(K=2)} = \frac{2r_{ij}}{1+r_{ij}}$$

Since it can be shown that this statistic fits the above stated definition of reliability, the resulting numerical values give a direct indication of the proportion of true variance in the observed average scores. There are different values for each pair of raters, task, and rating dimension. In the Hamburg data, there were no consistent differences between pairs of raters or rating dimensions, but there were differences between tasks: generally, writing achievement was measured less accurately for Tasks 1A, B, C, E, and Task 6 than it was for the remaining tasks. The best values for inter-rater agreement were obtained for Task 9.

Averaging the Alpha's between the first and second rating over all tasks and rating dimensions, a mean Alpha of 0.885 was obtained. This amounts to saying that, on the average, 11.5 percent of the variance in the outcome variables (arithmetic means from two independent scores on the same essay and rating dimension) must be attributed to error.

5. Estimates of intra-rater reliability

The above estimates of inter-rater reliability do not contain any reference to the fact that there may also be a certain amount of instability within the ratings of one and the same rater over time. In order to assess this source of error, a corpus of 138 compositions from all tasks was rated twice by all raters in the Hamburg jury. Assuming again that the two ratings from a given rater were equivalent in the statistical sense of the word, Alpha estimates the proportions of true and error variance in the obtained averages over time. The resulting average Alpha was 0.939; so 6.1 percent of the variance of within-rater average scores can be associated with intra-rater instability.

When trying to separate inter-rater from intra-rater effects, a correlation-based approach is more appropriate. Assuming hypothetically that perfect intra-rater agreement could be obtained, one could correct for attenuation on the basis of the usual formula

$$\text{true score correlation } r_{i_t j_t} = \frac{r_{ij}}{\sqrt{r_{ii}\ r_{jj}}} \quad \text{for raters } i,j$$

Using again data aggregated over tasks, dimensions, and raters from the Hamburg Study, the corrected estimate for inter-rater agreement would be r = 0.843 or Alpha = 0.915. This means that an average of only 8.5 percent of the variance of outcome scores can be attributed to inter-rater differences, whereas an additional 3.0 percent out of the total error component of 11.5 percent is estimated to be due to intra-rater instability.

6. Towards generalizability - the variance components model

An obvious draw-back of considerations so far has been that these were only concerned with the measurement accuracy for single tasks and rating dimensions. No reference was made to existing relationships between tasks/within students. It may be reiterated that in the Hamburg Study, all students were asked to complete four assignments (one more than internationally obligatory). There are 1,073 students for whom two valid independent scores exist for all four assignments. Without going into details here, it may be added that in Hamburg data also allow to combine "overall impression marks" and analytical scores (except mechanics and handwriting) into a single general merit score for each composition/rater. From now on, considerations will only refer to these general merit scores.

In the conceptualization of the IEA Study, it was attempted to have a sample of tasks from the domain of school writing (Vähäpassi 1982). Pragmatic constraints led to the rotation of 4 plus 3 of the 8 international tasks for Population B (modal grade before leaving compulsory school). In spite of the existence of acceptable measures for each constituent task it is necessary to ask whether - and if so, to what extent - the outcome variables measure a stable individual trait which can then be called 'general writing ability'.

Statistically speaking, this question is closely related to the identification of within-student/across-task variation, correcting for possible mitigating influences of rater performance. An appropriate technique is given by the analysis of variance components (cf. Thorndike 1982, pp. 156 ff).

The structure of the IEA rating data makes it difficult to conduct such an analysis for all assignments simultaneously: the fact that, for instance, no student has completed both Task 5 and Task 6 has left 'empty cells' in the overall design which should not be filled with estimated values, as long as virtually nothing is known about empirical relationships between achievement in these tasks. Thus, the following incomplete matrix of inter-task correlations (based on averages from two independent ratings) was obtained (Table 1).

Table 1: Correlation matrix for task-specific scales for nine tasks (total N = 1340 students; pairwise numbers of cases in parentheses)

Task	1A	1B	1C	1E	5	6	7	9	0
1B	* (0)								
1C	* (0)	* (0)							
1E	* (0)	* (0)	* (0)						
05	.12 (102)	.24 (96)	.32 (108)	.10 (110,					
06	.19 (92)	.29 (103)	.40 (98)	.37 (98)	* (0)				
07	.43 (103)	.26 (95)	.22 (107)	.39 (101)	* (0)	* (0)			
09	.28 (313)	.34 (312)	.32 (327)	.35 (308)	.25 (403)	.31 (376)	.37 (394)		
00	.23 (290)	.37 (299)	.35 (311)	.34 (304)	.27 (389)	.34 (379)	.46 (379)	.37 (1168)	
Total	(323)	(324)	(337)	(328)	(424)	(401)	(411)	(1270)	(1231)

Therefore, it seems advisable - at least at this stage - to disregard possible differences between rotation forms and include in the analysis only assumptions that **any** of the Tasks 1A, B, C, E and **any** of the Tasks 5, 6, 7 were completed. Moreover, only the information is retained that **any** two raters from the jury of five were involved. Given the high inter-rater reliabilities achieved, little appears to be lost as a consequence of that simplification, although there are certain implications for the subsequent analysis. With these modifications, a completely balanced factorial design - or design with these **'facets'**, to use the appropriate term - emerges. Figure 2 depicts the resulting analytic design and file structure:

Figure 2:

	first rating				second rating			
Task	1A,B,C,E	5,6,7	9	0	1A,B,C,E	5,6,7	9	0
Student								
1	4	2	3	3	4	3	3	3
2	2	2	1	3	3	3	1	3
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
1073	5	3	2	4	5	3	2	4

This is very much like a conventional three-way ANOVA design with a single observation per cell, except that obviously each case (student) is spread over eight cells. It can also be viewed as a MANOVA with a two-factor within-subject design and student as the breakdown variable.

Given the three facets "rating", "task", and "student", it is now possible to define and evaluate the respective main effects as well as the interaction terms. Hereby, it has to be taken into account that only "rounds of scoring" and not the individual raters are considered; therefore, the rater main effect is confounded with the rater-by-student interaction term and similarly the rater-by-task interaction effect with the rater-by-task-by-student term. So, the following variance components are defined (numerical results from the Hamburg Study):

Variance component	Notation		Results	
	σ^2	df	σ^2	df
between students	σ^2_s	n_s-1	.180	1072
between tasks	σ^2_t	n_t-1	.015	3
between raters/ rater-by-student interaction	$\sigma^2_{r,rs}$	$(n_r-1)+(n_r-1)(n_s-1)$	-.005	1073
student-by-task interaction	σ^2_{st}	$(n_s-1)(n_t-1)$	.347	3216
rater-by-task interaction/ rater-by-student-by-task interact.	$\sigma^2_{rt,rst}$	$(n_r-1)(n_t-1)+(n_r-1)(n_s-1)(n_t-1)$	.065	3219

with n_r = number of ratings
n_s = number of students
n_t = number of tasks

It can now be seen that the combined rater-effect/rater-by-student interaction term is virtually zero, as was, indeed, expected when the scoring design for the Hamburg Study was planned: since most, if not all, raters were likely to be involved with each student in the sample, this term was likely to disappear as a consequence of the scoring scheme. Similarly, this scheme would cancel out rater-by-task interaction effects except for a possible time-related factor. So, that the last variance component is almost exclusively related to what was labelled "inter-rater disagreement" above. Fortunately, this contribution to overall variance is minor.

The fact that there is no strong between-tasks effect in the data may be undesirable from a theoretical point of view, since it leaves little room for explanations referring to differential achievement over the tasks assigned. Methodogically, it may be a consequence of a tendency among raters to score to a normal curve, but it may, of course, also reflect a more fundamental difficulty, namely that the classical concept of "item difficulty" is not easily applied to tasks of school writing.

The remaining two variance components are those which are of primary interest for later multivariate analyses. Clearly, the relatively small amount of between-students variance (as compared with the student-by-task interaction, i.e. the "within student" component) will impose limitations on the attempt to find a single overall explanation for differences between students in terms of writing achievement. Within-student variation, on the other hand, may be related to many factors which were only partially controlled in this study - e.g. fluctuations in achievement over time, varying levels of motivation, familiarity with the tasks, etc. It remains to be seen whether some of the background data of the Study will help to explain this source of variation.

It is now possible to return to the guiding question of this paper: what can be said about the reliability of measuring 'general writing achievement' across the tasks used, or - in other words - <u>generalizability</u> of composition ratings in the Study of Achievement in Written Composition? It will be seen immediately that there is not a single and simple answer; instead, the solution depends on the kind of assumptions with respect to the tasks one is prepared to make. Statistically, the answer is a function of whether within-student variation is considered as true variance or error.

Assuming that Tasks 1A, B, C, E and Tasks 5, 6, 7 are strictly equivalent statistically as well as theoretically and that the set of four assignments per student represents exactly the domain to which one wishes to generalize (fixed effects model with randomly chosen raters), the following formula can be applied to estimate the achieved generalizability:

$$\text{Generalizability I} = \frac{\sigma^2_t}{\sigma^2_\delta} = \frac{\sigma^2_s + \frac{\sigma^2_{st}}{n_t}}{\sigma^2_s + \frac{\sigma^2_{st}}{n_t} + \frac{\sigma^2_t}{n_t} + \frac{\sigma^2_{r,rs}}{n_r} + \frac{\sigma^2_{rt,rst}}{n_r n_t}}$$

On the basis of this formula, a generalizability coefficient of 0.957 would be obtained for the Hamburg data. This value appears quite appealing, but it is not a very plausible one, given the doubts about the validity of the strong underlying assumptions. In fact, the already quoted specification of task types within the domain of school writing (Vähäpassi 1982) does not treat Tasks 1A, B, C, E and 5, 6, 7 as equivalent, and it would be difficult to find an expert/teacher in the City of Hamburg who would consider the tasks used in the Study as representative for all 11th-grade school writing there. These objections alone are really sufficient to reject the model as leading to grossly inflated estimates of the achieved generalizability. So, the generalizability formula must be changed to a random-effect model, deleting the term $\sigma^2_{\varepsilon}/n_t$ (the within-student component) from the numerator:

$$\text{Generalizability II} = \frac{\sigma^2_t}{\sigma^2_\delta} = \frac{\sigma^2_s}{\sigma^2_s + \frac{\sigma^2_{st}}{n_t} + \frac{\sigma^2_t}{n_t} + \frac{\sigma^2_{r,rs}}{n_r} + \frac{\sigma^2_{rt,rst}}{n_r \cdot n_t}}$$

Here, it is only assumed that **any** four tasks from the universe of school writing were completed and scored by **any** two raters from the jury assumed to produce valid ratings. On the basis of this formula, the generalizability estimate for the obtained ratings on the Hamburg sample is reduced rather drastically to 0.646. This means that, in spite of all efforts undertaken by students as well as raters, the measurement of general writing achievement was not very good. But perhaps it is comforting to see how much more effort it would have taken to achieve a satisfactory generalizability of above 0.85: everything else being equal, a minimum of 13 writing assingments would have been required.

Bibliography:

Ferguson, G.A. (1966): Statistical Analysis in Psychology and Education. 2nd ed.. New York (McGraw-Hill).

IEA (International Association for the Evaluation of Educational Achievement) (1984): International Study of Achievement in Written Composition: Scoring guides. Urbana, Ill. (IEA/WR/A59).

IEA (International Association for the Evaluation of Educational Achievement) (1985): IEA: Activities, Institutions, People. Oxford (Pergamon).

Thorndike, R.L. (1982): Applied Psychometrics. Boston (Houghton & Mifflin).

Vähäpassi, A. (1982): On the specification of the domain of school writing. Evaluation in Education, 5, (3), pp. 265-289.

Wesdorp, H., Bauer, B.A., & Purves, A.C. (1982): Toward a conceptualization of the scoring of written composition. Evaluation in Education, 5, (3), pp. 299-315.