DOCUMENT RESUME

ED 206 640	TM 810 528
AUTHOR	Nagy, Philip; Jarchow, Elaine McNally
TITLE	Estimating Variance Components of Essay Ratings in a Complex Design.
PUB DATE	Apr 81
NOTE	58p.; Paper presented at the Annual Meeting of the American Educational Research Association (65th, Los Angeles, CA, April 13-17, 1981).
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Analysis of Variance; Essay Tests; Foreign Countries; Grade 10; High Schools; Models; *Scoring; *Test Reliability; *Writing (Composition); *Writing Evaluation; Writing Skills
IDENTIFIERS	Generalizability Theory; Iowa; Newfoundland

ABSTRACT

This paper discusses variables influencing written composition quality and how they can best be controlled to improve the reliability of assessment of writing ability. The study aimed to concentrate on the effects of different imposed time limits on student performance in written composition at grade 10 level using methods of generalizability theory; however, as the study progresses the major focus becomes methodological issues. Research studies and literature on this subject are reviewed before the design of the study is described. The sample of 320 students from Iowa and Newfoundland schools who were divided into eight groups (each group being assigned eight judges, who would assess the students' performance on two written topics under different time constraints) is presented. Scoring procedures are outlined. Discussed are the analytic methods of both Model 1 (a seven way analysis of variance model) and its limitations, and Model 2, which was adopted to analyze the variance for a confounded model incorporating elements of topic, time limit, and group. The results and limitations of the study are documented in the concluding section. Data analysis statistics are appended. (AEF)

ESTIMATING VARIANCE COMPONENTS OF ESSAY RATINGS IN A COMPLEX DESIGN

Philip Nagy
Memorial University of Newfoundland

and

Elaine McNally Jarchow
Iowa State University

A Paper Presented to
the Annual Meeting, American Educational Research Association
Los Angeles, California
1981

This research was supported by an Iowa State University Research Grant and Memorial University's Institute for Educational Research and Development. Opinions expressed are those of the authors.

INTRODUCTION

This is a report on a large scale study of factors influencing grades assigned to written composition at the grade ten level, using the methods of generalizability theory (Cronbach et al, 1972). As originally conceived, the purpose of the study was to investigate the effects of differing time limits on student writing performance, along with differences in subjects and judges and other related factors. During the evolution of the investigation, it became clear that the focus of the report would be on methodological issues of generalizability theory, rather than on substantive issues of essay writing and grading. These dual purposes, then, are interwoven throughout the report.

The assessment of writing ability has been a long standing concern of teachers and researchers in language education. Coffman, in a review of the field, found that "few of the important questions have been answered in any final way." (1971, p. 296). Major overlapping themes of research can be discerned in the literature. Bereiter and his colleagues and students (e.g., Bereiter et al, 1979), among others, are concerned with the process of writing, with a view to understanding the skills necessary to produce good written work, and to improving the teaching of written composition in the classroom. Godshalk et al (1966) focus on the indirect assessment of writing ability by devices such as multiple-choice tests of components of the writing process. The major substantive concern of this paper is not on the process of writing, but on that of grading of writing products. While the direct motivation is the understanding and improvement of the grading of composition, progress in clarifying the various influences on grading

Thanks are due to Tom Maguire, Bill Spain, and others for discussions of the analysis problems. Remaining errors are the fault of the first author.

should be of interest to those whose main concerns are other aspects of written composition.

One main purpose of this study is to investigate the relationship between writing performance under two conditions, "at home" and "exam". The goal of teaching composition is to produce students who are able to write for personal purposes in later life. Such "personal purposes", of course, will often include later academic pursuits. Classroom assessment of student ability is usually done by grading work produced under two sets of circumstances: one, over a period of several days under "at home" or "assignment" conditions, and two, within a specific time limit, under "exam" conditions. The former set of conditions is more representative of the end goal of instruction, and the assumption has traditionally been that writing ability under exam conditions is an accurate reflection of the same ability under more relaxed conditions. This assumption has not been tested in any systematic manner.

The methods of generalizability theory, involving the calculation of variance component estimates and the accompanying generalizability coefficients, have been given their primary impetus by the work of Cronbach et al (1972). Cardinet et al (1976) and Rentz (1980) have produced considerably simpler "rules-of-thumb" views of the methods, making them more accessible to researchers.

The analysis began using a complex, nested design, (Model 1) which could be viewed either as incomplete, or as a Latin Square with confounding of some of the factors of major interest. Dissatisfaction with the confounding problem lead to a reconceptualization of the design in a manner which effectively removed the confounding. This second design, dubbed Model 2, allowed estimation of the variance components of interest. Both conceptualizations of the design contained several replications of an experimental unit. Comparisons across these replications allowed examination of the stability of variance component estimates. Our results support the concerns expressed by Smith (1978).

The structure of the nesting in the design limits the types of calculations of generalizability coefficients which can be made. Thus, methodologically, the interest of the report is in the two analysis of variance models used, the solving of the confounding problem, and in the data on variance component estimate stability.

BACKGROUND

Among the important questions to which researchers in the assessment of writing ability have sought answers are:

1. Can writing ability be assessed through indirect (amenable to machine scoring) methods?

2. Is it more efficient and reliable to score essays analytically or holistically?

3. What are the variables influencing written composition quality? How are they best controlled to improve reliability of estimates of writing ability?

This last question is of direct concern to this report, but accurate estimation of writing quality is a major tool in research on the first two. Any discussion of related research cannot entirely separate the three questions.

Godshalk (1966) undertook a major study of the relationship between direct and indirect assessment of writing ability at the high school level. Although his main concern is not central here, as part of the investigation, he sought reliability estimates for scores assigned to his direct measures. Without reliable direct estimates of writing ability, his investigation of their relationship to indirect methods would be difficult. To combat the well known inconsistency of essay grades, Godshalk employed a technique of obtaining multiple writing samples from each student, and having these graded by several graders. Godshalk commented, " . . . the unreliability of essay tests came from two major sources: the differences in quality of student writing from one topic to another, and the differences among readers in what they consider the characteristics of good writing."

In his study, Godshalk used five topics per student. The essays were timed, some forty minutes and some twenty minutes, and contained both descriptive and expository topics. The readers, with no experience in reading brief essays produced by high school students, were asked to make holistic judgments (3,2 or 1). Each essay was scored by 5 readers. Some twenty-five readers rated 646 papers and achieved a reliability estimate of 0.921 for scores on the samples and 0.841 for estimates of writing ability. Godshalk's calculational methods were based on analysis of variance, (see Ebel, 1951) but did not follow the methods of generalizability theory directly. Reanalysis of the data presented in Table 1 (page 12) of Godshalk produces a generalizability coefficient of student writing ability, for five topics crossed with five readers, both random, of 0.912. Two broad generalizations from this study are: "The reliability of essay scores is primarily a function of the number of different essays and the number of different readings included." "The most efficient predictor of a reliable direct measure of writing ability is one which includes essay questions or interlinear exercises in combination with objective questions." This second conclusion, while central to Godshalk's objective, is incidental here.

Many studies (e.g., Finlayson, 1951; Vernon and Millican, 1954) have considered essay scoring procedures primarily in terms of their reliability.

In a major review, Coffman (1971) concluded that "a) different raters tend to assign different grades to the same paper; b) a single rater tends to assign different grades to the same paper on different occasions; and c) the differences tend to increase as the essay question permits greater freedom of response" (p. 277). According to Coffman, raters differ in a number of ways: a) severity; b) extent to which they distribute grades throughout the score scales; and c) understanding of relative values assigned papers. Reliability becomes greater when a small group whose backgrounds of instruction are homogeneous act as scorers.

Bond and Tamor (1979) report that while differences in judges have been well documented, little is known about the variability of student performance over different occasions. In the present context, little is known about the relationship between performance with and without the pressure of time.

A concern over scoring procedures has led researchers to distinguish between holistic and analytic scoring. Prior to this distinction, scorers were often provided a list of do's and don'ts. For example, Ahman and Glock (1975) suggest "If the spelling, penmanship, grammar, and writing style of the responses are to be scored, it should be done independently. If this is not possible, score them a second time yourself, doing so without knowledge of your first judgments." (pp. 139-141).

More recently, though, the holistic scoring method has been accepted by both practitioners and researchers. The holistic method asks the rater to read for a judgment of the whole product and assign a rating on that basis. The analytic method might set forth a series of criteria similar to those identified by McNally (1977) and suggest weightings. In similar fashion, Diederich and Link, (1967) proposed five analytic characteristics: ideas, form, flavor, mechanics, and wording.

As early as 1950, Coward suggested that it is possible to get two global ratings in the time it takes to do one analytic score and the reliability is the same. "There is some evidence, for example, that for content examinations raters are able to distinguish more than 9 quality levels without an appreciable increase in the time required for making the ratings." (Coffman, 1971, p. 295).

Powills, Bowers, and Conlan (1979) describe an Indianapolis Writer's Clinic for teachers of seventh and eighth graders. The teachers, volunteers for the clinic, participated in a holistic scoring experiment of two controlled writing assignments. Although details of their methodology can be criticised, they concluded that the holistic method was an efficient and reliable method of scoring papers. Reading at the rate of 64 papers per hour, readers achieved an average reliability (Cronbach's alpha) on the sum of three independent ratings of about 0.80. Conlan (1980) reported scorer frustration with the time needed for analytic scoring.

Although the major sources of variation among scores on written composition are inter- and intra-student and judge differences, some investigations have demonstrated the effect of other variables on score variance.

In the present study, papers were typed to remove the effects of handwriting quality. The variable of handwriting and its effect on raters was addressed by Markham (1976). She asked the question, what is the effect of quality handwriting on elementary school teachers evaluation of elementary school children's written work? After analysis of forty-five teachers' and thirty-six student teachers' evaluations, she concluded that papers with better handwritting consistently received higher scores than did those with poor handwriting regardless of quality of content. This conclusion was also drawn at the secondary level by Chase (1968), Briggs (1970), and Soloff (1973).

Chase (1979) continued this consideration of the handwriting variable by examining the effect of achievement expectancy on scores given essay tests which varied widely in quality of handwriting. He concluded that handwriting alone was not the most significant variable in scoring. In fact, when high expectancy was coupled with poor handwriting, expectancy appeared to negate the effect of poor handwriting.

Another important variable in studying writing is that of mode of discourse. Cooper (1977) contends that different writing attributes characterize varied modes of discourse. For example, developing an argument in writing takes half as much time in words per minute as reporting a personal experience.

Crowhurst (1978) concluded that argument outranks both narration and description in syntactic complexity for grades six and ten. Rosen (1969) found that fifteen and sixteen year olds produced longer T-units in referential writing than they did in expressive or personal writing. Coffman (1971) in discussing the nature of the essay test also considered the nature of task complexity when he postulated, "The more complicated the question, the more time required to compose and record an answer." (p. 280).

Shale (1978), examining different modes of discourse, discusses criteria which have been identified for analytic scoring. He identified criteria common to both exposition and description. In his study 168 twelfth grade students wrote on two topics at home (one descriptive and one expository). Four well trained markers scored the essays. The findings prompted the researcher to conclude that there are criteria that characterize one type of writing and not another. Many of the same criteria are used to evaluate expository and descriptive writing but they are often perceived differently by raters.

Generalizability theory represents a powerful tool for the investigation of phenomena which are influenced by several sources of variance. Scores on written composition are an excellent example. The essence of generalizability theory is two-fold: first, by working through equations stating the theoretical expectations for the values of mean squares (EMS) in analysis of variance, it is possible to calculate values for each variance source in the EMS; second, by assigning variance component estimates (VCE's) to either true score or error score as appropriate, estimates of the generalizability of measures across various combinations of conditions can be made. In Cronbach's (1972) terminology, a G (generalizability) study is done to obtain the VCE's. Using these values, a D (decision) study can be planned to maximize generalizability across the domain of substantive interest. The present study was intended as a G study, done more for demonstration of a method than with any particular D study in mind.

While the theoretical underpinnings of the ANOVA calculations done below are available from many sources (e.g., Winer, 1971), reliance was placed on the Millman and Glass (1967) system and terminology. Rentz (1980) has provided guidelines for separation of VCE's into true and error scores for a complex, nested design. Analysis of the confounding problems, although done independently, is similar to that of Linhart (1979).

There have been a few studies which have applied the techniques of generalizability theory to essay grading. Godshalk et al. (1966) found reliabilitics of summated ratings (five readers) of individual student products from the same student (total of 25 grades), 0.92. He also found that there was little benefit to be gained by increasing the number of readers from four to five. The most important result to be drawn from this study, which focused on the relationship of indirect to direct measures of reading ability, is that considerably more reliability can be gained by increasing the number of writing samples drawn from students than by increasing the number of readings of each writing sample.

Steele (1979) reports generalizability coefficients in agreement with those of Godshalk. Starting with a base reliability of 0.58 for two samples read by two scorers, reliability could be increased to 0.65 by adding a third sample, but to only 0.62 by adding a third scorer. Steele's study used college students, but it is not clear from the report how many subjects were the basis for the above reported values.

Rentz (1980), in a primarily methodological study, reports generalizability coefficients for three judges scoring two samples per student using analytic scoring methods. Given the differences between his study and those of Godshalk and Steele, it is difficult to draw direct comparisons. However, his reported coefficients are in the range 0.85 and above.

Following the recommendations of Coffman (1972), the present study uses holistic ratings on a more detailed rating scale than those reported above. It includes estimates of the effects on essay grades of topics, time limits, the countries of both students and judges, as well as the influences of student and judge variations, as previously investigated by others. Among the most noteworthy omissions in the list of investigated variables is the influence of scoring context, in the sense of whether a paper is preceded in the scoring pile by high or low quality products. For a discussion of this variable, see Hughes et al. (1980).

PURPOSE

This study focuses on the effects of imposed time limits on student performance in written composition. Students wrote on two topics under two time limits, but no student wrote any one topic under both time limits. If we consider the students who wrote Topic 1 under Time Limit 1 and Topic 2 under Time Limit 2 as Group 1, and those who wrote Topic 2 under Time Limit 1 and Topic 1 under Time Limit 2 as Group 2, the sources of confounding in the design can be seen. Topic, Time Limit and Group are confounded, in a manner similar to the problem encountered by Linhart (1979).

Keeping in mind the shift from an initial dual interest in both G studies and essay grading to more of an emphasis on methodological problems, the purposes of the investigation, in chronological order rather than order of importance are:

1. To produce additive components for a complex analysis of variance, first for a confounded model (Model 1);
2. To discuss the assumptions necessary to untangle the confounding, and through preliminary data analysis, to show the limitations of this approach;
3. To produce additive components for an unconfounded analysis, Model 2;
4. To calculate variance component estimates under both Models;

5. To relate the analysis of variance results to issues in essay grading;

6. To investigate the instability of reported variance component estimates;

7. To calculate example generalizability coefficients.

The order of sections following in this report requires some explanation. The complexity of the design necessitates description in order to make sense of the sampling procedure. Then, the analytic methods of Model 1 must be fully discussed so that their limitations may be seen, and the necessity for abandonment of this procedure made clear before Model 2 can be described. Results and discussion follow description of Model 2.

DESIGN

The design as originally conceived was a balanced eight-way analysis of variance, with both nested and crossed factors, and three of the eight factors confounded. Because of the confounding, the design is presented below as a seven-way ANOVA (Model 1), with a discussion of the assumptions necessary to obtain estimates of the mean square for the eighth factor and its interactions. The lack of empirical support for these assumptions is part of the reason for the introduction of Model 2.

Grade 10 students were asked to write two short compositions on different topics. Topic 1 was descriptive: "Describe trying to get to sleep with a mosquito buzzing", and topic 2 was an argument: "Argue for or against the statement 'Children over the age of fourteen should have as much say as parents in reaching decisions which affect the entire family'". The first essay written was allowed a time limit of 30 minutes, while the second was allowed a class period, plus overnight for revision. Topics and Time Limits were crossed. Students operated under the assumption that the grades would "count", and participating teachers were encouraged to use the grades provided by the judges (almost all chose to do so, with some reserving the right to remark the papers themselves). The shorter time limit will be referred to below as Limit 1 and the overnight time limit as Limit 2. Students who wrote Topic 1 under Limit 1 and Topic 2 under Limit 2 will be called Group 1, while those writing Topic 1 under Limit 2 and Topic 2 under Limit 1 will be called Group 2. Students, topics, time limits and groups are symbolized S, T, L and G respectively. Students were balanced by country of origin: Country 1 being Canada (in fact, the Province of Newfoundland only) and Country 2 the United States (in fact, the State of Iowa only).

Each written composition was typed "as is", to eliminate problems due to handwriting quality, and graded by four judges, (J), two from each country. (In further discussion, the country of student will be symbolized C, and that of the judge K). [illegible]s were nested within topics.

In order to maintain complete crossing between judges and students, and at the same time require only a reasonable marking load from each judge, another factor was introduced, duplications (D). The more traditional terms, 'replications' and 'experiments', are both used in Model 2, with different meanings. The use of 'duplications' reduces confusion. (See Smith, 1978). Figure 1 displays the design within one duplication. Thus, the eight factors are Topics, Limits, Students, Judges, Country of Judge, Country of Student, Groups (which can be thought of as order of writing) and Duplications.

[Insert Figure 1 about here]

Ten students are nested within each group and country (C), and two groups within each country. Two judges are nested within topic and country (K). Each row of the design contains the eight scores assigned to each student, and each column represents the workload, forty papers, of each judge. Thus, there are forty students and eight judges in each duplication. In total, there are eight duplications, so that the entire sample consists of 640 papers written by 320 students, given 2560 grades by 64 judges.

Since no student was asked to write a given topic under both time limits, the design could have been conceived as balanced and incomplete. Winer (1971, p. 711-713) suggests, however, that it be considered as a repeated measures Latin Square, eliminating the incompleteness, but not the confounding. This problem will be taken up again after discussion of other issues.

SAMPLE AND DATA COLLECTION

Selection of students within schools was random, but selection of schools was not. Sample schools were selected in Newfoundland first. As the population of the Province is largely rural, an attempt was made to solicit cooperation from schools in the larger centres, in order to make the samples from the two locations more comparable. Using available standardized test data from the areas in which schools were located, an attempt was made to locate Iowa schools which were approximately matched on test scores, school size, and community size. Thus, the samples are cross-sectional in nature, tending more towards rural than urban, and on balance, slightly below national norms. We make no claim that the samples are "matched".

Schools were assigned, in pairs, one from each location, to Group 1 or Group 2 treatment. In Figure 1, Groups 3 and 4 received treatment of Groups 1 and 2 respectively. The different subscripts are intended to emphasize that they are different students. Of the 22 schools used, approximately half were large enough to have more than one grade ten class. In these cases, a middle ability academic class was chosen. In four schools, two classes were used.

The data were collected and instructions given to the students by either school or board office personnel, or a research assistant. Written instructions were given to the administrators of the writing samples, and a form provided for them to report any difficulties. Four classes were lost, one in which the administrator allowed use of a dictionary, another in which the time limit was not adhered to, and two whose school administration changed their minds about participation after the first writing. As there was considerable oversampling in the data gathering phase, these losses were not a problem.

The written samples were checked, and any which gave away the author's location were eliminated from the sample. We were prepared also to eliminate those which, despite instructions, were too long to have been written in thirty minutes. As it happened, the longest selection, almost 500 words, was written in the short time limit. Thus, it was judged not necessary to eliminate any for this reason.

Eighty students (ten for each duplication) were needed to fill each of the country-group cells for the eight duplications. After removal of data as described above, and removal of single, unpaired samples, from 85 to 140 students remained for each of the cells. The design cells were filled by random selection from the data pool.

Sixty-four judges, 32 from each country, were selected through personal recommendation of school board personnel and acquaintance with the investigators. All were qualified, experienced teachers of grade ten English. Approximately 80 teachers were approached to fill the 64 vacancies. Once they had agreed, no one backed out. The teachers were paid an honorarium for their cooperation.

SCORING

Judges were given no specific instructions for scoring. In keeping with the intention of generalizing as much as possible to everyday classroom practice, they were asked to grade on the scale F, D-, D, ... A, A+ according to the implicit standards they had developed in their experience with grade ten students. Judges were given the topic title, but no information on time limits or the origins of the students in the sample. Papers were presented in order, according to the student numbers, 1-40, of Figure 1, with coding disguised. Judges were not instructed on order of marking, and most of the judges mailed back the papers in a different order than the one in which they were sent out.

Lack of information on the time limits made the situation slightly artificial, but given the circumstances, this seemed a reasonable compromise. For analysis, scores were converted to a scale of 1-13, ranging from F to A+.

ANALYSIS (MODEL 1)

The design as depicted in Figure 1 may be conceptualized as a Latin Squares design with repeated measures (Winer, 1971, page 711). Alternatively, the designation of Groups may be eliminated, and the design viewed as a balanced incomplete design, with the incompleteness resulting from the fact that no student approached the same topic under both time limits. It is this impossibility which confounds the factors, and makes the analysis less than straightforward. Temporarily ignoring factor L, Figure 2 gives additive components of the design for the other seven factors. Millman and Glass' (1967) terminology has been used. To simplify depiction of the expected mean squares of the sources, the usual "Within" and "Between" distinction has not been used. Also, in order to simplify and clarify the introduction of the time limit factor into the design, a complete model is presented. That is, all non-interpretable higher order interactions have been left in the model, and have not been pooled. While this limits the power of the design to find significant effects, such findings are not of major concern. For consistency and ease of reference, names of sources are arranged alphabetically, on each side of the colon. The only exception is in Model 1, where D (duplications) is treated as if it were in the position of R (replications) in the alphabet. This simplifies comparison of Models 1 and 2.

[Insert Figure 2 & 3 about here]

By ignoring all but the three factors involved, Figure 3 shows the relationship between G, T and L. Each is confounded with the interaction of the other two. The Group variable is not, of course, of any substantive interest, so it would be preferable if interpretations of the sources of variation could be converted from "Group" to "Time Limit" terminology. The additive model of Figure 2 is only one possibility for the analysis. As a logical alternative, Topics could have been ignored, and Limits used in designating the components. Had this been the case, our interpretation problem would involve trying to eliminate Groups in favor of Topics, rather than, as at present, in favor of Limits.

Confounding means that the factors cannot be distinguished from each other, either conceptually or analytically. While it is clear that Linhart (1978) understands this, her use of the term 'aliases' to describe the relationship between, for example, G and IT, is potentially misleading. In switching from one name to another, certain assumptions are made which can be exposed, and to some extent, analyzed. The translation is complicated by the nesting of the present design.

Figure 4 gives a breakdown of the terms which are confounded due to nesting, within the five sources from Figure 2 which involve G on the left of the colon. Applying the simple rule of eliminating G suggested by the relationships depicted in Figure 3, we can "translate" each of these terms to the expressions found on the right hand side of Figure 4. This process produces a list of sources which range from main effects (L) to six-way interactions (CJKLDT). Rational interpretation of a nested source, such as G:CD assumes that the confounded sources, GD, CG and CGD are all zero. Parallel reasoning applied to the sources on the right hand side of the table suggests that higher order interactions be allowed to "defer" to lower order interactions or main effects. This allows us to allocate variance attributed to, for example, G:CD directly to LT, under the assumption that LDG, CLT and CLDT are equal to zero.

[Insert Figure 4 about here]

The reasonableness o such assumptions can be verified, in part, by further analysis. Table 1 presents some results for an investigation of the assumptions discussed above. Temporarily assuming a completely crossed five-way design, the data was analyzed as a C by K by L by D by T design. This makes the false assumption that the groups nested within countries (C) and duplications were, in fact, composed of the same people. In this analysis, presented only for two sources drawn from Figure 4, G:CD and GT:CD, we can verify that the sums of squares add as expected, but that the assumption that higher order interactions are less than those of lower order does not hold. In fact, the Group effect, G or LT, is by chance, precisely zero. The implications are: one, that interpretations of lower order interactions and main effects are dubious at best; and two, Model 1 is only partly functional for answering the questions asked of it.

[Insert Table 1 about here]

Large sample size and luck rather than planning made abandonment of Model 1 an option. Discussion of this model will be pursued for two reasons: one, the discussion should prove beneficial to those whose circumstances do not allow alternative analyses; and two, the problem is inherently interesting.

Returning to Figure 2, five of the sources remain nested within G. In attempting to translate these, we would have to introduce complex terminology nesting students within LT interactions. Students are crossed with both L and T, so such a translation is misleading at best, and probably not permissible. Note, finally, that the confounding of T with GL exists as well.

The translation is an aid to understanding and interpretation only, and can in no way affect the calculations which follow. Thus, use of the Millman and Glass (1967) rules of thumb for EMS equations must incorporate

G rather than L, as much the Rentz (1980) rules for VCE's. As an aside, note that straightforward application of the Millman and Glass rules to Linhart's analysis problem produces her summary table.

Expected contributions to mean square values are summarized in Table 4. This Table follows Millman and Glass (1967), and assumes that S, J and D are random, while C, G, K and T are fixed. While it may seem that G should be considered a random factor, it is being considered as confounded with the interaction of two fixed factors, L and T. Rentz (1980) suggests that, for calculation of VCE's, all factors may be considered random initially, as generalizability coefficients can be calculated later under the assumption that factors are either fixed or random. However, there is no possibility of either K or C being considered random, and although Topics could logically be considered random, Winer (1971) suggests that it must be considered fixed in order that the assumptions involved in the G, L and T translation be reasonable.

[Insert Table 2 about here]

Although complex enough, Table 2 does not provide the coefficients by which each of the sources of variance are multiplied in the expectancy table. These are arrived at by multiplying together, for each term in the EMS expression, the number of levels of each factor which does not appear in the designation of the term. For example, for source 18, GK:CD, $E(MS_{GK:CD}) = jst\,\sigma^2_{GK:CD} + jt\,\sigma^2_{KS:CGD} + s\,\sigma^2_{GJ:CKDT} + \sigma^2_{JS:CGKDT}$.

Figure 5 gives appropriate calculations for both tests of significance and variance component estimates. The tests of significance are not an integral part of the approach taken in this study, but they can be easily calculated on the way to the VCE's. For both F- and quasi-F tests, the numerator becomes the additive component, and the denominator the subtractive component of the VCE. The divisor consists of the product of the number of levels of factors not included in the name of the source for which the calculation is being done. For example, for source 14, CDT, the quasi-F ratio is formed by:

$$(MS_{CDT} + MS_{JS:CGKDT})/(MS_{J:KDT} + MS_{ST:CGD})$$

and the VCE by:

$$(MS_{CDT} + MS_{JS:CGKDT} - MS_{J:KDT} - MS_{ST:CGD})/gjks$$

The value gjks equals 80.

[Insert Figure 5 about here]

Still within the confines of Model 1, an alternative method of analysis, involving a different treatment of the duplications, was considered, and judged inferior. It is possible to ignore the duplications factor, perform eight separate six-way analyses, and average the results. Such a method should, because of the averaging, yield highly stable estimates of the variance components. We report this analysis only for the light it may shed on variance component estimates, and the inner workings of complex analyses of variance.

This alternative approach is based on the additive components displayed in Figure 6. The first insight, not major in the present context because of the low priority we placed on tests of significance, is that all of the sources in Figure 5 which are tested against their interaction with D have to be approached through the use of a quasi-F test in the six-factor analysis. The second insight involves the relationship among the VCE's as calculated by the two methods. The eight extra sources in the seven-way analysis are D and the interactions of D, with combinations of the totally crossed factors, C, K and T. (L must be ignored in this discussion, which is in terms of G only). The remaining sources give identical VCE's in both analyses, while each source involving crossed factors in the six-way analysis gives a VCE equal to the sum of the corresponding VCE from the seven-way analysis plus its interaction with D. That is, for example, VCE (CK, six-way) = VCE (CK, seven-way) + VCE (CKD, seven way). The remaining VCE, D itself, directly reflects the differences in duplications.

(Insert Figure 6 about here)

It seems therefore, that there is little advantage to be gained by the averaging of the eight six-way analyses. The expected stability of such averaging turns up in the seven-way design. We not only get a direct estimate of stability in the D factor itself, but we also remove random error from our estimates of C, K, T and their interactions. Failure to remove random error from VCE's for fixed factors would, in turn, inflate generalizability coefficients.

ANALYSIS (MODEL 2)

The relationship between Model 1 and Model 2 is best understood by considering a set of four duplications, say 1 to 4. Sets of results are selected from these four duplications in the following manner: T_1L_1 from D_1; T_1L_2 from D_2; T_2L_1 from D3; and T_2L_2 from D_4, producing the data set as depicted in Figure 7. Judges and students have been given their sequence numbers in the total design, rather than within duplication, to clarify the rearrangement. Also, the reason for numbering out of sequence in Figure 1 should now be clear.

(Insert Figure 7 about here)

In this new design, Limits and Topics are completely crossed, with each of the LT combinations produced by independent sets of judges and students. The proxy, G, is no longer needed.

This arrangement (called Replication 1) has used one half of the data from one half of the students in Duplications 1 to 4. One half of the data from the other students forms a second, independent set of data similar to figure 1 (Replication 2). Thus far, we have used one half of the data from all students in Duplications 1 to 4. The other half of their data forms two more replications, independent of each other, but not indepentent of Replications 1 and 2. The non-independent partitioning will be referred to as division into Experiments 1 and 2. Thus, Duplications 1 to 4 have yielded Replications 1 and 2, independent of each other within Experiment 1, but dependent on Replications 1 and 2 of Experiment 2, which are formed by the other half of the data from the same students.

A corresponding process in Duplications 5 to 8 yields Replications 3 and 4 for each of Experiments 1 and 2. Figure 8 clarifies the reallignment of the data. The reader should note that this partitioning of the data is not unique.

(Insert Figure 8 about here)

The new design removes the confounding, allows complete crossing of L and T, and simplifies analysis. Much of the preceding discussion on Model 1 is applicable to Model 2, so that detailed discussion of Model 1 is not redundant. Figures 9 and 10, and Table 3, for Model 2, correspond to Figures 2 and 5, and Table 2 for Model 1. These give, respectively, additive components, calculations for VCE's and F-tests, and EMS expressions.

[Insert Table 3 and Figures 9 and 10 about here]

RESULTS

Consistency of Judges - The judges were quite inconsistent in their scoring, as would be expected based onthe literature. Using the 13-point scale for averaging, and then converting back to the nearest letter, the average grade awarded by a judge over the 40 papers graded varied from a 'D' to a 'B'. In average grade awarded, the 64 judges broke down as follows: D,5; D+,7; C-,13; C,18; C+,12; B-,6; B,3. These results show great differences in expectations of experienced, qualified teachers of grade ten English.

The above analysis tends to exaggerate the differences in grading among teachers. Even though the sets of students were formed randomly, the sets would be expected to differ in performance. Thus, not all of the variation reported in the above paragraph is due to inconsistent grading practices. However, sixteen sets of four judges graded the same forty papers. Within these sets, any spread across judges is a direct reflection of differing standards. In only one of the sixteen sets was the spread across all four judges near one unit on the 13-point scale. Most sets of judges had a spread of two, three, or four points, while in one set of judges the spread was five points. Given that these grades are averaged over 40 papers, the degree of inconsistency is remarkable.

Average grade awarded is only one possible way of considering consistency. Looking at the 16 sets of four judges who scored identical papers, we can produce six correlation coefficients for each set, for a total of 96. Of these, 21 are less then 0.50, 52 between 0.50 and 0.70, and 23 above 0.70. These figures give a much more optimistic picture of judge consistency, which, in summary, seems to be more a problem of difference in average grade than differences in rank-ordering.

Summary Statistics - A summary of grades awarded, ignoring the data partitions of Duplications, Replications, and Experiments, is presented in Table 4. Column 1 gives a breakdown of the scores of the entire sample (320 students X two topics X four judges/topic). The scores are positively skewed, with a mean of 5.78, between 'C- and 'C'. Considerably more low grades have been awarded than high grades. There are more grades of 'F' than 'A-', 'A' and 'A+' combined. As experience suggests, these results confirm that it is difficult for a student to score high in English.

[Insert Table 4 about here]

Columns 2 and 3 give the results broken down by time limits. Against expectations, there is a small difference in grades in favour of the short time limits. Since students were writing under the understanding that the grades would "count", this cannot be explained by suggesting that students would not bother with any effort at home merely to cooperate with a research project. The difference in grades under the two conditions is quite small, and probably of no practical significance. Such an unexpected finding will be discussed below. It would be an interesting extension of this study to see if lengthening the time limits (and the length of the writing sample) might change this tentative conclusion. For example, would grades differentiate between a one hour time limit and a one week time limit. Also, this study does not look at the length of production under different time limits.

Columns 4 and 5 give a breakdown for the two topics. There is a difference in grades, in favour of the descriptive topic. Unlike the time limit effect, inspection of thc pattern of grades shows a trend in the difference. The descriptive topic was awarded fewer grades in the range D+ to B-, and more in the B to A range. The number of very low grades was about equal in both cases.

Although "Country" has been so dubbed, we are considering students from only one state and one province. This, of course, limits generalizability of the findings. Columns 6 and 7 demonstrate that the Newfoundland students scored considerably higher than the Iowa students. The difference appears to be real, statistically significant, and large enough to be educationally important. The cause, however, is open to question. Matching of the two groups was done roughly, on the basis of school and community size, and on the basis of standardized test scores available for earlier grades in the same communitie The finding could be attributed to a poor match across countries. Or, since the matching was on scores from earlier years, the higher dropout rate in Newfoundland could lead to a better student, on average, remaining in school there. As a closely related alternative, the schools in Iowa are in most cases composite high schools. Due to widespread population in Newfoundland, a system of regional vocational schools draws vocational students away from the largely academic high schools. Again, the population in the Newfoundland schools could, on average, be more academic than in Iowa. Choosing among these explanations, and the controversial and unlikely alternative that Newfoundland education is simply better, is without further investigation, merely speculation.

On average, Newfoundland students scored more than 0.50 points higher than Iowa students, on the 13-point scale. Inspection of the table shows that they were awarded far fewer failing grades, which was compensated for by more grades in the range C- to B-. There were about equal numbers of good grades.

Columns 8 and 9 report the grades awarded by judges from the two countries. As can be seen, Iowa judges, on average, gave scores a full point higher than Newfoundland judges. The awarding of far fewer failing grades by the Iowa judges is balanced by more very good grades. Numbers of grades in the average range are about equal.

Individual variations among judges are very large. Four judges gave failing grades to 15 or more of the 40 papers they scored. All were from Newfoundland. Twenty judges gave no failing grades, and 17 of these were from Iowa. At the other end of the scale, considering A-, A, and A+ as one grade of 'A', only one judge, from Iowa, gave more than ten A's-twelve. Ten judges gave no A grades, nine from Newfoundland.

This phenomenon reflects, we believe, not so much a difference in expectations of quality of work, but in the language through which work is judged. That is, Iowa judges tend to consider a "typical" or average paper to be worth a higher grade on the F ... A+ scale. As long as the language and expectations are understood, and internally consistent, no problems arise. Problems do arise, however, when between country comparisons are made.

It is important to realize that, for the values in Table 4, judges from both countries scored essays from both countries, but were unaware of country of origin of the papers. If the Iowa papers had been marked only by the Iowa judges, scores would not be comparable with Newfoundland papers marked only by Newfoundland judges. This seems an important finding, and is worthy of further investigation.

Analysis of Variance - Before abandoning Model 1, we argued for inclusion of the Duplication effect in the ANOVA. The Duplications were independent, as are the Replications of Model 2. The Experiments of Model 2 are not independent, however, as they involve the same students' data, and thus should not be treated in the same manner as Replications. Replications, as a factor, has been included in the design of Model 2. In principle, Experiments could be included as well, as a factor distinct from Replications, but this would produce a summary table with more than 50 sources, and might lead to difficulty in interpreting VCE's due to the non-independence problem. Thus, Experiments 1 and 2 are reported separately and averaged for calculation of generalizability coefficients.

Interest in Model 1 is confined largely to the ANOVA questions raised in the preceeding discussion, but results are included for the sake of comparison. Results of the analysis of variance are reported in Tables 5, 6 and 7 for Model 1, Model 2, (E_1), and Model 2, (E_2) respectively. The fact that sums of squares for some of the crossed effects, such as C or K, do not sum in E_1 and E_2 from Model 2 to the value found in Model 1 can be attributed to the unexamined Experiment effect, and its interaction with C and K. The discussion parallels that involving the designs in Figures 2 and 6, involving omission of the Duplication factor.

(Insert Tables 5, 6 and 7 about here)

Significant effects were not consistent across the three analyses. Using a conservative level of 0.01, only the student and judge effects were significant in all three cases. Under Model 1, there were significant effects as well for K, the country of judges, and the CJ and ST interactions. Both Experiments of Model 2 showed significant effects for the JL interaction, and Model 1 for the LRT interaction.

The most disturbing characteristic of these data is the lack of consistency of results across E_1 and E_2, a random partitioning of the data. For example, the top half of Table 8 shows a breakdown of means for the LJ interaction, which was significant at the 0.01 level in both Experiments. Inspection of the cell means shows that, in fact, these two highly significant interactions are in opposite directions, on average cancel each other, and are nothing more than artifacts of this particular random splitting of the data. This is even more disturbing when one considers that either half of the data set is still quite large, larger than the data sets on which are based many conclusions in the literature.

With a smaller data set, of course, differences of the size found here might not be judged significantly different. The logical outcome of this line of reasoning is that we should be reporting effect sizes rather than levels of significance. Indeed, such a stance forms much of the basis for the estimating of variance components.

This particular split of the data is one of many possibilities. A few different partitionings of the data have been carried out in preliminary analysis, and, in every case, "significant" interactions have been found in one or the other half of the data. Only the student and judge effects have been consistent throughout this data probing. While our results seem to justify the general trend away from reporting of significant differences to reporting of effect sizes and variance component estimates, as will be seen below, little comfort can be found in the stability of the VCE's resulting from E_1 and E_2.

(Insert Table 8 about here)

The second half of Table 8 reports data on the interesting speculation that markers from each country show preference either for or against student writing from their own country. While the large C effect has been discussed above, CK cell means might tend to show some preference one way or the other. As can be seen, at least in this particular random division by E_1 and E_2, this has not occurred.

Table 9 reports VCE's for Model 1, E_1 and E_2 of Model 2, and Model 2 in total. As much as possible, sources from Model 1 and 2 have been paired. Those in brackets for Model 1 are confounded, and as has been discussed, cannot be estimated independently. Similarly, those same sources, without brackets, contain confounding. These estimates should be viewed from this perspective. Theoretically, VCE's for totally crossed factors, such as C differ in Model 1 from Model 2 by an amount equal to the interaction of that factor with the uncalculated Experiment effect, in this case, CE. Further discussion and generalizability coefficients will use Model 2 data only.

The variability in VCE's across the two Experiments of Model 2 parallels that in the ANOVA summary tables. There are, as expected, consistent estimates for J, S, and the JS interaction, although the VCE for the judge effect is considerably smaller than in Model 1. In each Experiment, there are large effects which are not duplicated in the other experiment, notably K, KLR, LRT, LR and LT (these latter two being very large negative in E_1). It should be noted that interdependence of E_1 and E_2 confounds these between Model comparisons.

While the number of negative VCE's suggests that the model may be inappropriate, and might be improved by elimination of many of the higher order interactions, this route was avoided in order to facilitate comparisons of the two models. In general, negative VCE's recur, and in those cases where a negative from one model is paired with a positive from the other, both are small. There are exceptions, as noted. While the two values for the student and judge effects vary enough that generalizability coefficients calculated on the basis of each model differ substantially, they at least are of the same order of magnitude. Generalizability coefficients will be calculated on the basis of each experiment of Model 2. Negative VCE's will be replaced by zero.

This study was reasonably well designed for estimating variance components of a great many factors and for gathering information on the stability of these estimates. However, in retrospect it appears as if the design is not amenable to the production of many generalizability coefficients which will in turn shed light on issues of essay grading. Model 1 nested judges within topics, and Model 2 nested both judges and students within topics. Thus, no comments can be made about how generalizability coefficients would vary with manipulation of number of samples and judges. A few calculations, nevertheless, can be made. Taking S:CLRT as the facet of differentiation, with K fixed, only six of the 38 sources are not included in the true score, or numerator of the generalizability coefficient (Rentz, 1980). Of these six, four are interactions of random factor J:KRT with components of true score, and the remaining two are K and J:KRT itself. This discussion assumes division by appropriate divisors in each case. The only options seem to be whether to include in the denominator J:KRT or not. If we ignore J:KRT, (and K), presumably we have an estimate of generalizability across students ignoring judge variability, that is, for the score given by an average judge, or the sum score of four judges. These values, for E_1, E_2 and the average of Model 2, are 0.95, 0.93 and 0.92. On the other hand, if we include J:KRT in the denominator, this gives a coefficient generalizing over students for the grade assigned by a random judge from a particular country. In this case, the values just quoted drop to 0.88, 0.86 and 0.87 respectively. This interpretation, representing a unique blend of Rentz, Cronbach et al. and Linhart, is open to discussion.

DISCUSSION AND CONCLUSIONS

We made no attempt to read the literature in comparative education. Thus, we dc not know the importance of the two country effects. Certainly, the country of student effect would have to be investigated thoroughly, in an investigation specifically designed to do so, before any conclusions could be drawn. On the basis of this study, the result obtained must be attributed to sample mismatch between the two countries on the basis of selective attrition, differences in standardized test validity, or some such. The country of judge phenomenon is probably more firmly supported by our data, and as such is worthy of further investigation.

Student and judge variability was, as fully expected, large. Our generalizability coefficients, interpreted as four judges scoring one writing sample, are larger than earlier reported. Given the nesting of factors in this design, we can put little faith in direct comparisons of our coefficients with those from the literature. Our design was unable to address the usual question of studies such as this, the relative efficiency of increasing writing samples or judges in estimating writing ability.

The time limit effect was one of the major objects of investigation. While this was initially one of the goals, as analysis proceeded, it diminished in importance. The time limit results are unexpected, and as such, we doubt their validity. As one reviewer of the study suggested, our treatment may not have been strong enough. Students may be so out of practice at doing homework that they would revise a paragraph at home only if it clearly was not "finished" (perhaps, as a cynic suggested, as indicated by the absence of a period after the last word). Further investigation of the time limit phenomenon should utilize a greater distinction in conditions, perhaps putting the longer, untimed conditions first in order to remove the perception that what is required in the second situation is merely a minor modification of an in-class, timed work.

The ANOVA situation encountered interfered with substantive issues, but was of interest itself. The solution proposed, Model 2, solved the confounding problems, with some sacrifice, notably loss of within student-across topic information. Also, by forcing the nesting of students within topics, it complicated already complex generalizability coefficient calculations.

We consider the most important outcomes of this study to be related to the work of Smith (1978). Smith has done substantial work on the stability of variance component estimates, and reports two general conclusions: first, that stability decreases as the complexity of the design and expected mean square expressions increases; second, stability increases as the number of levels of each source increases. By Smith's criteria, this study fails on both counts. More complex designs than this are few and far between, and most of our factors were represented by only two levels. Had we been aware of Smith's work earlier, we would have proceeded differently. However, all is not lost. As we mentioned in passing, our splitting of the data into E_1 and E_2 was done in one of many equally acceptable ways. Each split will yield two sets of variance component estimates, producing a source of data for a study of the stability of the VCE's under the circumstances of this design. Since such sets of estimates would not be independent, the product would not be as useful a tool as a Monte Carlo study, but it would have use as a generator of a distribution of VCE's around a "true" value.

Having found a way to circumvent the confounding, we are confident that other, yet to be discovered methods of cutting our data will yeild more fruitful results. While flaws in this study require that further data be collected with different designs to answer the questions posed, the potential of the present data set clearly has not yet been exhausted.

			T_1		T_2	
			K_1	K_2	K_1	K_2
			J_1 J_2	J_3 J_4	J_5 J_6	J_7 J_8
C_1	G_1	S_1 . . . S_{10}	L_1	L_1	L_2	L_2
	G_2	S_{21} . . S_{30}	L_2	L_2	L_1	L_1
C_2	G_3	S_{11} . . . S_{20}	L_1	L_1	L_2	L_2
	G_4	S_{31} . . . S_{40}	L_2	L_2	L_1	L_1

Figure 1[1]

One Duplication of the Design, Model 1

[1] Students are numbered out of sequence to assist in conversion to Model 2.

Additive Components, Model 1
(Without "Limits")

Source	degrees of freedom
C	1
K	1
D	7
T	1
CK	1
CD	7
CT	1
KD	7
KT	1
DT	7
G:CD	16
CKD	1
CKT	7
CDT	7
KDT	7
S:CGD	288
J:KDT	32
GK:CD	16
GT:CD	16
CKDT	7
CJ:KDT	32
KS:CGD	288
ST:CGD	288
GKT:CD	16
GJ:CKDT	64
KST:CGD	288
JS:CGKDT	1152
TOTAL	2559

Figure 2

	T_1	T_2
G_1	L_1	L_2
G_2	L_2	L_1

	L_1	L_2
G_1	T_1	T_2
G_2	T_2	T_1

	L_1	L_2
T_1	G_1	G_2
T_2	G_2	G_1

This diagram ignores factors C, J, K, D, S

Relationship Between "Group", "Limit" and "Topic" Effects

Figure 3

Relationship between "Groups"
and "Limits", Model 1

Source	Confounded Sources	Translation	Confounded (assumed = 0)
G:CD	G	LT	
	GD		LDT
	CG		CLT
	CGD		CLDT
GK:CD	GK	KLT	
	CGK		CKLT
	GKD		KLDT
	CGKD		CKLDT
GT:CD	GT	L	
	CGT		CL
	GDT		DL
	CGDT		CLD
GKT:CD	GKT	KL	
	CGKT		CKL
	GKDT		KLD
	CGKDT		CKLD
GJ:CKDT	GJ		JLT
	CGJ		CJLT
	GJK		JKLT
	GJD		JLDT
	GJT	JL:KDT	
	CGJK		CJKLT
	CGJD		CJLDT
	CGJT		CJL
	GJKD		JKLDT
	GJKT		JKL
	GJDT		JLD
	CGJKD		CJKLDT
	CGJKT		CJKL
	CGJDT		CJLD
	GJKDT		JKLD
	CGJKDT		CJKLD

Figure 4

Tests of Significance and VCE Denominators
Model 1

Source	MS's, Numerator	MS's, Denominator	VCE Denominator
1 C	C	CD	1280
2 K	K	KD	1280
3 D	D, JS:CGKDT	S:CGD, J:KDT	320
4 T	T	DT	1280
5 CK	CK	CKD	640
6 CD	CD, JS:CGKDT	S:CGD, CJ:KDT	160
7 CT	CT	CTD	640
8 KD	KD, JS:CGKDT	J:KDT, KS:CGD	160
9 KT	KT	KDT	640
10 DT	DT, JS:CGKDT	J:KDT, ST:CGD	160
11 G:CD	G:CD, JS:CGKDT	S:CGD, GJ:CKDT	80
12 CKD	CKD, JS:CGKDT	CJ:KDT, KS:CGD	80
13 CKT	CKT	CKDT	320
14 CDT	CDT, JS:CGKDT	J:KDT, ST:CGD	80
15 KDT	KDT, JS:CGKDT	J:KDT, KST:CGD	80
16 S:CGD	S:CGD	JS:CGKDT	8
17 J:KDT	J:KDT	JS:CGKDT	40
18 GK:CD	GK:CD, JS:CGKDT	KS:CGD, GJ:CKDT	40
19 GT:CD	GT:CD, JS:CGKDT	ST:CGD, GJ:CKDT	40
20 CKDT	CKDT, JS:CGKDT	CJ:KDT, KST:CGD	40
21 CJ:KDT	CJ:KDT	JS:CGKDT	20
22 KS:CGD	KS:CGD	JS:CGKDT	4
23 ST:CGD	ST:CGD	JS:CGKDT	4
24 GKT:CD	GKT:CD, JS:CGKDT	GJ:CKDT, KST:CGD	20
25 GJ:CKDT	GJ:CKDT	JS:CGKDT	10
26 KST:CGD	KST:CGD	JS:CGKDT	2
27 JS:CGKDT	JS:CGKDT		1

Figure 5

Alternative 6 Way Analysis

Source	d.f.
C	1
K	1
T	1
CK	1
CT	1
KT	1
G:C (LT)	2
CKT	1
S:CG	36
J:KT	4
GK:C (KLT)	2
GT:C (L)	2
CJ:KT	4
KS:CG	36
ST:CG	36
GKT:C (KL)	2
GJ:CKT (JL:KT)	8
KST:CG	36
JS:CGKT	144
	319

Figure 6

		T_1		T_2	
		K_1	K_2	K_1	K_2
		J_1 J_2	J_1 J_2	J_{21} J_{22}	J_{23} J_{24}
L_1	C_1	S_{1-10}	S_{1-10}	$S_{101-110}$	$S_{101-110}$
	C_2	S_{11-20}	S_{11-20}	$S_{111-120}$	$S_{111-120}$
		J_9 J_{10}	J_{11} J_{12}	J_{29} J_{30}	J_{31} J_{32}
L_2	C_1	S_{61-70}	S_{61-70}	$S_{121-130}$	$S_{121-130}$
	C_2	S_{71-80}	S_{71-80}	$S_{131-140}$	$S_{131-140}$

Figure 7[1]

One Replication of one Experiment, (E_1) based on one quarter of Duplications 1 to 4, Model 2

D	T	L	R_1	R_2	R_3	R_4	R_1	R_2	R_3	R_4
1	1	1	X							
		2		X						
	2	1						X		
		2					X			
2	1	1		X						
		2	X							
	2	1					X			
		2						X		
3	1	1						X		
		2					X			
	2	1	X							
		2		X						
4	1	1					X			
		2						X		
	2	1		X						
		2	X							
5	1	1			X					
		2				X				
	2	1								X
		2							X	
6	1	1				X				
		2			X					
	2	1							X	
		2								X
7	1	1								X
		2							X	
	2	1			X					
		2				X				
8	1	1							X	
		2								X
	2	1				X				
		2			X					

Figure 8

Conversion of Model 1 to Model 2

Additive Components

Model 2

Source	df
C	1
K	1
L	1
R	3
T	1
CK	1
CL	1
CR	3
CT	1
KL	1
KR	3
KT	1
LR	3
LT	1
RT	3
CKL	1
CKR	3
CKT	1
CLR	3
CLT	1
CRT	3
KLR	3
KLT	1
KRT	3
LRT	3
J:KRT	16
CKLR	3
CKLT	1
CKRT	3
CLRT	3
KLRT	3
S:CLRT	288
CJ:KRT	16
JL:KRT	16
CKLRT	3
KS:CLRT	288
CJL:KRT	16
JS:CKLRT	576
	1279

Figure 9

Tests of Significance and VCE Denominators
Model 2

Source	MS's, Numerator	MS's, Denominator	VCE Denominator
1 C	C, CRT	CR, CT	640
2 K	K, KRT	KR, KT	640
3 L	L, LRT	LR, LT	640
4 R	R	RT	320
5 T	T	RT	640
6 CK	CK, CKRT	CKR, CKT	320
7 CL	CL, CLRT	CLR, CLT	320
8 CR	CR	CRT	160
9 CT	CT	CRT	160
10 KL	KL, KLRT	KLR, KLT	320
11 KR	KR	KRT	160
12 KT	KT	KRT	320
13 LR	LR	LRT	160
14 LT	LT	LRT	320
15 RT	RT, JS:CKLRT	J:KRT, S:CLRT	160
16 CKL	CKL, CKLRT	CKLR, CKLT	160
17 CKR	CKR	CKRT	80
18 CKT	CKT	CKRT	160
19 CLR	CLR	CLRT	80
20 CLT	CLT	CLRT	160
21 CRT	CRT, JS:CKLRT	S:CLRT, CJ:KRT	80
22 KLR	KLR	KLRT	80
23 KLT	KLT	KLRT	160
24 KRT	KRT, JS:CKLRT	J:KRT, KS:CLRT	80
25 LRT	LRT, JS:CKLRT	S:CLRT, JL:KRT	80
26 J:KRT	J:KRT	JS:CKLRT	40
27 CKLR	CKLR	CKLRT	40
28 CKLT	CKLT	CKLRT	80
29 CKRT	CKRT, JS:CKLRT	CJ:KRT, KS:CLRT	40
30 CLRT	CLRT, JS:CKLRT	S:CLRT, CJL:KRT	40
31 KLRT	KLRT, JS:CKLRT	JL:KRT, KS:CLRT	40
32 S:CLRT	S:CLRT	JS:CKLRT	4
33 CJ:KRT	CJ:KRT	JS:CKLRT	20
34 JL:KRT	JL:KRT	JS:CKLRT	20
35 CKLRT	CKLRT, JS:CKLRT	KS:CLRT, CJL:KRT	20
36 KS:CLRT	KS:CLRT	JS:CKLRT	2
37 CJL:KRT	CJL:KRT	JS:CKLRT	10
38 JS:CKLRT	JS:CKLRT		1

Figure 10

Table 1

Investigation of Translation Assumptions, Model 1

Group Version				Limit Version			
Source	SS	df	MS	Source	SS	df	MS
G:CD	336.4	16	21.0	LT	0.0	1	0.0
				LDT	210.7	7	30.1
				CLT	21.8	1	21.8
				CLDT	103.9	7	14.8
					336.4		
GT:CD	195.9	16	12.2	L	9.5	1	9.5
				CL	10.0	1	10.0
				LD	64.0	7	9.1
				CLD	112.4	7	16.1
					195.9		

Table 2
Expected Mean Squares, Model 1[1]

Source No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
1 C	X						X									X					X						X
2 K		X							X								X					X					X
3 T			X							X							X						X				X
4 D				X												X	X										X
5 CK					X								X								X	X					X
6 CT						X								X							X		X				X
7 CD							X									X					X						X
8 KT								X							X		X									X	X
9 KD									X								X					X					X
10 DT										X							X						X				X
11 G:CD											X					X									X		X
12 CKT												X								X	X					X	X
13 CKD													X								X	X					X
14 CDT														X							X		X				X
15 KDT															X		X									X	X
16 S:CGD																X											X
17 J:KDT																	X										X
18 GK:CD																		X				X			X		X
19 GT:CD																			X				X		X		X
20 CKDT																				X	X					X	X
21 CJ:KDT																					X						X
22 KS:CGD																						X					X
23 ST:CGD																							X				X
24 GKT:CD																								X	X	X	X
25 GJ:CKDT																									X		X
26 KST:CGD																										X	X
27 JS:CGKDT																											X

[1] S, J, and D random, C, G, K, T fixed

Table 3
Expected Mean Squares, Model 2[1]

Source No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
1 C	X							X	X												X											X	X					X
2 K		X									X	X												X		X										X		X
3 L			X										X	X											X							X		X				X
4 R				X											X											X						X						X
5 T					X										X											X						X						X
6 CK						X											X	X											X				X			X		X
7 CL							X												X	X										X		X						X
8 CR								X													X											X	X					X
9 CT									X												X											X	X					X
10 KL										X												X	X								X			X		X		X
11 KR											X													X		X										X		X
12 KT												X												X		X										X		X
13 LR													X												X							X		X				X
14 LT														X											X							X		X				X
15 RT															X											X						X						X
16 CKL																X											X	X							X	X		X
17 CKR																	X												X				X			X		X
18 CKT																		X											X				X			X		X
19 CLR																			X											X		X					X	X
20 CLT																				X										X		X					X	X
21 CRT																					X											X	X					X
22 KLR																						X									X			X		X		X
23 KLT																							X								X			X		X		X
24 KRT																								X		X										X		X
25 LRT																									X							X		X				X
26 J:KRT																										X												X
27 CKLR																											X								X	X	X	X
28 CKLT																												X							X	X	X	X
29 CKRT																													X				X			X		X
30 CLRT																														X		X					X	X
31 KLRT																															X			X		X		X
32 S:CLRT																																X						X
33 CJ:KRT																																	X					X
34 JL:KRT																																		X				X
35 CKLRT																																			X	X	X	X
36 KS:CLRT																																				X		X
37 CJL:KRT																																					X	X
38 JS:CKLRT																																						X

[1] R, J, S, and T random, C, K, and L fixed.

TABLE 4

Summary of Scores Awarded
(% of Total Sample)

	1	2	3	4	5	6	7	8	9
		SHORT TIME LIMIT	LONG TIME LIMIT	TOPIC 1	TOPIC 2	NFLD. STUDENTS	IOWA STUDENTS	NFLD. JUDGES	IOWA JUDGES
SCORE	N=2560	N=1280	N=1280	N=1280	N=1280	N=1280	N=1280	N=1280	N=1280
1-F	10.2	9.5	10.9	9.2	11.2	6.6	13.8	16.3	4.1
2-D-	6.6	6.3	7.0	7.6	5.6	6.0	7.2	8.1	5.1
3-D	10.3	10.5	10.1	10.3	10.3	10.5	10.1	9.3	11.5
4-D+	8.2	9.1	7.4	7.7	8.8	8.3	8.2	9.1	7.3
5-C-	13.1	12.3	13.8	12.7	13.4	13.6	12.6	12.2	14.0
6-C	11.7	12.5	10.9	10.9	12.4	11.7	11.6	10.9	12.5
7-C+	9.5	8.5	11.2	9.5	9.4	10.9	8.0	8.5	10.4
8-B-	9.6	10.2	9.1	9.3	10.0	10.5	8.8	8.6	10.7
9-B	7.6	7.2	8.0	8.4	6.9	8.4	6.9	6.3	8.9
10-B+	5.1	5.0	5.2	5.9	4.3	5.5	4.7	5.1	5.2
11-A-	4.3	4.8	3.8	4.4	4.2	4.2	4.4	2.7	5.9
12-A	3.4	3.7	3.1	3.8	3.0	3.5	3.3	2.6	4.2
13-A+	0.4	0.4	0.3	0.3	0.4	0.2	0.5	0.3	0.4
Average	5.78	5.85	5.72	5.88	5.69	6.04	5.53	5.24	6.34

Table 5

Summary of Anova, Model 1

Source	Mean Square	F-ratio	Quasi-F-ratio	df(test)
1 C	171.1891	3.15		1,7
2 K	776.6016	13.06[1]		1,7
3 D	139.8123		1.57	7,75
4 T	23.6396	0.29		1,7
5 CK	10.7641	2.18		1,7
6 CD	54.4533		1.48	8,305
7 CT	9.7516	2.53		1,7
8 KD	59.4480		1.01	8,36
9 KT	159.0016	1.76		1,7
10 DT	82.8266		1.28	7,42
11 G:CD(LT)	21.0227		0.66	21,337
12 CKD	4.9337		0.79	17,74
13 CKT	9.2641	0.97		1,7
14 CDT	3.8516		0.46	21,142
15 KDT	90.1444		1.52	7,35
16 S:CGD	32.3617	11.26[1]		288,1152
17 J:KDT	58.4836	20.34[1]		32,1152
18 GK:CD(KLT)	4.4555		1.04	43,207
19 GT:CD(L)	12.2461		1.28	24,321
20 CKDT	9.4980		1.35	12,64
21 CJ:KDT	6.4180	2.23[1]		32,1152
22 KS:CGD	3.4841	1.21		288,1152
23 ST:CGD	8.2749	2.88[1]		288,1152
24 GKT:CD(KL)	4.3289		1.14	44,178
25 CJ:CKDT(JL:KDT)	3.5617	1.24		64,1152
26 KST:CGD	2.7716	0.96		288,1152
27 JS:CGKDT	2.8752			

[1] significant at the 0.01 level

TABLE 6

SUMMARY OF ANOVA, MODEL 2, E_1

		Mean Square	F-ratio	Quasi-F-ratio	df(test)
1	C	63.457		5.21	2,4
2	K	192.976		1.26	1,2
3	L	1.313		9.48	3,3
4	R	9.149	0.38		3,3
5	T	1.582	0.07		1,3
6	CK	12.601		2.40	1,3
7	CL	14.238		0.78	4,4
8	CR	15.978	0.41		3,3
9	CT	3.720	0.10		1,3
10	KL	6.757		0.54	4,3
11	KR	52.942	1.61		3,3
12	KT	125.626	3.83		1,3
13	LR	34.047	0.10		3,3
14	LT	0.488	0.00		1,3
15	RT	24.145		0.56	4,43
16	CKL	0.132		0.43	3,4
17	CKR	5.472	5.91		3,3
18	CKT	0.176	0.19		1,3
19	CLR	34.884	1.59		3,3
20	CLT	11.438	0.52		1,3
21	CRT	39.103		1.77	4,225
22	KLR	28.640	2.91		3,3
23	KLT	2.195	0.22		1,3
24	KRT	32.800		1.10	4,20
25	LRT	326.092		6.38[1]	3,40
26	J:KRT	29.702	9.02[1]		16,576
27	CKLR	7.170	1.46		3,3
28	CKLT	4.632	0.94		1,3
29	CKRT	0.926		0.56	58,46
30	CLRT	21.947		1.00	4,191
31	KLRT	9.832		0.37	5,19
32	S:CLRT	19.489	5.92[1]		288,576
33	CJ:KRT	4.409	1.34		16,576
34	JL:KRT	32.121	4.75[1]		16,576
35	CKLRT	4.915		0.93	8,38
36	KS:CLRT	3.137	6.95	0.95	288,576
37	CJL:KRT	5.652	1.72		16,576
38	JS:CKLRT	3.294			

[1] significant at the 0.01 level

TABLE 7

		Mean Squares	F-ratio	Quasi-F-ratio	df(test)
1	C	111.038		1.79	1,2
2	K	651.226		5.61	1,4
3	L	10.332		1.98	4,3
4	R	23.965	0.62		3,3
5	T	31.563	0.82		1,3
6	CK	1.188		0.11	4,3
7	CL	0.488		1.05	3,4
8	CR	22.688	12.19		3,3
9	CT	40.257	21.63		1,3
10	KL	3.301		0.18	4,3
11	KR	73.597	8.87		3,3
12	KT	43.882	5.29		1,3
13	LR	53.353	0.55		3,3
14	LT	0.488	0.01		1,3
15	RT	38.659		0.91	3,55
16	CKL	0.063		0.57	3,3
17	CKR	12.259	6.71		3,3
18	CKT	15.095	8.26		1,3
19	CLR	23.051	0.67		3,3
20	CLT	10.332	0.30		1,3
21	CRT	1.861		0.61	16,220
22	KLR	121.076	6.34		3,3
23	KLT	2.720	0.14		1,3
24	KRT	8.300		0.40	3,37
25	LRT	96.238		1.62	3,37
26	J:KRT	23.868	9.72[1]		16,576
27	CKLR	9.455	1.71		3,3
28	CKLT	0.282	0.05		1,3
29	CKRT	1.828		0.53	16,41
30	CLRT	34.557		1.50	3,263
31	KLRT	19.099		0.50	4,19
32	S:CLRT	21.147	8.61[1]		288,576
33	CJ:KRT	4.968	2.02		16,576
34	JL:KRT	39.865	16.22[1]		16,576
35	CKLRT	5.528		1.21	
36	KS:CLRT	3.118	1.27		288,576
37	CJL:KRT	3.465	1.41		16,576
38	JS:CKLRT	2.456			—

[1] significant at the 0.01 level

TABLE 8

SELECTED MEANS FROM E_1 AND E_2

	E_1		E_2		TOTAL	
	J_1	J_2	J_1	J_2	J_1	J_2
L_1	5.88	5.68	5.99	5.83	5.94	5.75
L_2	6.03	5.40	5.61	5.94	5.82	5.67

	E_1		E_2		TOTAL	
	K_1	K_2	K_1	K_2	K_1	K_2
C_1	5.68	6.26	5.44	6.82	5.56	6.54
C_2	5.03	6.03	4.79	6.25	4.91	6.14

TABLE 9

SUMMARY OF VARIANCE COMPONENT ESTIMATES

	Sources (Model 2)	Sources (Model 1)	Model 1	Model 2(E_1)	Model 2(E_2)	Model 2(Avg)
1	C	C	0.0912	0.1295	0.0781	0.1038
2	K	K	0.5603	0.0738	0.8469	0.4604
3	L	GT:CD	0.0821	0.4576	0.0824	0.2700
4	R	D	0.1620	-0.0469	-0.0459	-0.0464
5	T	T	-0.0462	-0.0353	-0.0111	-0.0232
6	CK	CK	0.0091	0.0246	-0.0761	-0.0258
7	CL	(GT:CD)		-0.0317	0.0052	-0.0133
8	CR	CD	0.1159	-0.1445	0.1302	-0.0072
9	CT	CT	0.0092	-0.2211	0.2400	0.0095
10	KL	CKT:CD	0.0435	-0.0445	-0.3169	-0.1897
11	KR	KD	0.0022	0.1259	0.4081	0.2670
12	KT	KT	0.1076	0.2901	0.1112	0.2007
13	LR	(GT:CD)		-1.8253	-0.2680	-1.0467
14	LT	G:CD	-0.1503	-1.0175	-0.2992	-0.6584
15	RT	DT	0.1184	-0.1360	-0.0244	-0.0802
16	CKL	(GKT:CD)		-0.0422	-0.0259	-0.0341
17	CKR	CKD	-0.0262	0.0568	0.1304	0.0936
18	CKT	CKT	-0.0007	-0.0047	0829	0.0391
19	CLR	(GT:CD)		0.1617	-0.1438	0.0090
20	CLT	(GT:CD)		-0.0657	-0.1514	-0.1085
21	CRT	CDT	-0.0996	0.2312	-0.2725	-0.0207
22	KLR	(GKT:CD)		0.2351	1.2747	0.7549
23	KLT	GK:CD	0.0071	-0.0477	-0.1024	-0.0750
24	KRT	KDT	0.3971	0.0407	-0.2029	-0.0811
25	LRT	(G:CD)		3.4722	0.4710	1.9716
26	J:KRT	J:KDT	1.3902	0.6602	0.5352	0.5977
27	CKLR	(GKT:CD)		0.0564	0.0982	0.0773
28	CKLT	(GK:CD)		-0.0035	-0.0656	-0.0346
29	CKRT	CKDT	0.0796	-0.0832	-0.0950	-0.0891
30	CLRT	(G:CD)		0.0025	0.3100	0.1563
31	KLRT	(GK:CD)		-0.5533	-0.5357	-0.5445
32	S:CLRT	S:CGD	3.6858	4.0488	4.6727	4.3608
33	CJ:KRT	CJ:KDT	0.1771	0.0558	0.1256	0.0907
34	JL:KRT	CJ:CKDT	0.0687	1.4414	1.8704	1.6559
35	CKLRT	(GK:CD)		-0.0290	0.0701	0.0206
36	KS:CLRT	KS:CGD	0.1522	-0.0785	0.3309	0.1262
37	CJL:KRT	(GK:CKDT)		0.2358	0.1009	0.1684
38	JS:CKLRT	JS:GKDT	2.8752	3.2940	2.4563	2.8752

References

Ahman, J.S. & M.D. Glock. Evaluating Pupil Grcwth. Boston: Allyn & Bacon, 1975.

Bereiter, C. et al. An applied cognitive-developmental apprpa[illegible] to writing research. Presented to the American Educational Research Association, San Francisco, 1979.

Bond, J.T. & L. Tamor. Determining the validity and dependability of writing tests. Presented to the annual meeting, American Educational Research Association, San Francisco, 1979.

Briggs, D. Influence of handwriting on assessment. Educational Research, 1970, 13, 50-55.

Cardinet, J. et al. The symmetry of [illegible]eralizability theory: applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135.

Chase, C.I. The impact of some obvious variables on essay test scores. Journal of Educational Measurement, 1968, 5, 315-318.

Chase, C.I. The impact of achievement expectations and handwriting quality on scoring essay tests Journal of Educational Measurement, 1979. 16, 39-42

Coffman, W.E. Essay examinations. In R.L. Thorndike (ed.) Educational Measurement Washington: American Council on Education, 1971.

Coffman, W.E. On the reliability of ratings of essay examinations. Measurement in Education, 1972, 3(3).

Conlan, G. Comparison of analytic and holistic scoring techniques. Presented to the annual meeting, American Educational Research Association, Boston, 1980.

Cooper, C.R. Holistic evaluation of writing. In C.R.Cooper & L. Odell (eds.) Evaluating Writing. Urbana, Ill.: National Council of Teachers of English, 1977.

Cronbach, L.J. et al. The Dependability of Behavioral Measures. Toronto: Wiley, 1972.

Crowhurst, M. The effect of audience and mode of discourse on the syntactic complexity of written composition at two grade levels. Presented to the annual meeting, Canadian Educational Research Association, London, 1978.

Diederich, P.B. & F.R. Link. Cooperative evaluation in English. In F. T. Wilhelms (ed.) *Evaluation as Feedback and Guide* Washington: Association for Supervision and Curriculum Development, 1967.

Ebel, R.L. Estimation of the reliability of ratings. *Psychometrika*, 1951, *16*, 407-424.

Finlayson, D_S. The reliability of marking essays. *British Journal of Educational Psychology*, 1951, *21*, 126-134.

Godshalk, F.I. *et al*. *The Measurement of Writing Ability*. New York: College Entrance Examination Board, 1966.

Hughes, D.C. *et al*. The influence of context position and scoring method on essay scoring. *Journal of Educational Measurement*, 1980, *17*, 131-136.

Linhart, C.A. Application of generalizability theory to a complex rating situation. Presented to the annual meeting, American Educational Research Association, San Francisco, 1979.

Markham, L.R. Influence of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 1976, *13*, 277-284.

McNally, E.F. A study to determine through content analysis selected criteria for open-ended examinations. Presented to the annual meeting, American Educational Research Association, Toronto, 1978.

Millman, J. & G.V. Glass. Rules of thumb for writing the Anova table. *Journal of Educational Measurement*, 1967, *4(2)*, 41-51

Powills, J.A. *et al*. Holistic essay scoring: an application of the model for the evaluation of writing ability and the measurement of growth in writing ability over time. Presented to the annual meeting, American Educational Research Association, San Francisco, 1979.

Rentz, R.R. Rules of thumb for estimating reliability coefficients using generalizability theory. Educational and Psychological Measurement, 1980, 40, 575-592.

Rosen, H. An investigation of the effects of differentiated writing assignments on the performance in English composition of a selected group of 15/16-year-old pupils. University of London, Unpublished doctoral dissertation, 1969.

Shale, D. A factorial analysis of evaluations of English composition. Presented to the annual meeting, National Council on Measurement in Education, Toronto, 1978.

Smith, P.L. Sampling errors of variance components in small sample multifacet generalizability studies. Journal of Educational Statistics, 1978, 3, 319-346.

Soloff, S. Effect of non-content factors on the grading of essays. Graduate Research in Education and Related Disciplines, 1973, 6, 44-54.

Steele, J.M. The assessment of writing proficiency via qualitative ratings of writing samples. Presented to the annual meeting, National Council on Measurement in Education, San Francisco, 1979.

Vernon, P.E. & G. D. Millican. A further study of the reliability of English essays. British Journal of Statistical Psychology, 1954, 7, 65-74.

Winer, B.J. Statistical Principles in Experimental Design (2nd ed.) New York: McGraw-Hill, 1971.