DOCUMENT RESUME

ED 274 673 TM 860 527

AUTHOR Braun, Henry I.
TITLE Calibration of Essay Readers. Program Statistics Research. Final Report.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-86-9; ETS-TR-86-68
PUB DATE Jun 86
NOTE 63p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Essay Tests; Estimation (Mathematics); *Mathematical Models; *Research Design; Research Methodology; Sampling; *Scoring; Secondary Education; Statistical Studies; *Test Reliability
IDENTIFIERS Advanced Placement Examinations (CEEB); *Calibration

ABSTRACT

This report describes a statistically designed experiment that was carried out in an operational setting to determine the contributions of different sources of variation to the unreliability of scoring. The experiment made novel use of partially balanced incomplete block designs that facilitated the unbiased estimation of certain main effects without requiring readers to assess the same paper several times. In addition, estimates were obtained of the improvement in reliability that results from removing variability from systematic sources of variation by an appropriate adjustment of the raw scores. This statistical calibration appears to be a cost-effective approach to enhancing scoring reliability when compared to simply increasing the number of readings per paper. The results of the experiment also provide a framework for examining other, simpler calibration strategies. One such strategy is briefly examined. (Author/JAZ)

RR-86-9

Calibration of Essay Readers
Final Report

Henry I. Braun

PROGRAM STATISTICS RESEARCH

TECHNICAL REPORT NO. 86-68

EDUCATIONAL TESTING SERVICE
PRINCETON, NEW JERSEY 08541

Calibration of Essay Readers
Final Report

Henry I. Braun

Program Statistics Research
Technical Report No. 86-68

Research Report No. 86-9

June 1986

Funding for this research was provided by the Program Research Planning Council of the Educational Testing Service.

The Program Statistics Research Technical Report Series is designed to make the working papers of the Research Statistics Group at Educational Testing Service generally available. The series consists of reports by the members of the Research Statistics Group as well as their external and visiting statistical consultants. Reproduction of any portion of a Program Statistics Research Technical Report requires the written consent of the author(s).

Table of Contents

Acknowledgements

The success of a complex project depends on the efforts of many individuals. This project was no exception. Carl Haag, Lucy Haagen and Lou Kremer of Advanced Placement (AP) Program Direction provided encouragement and cheerful cooperation whenever it was needed, as did Professor Alden Vaughn (Columbia) and Professor Charles Long (Yale), chief readers for American History and English Literature and Composition, respectively. Dolores Wunder and her staff did an excellent job of photocopying and sorting. Linda DeLauro, Jan Flaugher and Faustino Romero did most of the field work and were instrumental in getting everything to run smoothly. John Young assisted in the early parts of the data analyses. Special thanks to Jim Ferris who contributed to the selection of the design and assisted in the data analyses. Paul Rosenbaum provided many helpful comments on an earlier draft. Last, but not least, thanks to sixty-nine readers who participated in the experiment and whose efforts provided the raw material for the analysis.

Abstract

Scoring reliability of essays and other free response questions is of considerable concern. This report describes a statistically designed experiment that was carried out in an operational setting to determine the contributions of different sources of variation to the unreliability of scoring. The experiment made novel use of partially balanced incomplete block designs that facilitated the unbiased estimation of certain main effects without requiring readers to assess the same paper several times. In addition, estimates were obtained of the improvement in reliability that result from removing variability from systematic sources of variation by an appropriate adjustment of the raw scores. This statistical calibration appears to be a cost-effective approach to enhancing scoring reliability when compared to simply increasing the number of readings per paper. The results of the experiment also provide a framework for examining other, simpler calibration strategies. One such strategy is briefly examined.

1. Introduction

The purpose of this study was to investigate the feasibility and efficacy of a new approach to enhancing the scoring reliability of essays and open-ended questions in general. Scoring reliability is less than unity through the action of different sources of variability that arise in carrying out the grading. Some sources can be termed idiosyncratic: a particular reader may think highly of a particular essay that most other readers would rate rather low. Other sources can be termed systematic: one reader mav consistently give higher grades than another reader or grades given on one day may be consistently lower than those given on another day.

The new approach involves effectively removing variability from systematic sources by appropriately adjusting scores. The information for the adjustments comes from a statistically designed experiment embedded within the operational grading process. The data from the experiment can be employed in a variance components analysis that yields estimates of the relative contributions of the different sources of variability to the loss in reliability. Moreover, this same data can be used to estimate an upper bound for the reliability that can be attained through calibration.

Two examinations in the Advanced Placement (AP) Program, American History and English Literature and Composition, were chosen to field test the new methods. Both examinations have a substantial essay component and scores are based on a single reading. While great care is taken to maintain uniform standards across readers and through the five or six day grading period, the single reading reliability for one essay score is typically less than 0.5. (Estimates of the single reading reliability are periodically obtained by the AP program through the double reading of a randomly selected subset of papers.) Thus, improved reliability is both possible and desirable.

The results of the experiment and subsequent analysis indicate that systematic sources of variability do indeed contribute substantially to the unreliability of essay scores and that the proposed method is successful in eliminating most of their effect. However, most of the unreliability is due to idiosyncratic sources and it is estimated that this calibration procedure can recover at most some thirty percent of the reading reliability lost in moving from two readings to one. This improvement in reliability, which holds for both single essays and total essay score, is obtained with only a five percent increase in effort; i.e., one-twentieth of the effort involved in doubling the number of readers. Thus, calibration of essay scores promises to be a feasible and cost-effective way of enhancing reliability. It appears also that a calibration based solely on the results of the operational grading works as well as that based on the experiment and would involve minimal additional cost. It should be noted that the reliability estimates discussed in this paper do not include the effects of form-to-form variability (i.e., variability due to the use of different essay questions on different forms of the test) and only scoring reliability is considered.

2. Review of the Literature

Most of the published work on the use of essays in assessment deals with the problem in the context of the measurement of writing skills. Two excellent reviews are Coffman (1971) and Breland (1983). Both authors review the evidence on reader reliability and conclude that one of the major drawbacks to the general use of essays is the extraneous variability introduced by the necessity of involving multiple judges over extended periods of time. Such direct assessment is often very costly and sometimes yields results of rather low reliability. Bejar (1985) has also studied rater disagreements.

The use of statistically designed experiments to study the reliability of essay examinations has a long if somewhat sparse history. Finlayson (1951) and Vernon and Millican (1954) both carried out very large experiments intended to assess both reader reliability and the consistency of examinee performance across different topics. Stanley (1962) provides the details of the analysis of variance (ANOVA) for a very complex experiment and includes formulas for different reliability coefficients in terms of the mean squares from the ANOVA. In an earlier article, Stanley (1961) also suggests adjusting raw scores for systematic differences between raters to improve reliability. The use of variance components in this setting was developed extensively by Cronbach, et.al. (1972) in their work on generalizability theory. Another large scale experiment on the measurement of writing ability is described by Godshalk, et al. (1966). Since the studies mentioned above deal with essay questions that are rather different from each other and from those considered here, comparisons among the findings will not be made here.

Ebel (1951) was probably the first to consider the problem of estimating reliability from incomplete designs. More recently, Fleiss (1981) has described how balanced incomplete block designs may be employed to obtain unbiased estimates of differences in readers' standards. Paul (1981) suggests a Bayesian approach to the calibration of essay readers while de Gruijter (1984) proposes a formulation closely related to the Rasch model in item response theory (Lord, 1980). Another study of the reliability of essay grades and the patterns in reader reliability is reported by Blok (1985).

Apparently, a systematic study of the effects of different kinds of calibration on score reliability is yet to be carried out. Moreover, the ETS context is unique in that, operationally, adjustments based on a small sample of essays included in an experiment would have to be applied to a large set of essays. Questions of precision and cross-validation consequently arise in the assessment of the efficacy of any such procedure. These issues are considered in the analysis presented below.

3. Study Design

The design of the study was driven by two considerations: the structure of the grading process of the AP examinations and the nature of the information required. We begin with a description of the former.

The free-response portion of the AP examination in American History consists of two essay questions. The first, called the document based question (DBQ), must be answered by all candidates. The student is offered some eight pieces of information on a topic in the form of excerpts from historical documents, quotes, maps and the like. A question related to that topic is then posed. For the second essay question, the candidate must choose one question from among five options.

The grading is carried out over a six day period. All readers are trained for and begin grading the DBQ. After two days, readers are trained for one of the optional questions and begin grading that question. Toward the end of the fifth day or the beginning of the sixth day, the readers return to grading the DBQ.

The free-response portion of the AP examination in English Literature and Composition consists of three essay questions that all candidates must answer. Grading is carried out over a six day period and each reader is trained for and reads only one question throughout the grading period.

For both examinations, readers are assigned to "tables" consisting of six or seven readers. One of the readers is designated the "table leader" and helps to coordinate the readings and maintain standards.

It was not considered feasible or necessary to include all readers in this type of experimental study. Acccrdingly, for the American History exam (AH) two groups of readers were selected. One group of twenty-one readers (3 tables), including the table leaders, read the DBQ and optional question number 3. In the analysis these questions are denoted "D" and "X", respectively. A second group of twelve readers (2 tables), excluding the table leaders, read the DBQ and optional question number 6. In the analysis these questions are denoted "E" and "Y". Note that "D" and "E" refer to the same document-based question but "X" and "Y" refer to different optional questions. The grading of each question over a three day period was to be investigated.

For each question on the English Literature and Composition exam (ELC), twelve readers (2 tables) were selected to participate in the study. In the analysis, these questions are denoted "A", "B" and "C". The grading of each question over the central four day period was to be investigated. A graphical representation of this aspect of the design is presented in Figure 1.1.

The actual design employed was motivated by the information required to carry out the calibration. Estimates of the average difference between readers on each day as well as the average difference over the entire grading period were needed. In addition, the average difference between scores given on different days was considered essential.

One possible design would involve selecting a small sample of essays at random from the population of essays, with each essay read by each reader on each day. There are at least two drawbacks to this approach. The first is that the number of essays would have to be very small, say ten or so, in order to keep the experiment at a manageable size. So small a sample would raise doubts about the representativeness of the results. More important, the estimates of the between day differences would be confounded with carryover effects from repeated readings of the same essay.

An alternative to the complete factorial design described above is a balanced incomplete block (BIB) design, one in which unbiased estimates of reader effects can be obtained although each reader does not read all the essays. Unfortunately, the number of readings required of each reader to achieve this balance is prohibitively large. Consequently, a partially balanced incomplete block (PBIB) design was chosen (John, 1971). Such designs yield unbiased estimates of reader effects with fewer readings than required by a comparable BIB design at the cost of an increase in variance in the estimates of the difference between some pairs of readers. While a PBIB design employing a different set of essays on each day would permit the adjustment of scores for differences between readers, estimates of between day differences would be confounded with differences between sets of essays. Ideally, one would prefer selecting a single set of essays and constructing a different PBIB design for each day using these essays as blocks. Readers would read each essay only once during the entire period so that there would be no carry-over effect.

Fortunately, a particular class of PBIB designs, called semi-regular, group divisible designs (John, 1971) is perfectly suited to these requirements. To make matters more concrete, the design employed for essay E will be described in detail.

Figure 3.1 displays the allocation plan for the first day of the experiment. The twenty-seven essays (playing the role of "blocks") are represented by rows while the twelve readers (the "treatments") are represented by columns. In each row there are four "X's" denoting which readers were assigned to read that essay on day one. Similarly, in each column there are nine "X's" denoting the essays that were read by the reader on day one. Beyond these two obvious kinds of balance, there is a very delicate choice of reader-essay combinations so that each pair of readers either read three essays (" λ_2") in common or no essays ("λ_1") in common. For example, readers 1 and 4 read essays 1, 10 and 19 in common, while readers 2 and 5 read no essays in common. Data collected according to this partially balanced design can be used to obtain unbiased estimates of the systematic differences in readers' grading standards. This is quite remarkable: ordinarily in such an incomplete design estimates of readers' effects would be confounded with differences between essays since no two readers read the same set of essays.

Another view of the allocation is displayed in Figure 3.2a. Here the twelve readers are represented by rows and in each row the essays read by that reader on day one are listed in numerical order. (Note: In practice the essays were packaged and read in random order.) This version of the field plan makes it easy to see how the allocation for the next two days is organized. For day two, the essay lists remain fixed but the readers are shifted down four steps. Thus, reader 1 reads the essays read by reader 5 on the previous day, reader 2 those of reader 6, etc. Readers 9 through 12 are assigned the essays read previously by readers 1 through 4. The plan is displayed in Figure 3.2b. Similarly, for day three, the readers are shifted another four steps. Figure 3.2c contains the plan.

It is the special group divisible structure of the original plan that facilitates the generation of the plans for the last two days by simple shifts. Each day's plan is a PBIB permitting unbiased estimates of reader differences on that day. Over the three day period, each reader reads each essay exactly once. In fact, each day's plan is a one-third replicate of a complete factorial design. That is, if the data are pooled over days an ordinary complete factorial design (essays x readers) is obtained. This simplifies the estimation of certain variance components, as will be seen in Section 5. Equally important, unbiased estimates of systematic differences between days can be easily derived by computing the averages of scores given on each day. This follows because each reader reads nine essays on each day, each essay is read four times on each day and there is no overlap. An enumeration of the plans employed for each essay, with references is given in Appendix 1.

It was also decided to investigate time of day effects for the grading in ELC. Accordingly, the essays to be [illegible] each day were further divided into two groups of four with the additional constraint that these essays would be read an equal number of times in the morning and in the afternoon.

One feature of the overall design deserves mention. Substantial redundancy was incorporated at each level of the design so that even if part of the experiment failed for one reason or another, the remainder of the data would provide usable information. Thus, two different examinations and six essays in all were involved. Furthermore, for each essay, the data was collected over three or four days with each day's data providing unbiased estimates of relevant parameters.

4. Data Collection

Some ten days before the operational grading, three samples of examinees were obtained. The samples were all drawn randomly from the appropriate population of candidates and no books were discarded because of poor writing or the like. One sample consisted of twenty-seven candidates in American History who had written the DBQ and optional question 3. A second sample of twenty-seven candidates in American History who had written the DBQ and optional question 6 was also selected. The third sample consisted of thirty-two candidates in English Literature and Composition. Each candidate was assigned a unique code that was copied to the first page of each essay. The entire book was then photocopied to provide sufficient copies for all the readings. The separate essays were then bundled together in groups comprising a day's reading for each reader according to the plans similar to those given in Figure 3.2. A code for the day as well as the reader was also included on the copy. The sequence of the essays in each bundle was separately randomized to avoid confounding with order effects. Substantial effort was devoted to insuring that the plan was followed exactly and, in fact, only one error was made.

Tables of readers participating in the experiment were selected by the chief readers. Although the selection was not random, each table, by design, contained both new and experienced readers as well as both high school and college teachers. At the beginning of the grading of each examination, the readers selected to participate were addressed by both the chief reader and the author. They were informed of the nature of the experiment and that the grades they assigned would not affect the operational scores of the candidates. It was emphasized, however, that as much as possible they should carry out the grading in the experiment as if it were operational.

The bundled essays for a given day's reading were distributed to each reader by the author or his assistants. They were graded just after morning or afternoon coffee break and were interspersed among operationally read essays. At least one person from the experiment staff was available to answer questions or resolve problems. The graded essays were then retrieved, counted and aggregated for shipment to key entry. The data files produced by key entry, including all identification codes as well as scores, were carefully checked against the plans. The final analysis file of some three thousand records contains only three missing values. One value was missed because the wrong essay was inserted in a bundle, while the other two missing values resulted from readers inadvertently skipping a paper.

It should be noted that some readers displayed considerable resistance to the experiment, principally because they felt it distracted them from their true task. Others asserted that the experimental gradings would not be strictly comparable to the operational gradings because the former would be done more slowly and with perhaps greater care. That the experimental grading involved the use of photocopies rather than originals was also held to be a distorting factor. These issues are considered in the discussion in Section 7 of the general validity of the results.

5. Analysis of English Literature and Composition

5.1. Models and Estimates

The process of calibration depends on the precise estimation of any systematic differences between the levels of the factors in the experiment. For the first analysis we propose the following model: For each essay (A, B or C) let

y_{erdm} = grade assigned to examinee e by reader r on day d at time-of-day m,

where

$e = 1, \ldots, 32$

$r = 1, \ldots, 12$

$d = 1, \ldots, 4$

$m = 1, 2.$

(The grades y are on a scale of zero to nine.)

Then

$$y_{erdm} = k + u_e + v_r + w_d + x_m + \text{error}, \qquad (1)$$

where

k is the overall mean score for the essay

u_e is the deviation due to examinee e

v_r is the deviation due to reader r

w_d is the deviation due to day d

x_m is the deviation due to time-of-day m.

Note that the data available to estimate the model does not constitute a complete replicate of the four factor design, since each examinee-reader combination occurs on only one day/time-of-day combination rather than on eight days, as would be required for a complete replicate. Consequently, certain interactions can not be estimated. For the present, we exclude from the model even the estimable interactions since they prove to be rather small.

Missing values were imputed using standard techniques (Cochran and Cox, 1980) and the analysis was carried out with these imputed values, but no account was taken of the imputation process. The effect on the results was negligible since there were only three missing values overall.

Table 5.1 presents estimates of k,$\{w_d\}$, $\{x_m\}$ and $\{v_r\}$ for essays A, B and C. For A and B, the estimated day effects $\{\hat{w}_d\}$ are quite small absolutely as well as relative to the estimated standard errors. This means that the average score assigned was quite stable over the four days. For C, however, there was a sizeable shift in average score from day two to day three. The increase of 0.82 points (0.82 = .2786 + .5235) was followed by an increase of 0.15 points (0.15 = .4349 - .2786) on day four. We have not been able to find a good explanation for this phenomenon. In all three essays, the time-of-day deviations are quite small.

The estimated reader effect represents the difference between the average grade assigned by the reader over the course of the experiment and the average grade assigned by all readers over the experiment. They display considerable variability with extreme differences ranging from about 1.3 points in A to nearly 1.8 points in C. The distribution of the deviations is presented in Figure 5.1 which contains stem-and-leaf displays of the data. While most readers' deviations cluster about zero (corresponding to little bias), fully one-third (13/36) have average deviations that are 0.5 points or more away from zero. This suggests that adjusting scores for differences between readers should improve the reliability.

Before turning to the estimation of variance components and reliabilities, we want to take advantage of our PBIB design by estimating reader effects separately for each day. This allows us to study how differences between readers may have changed from day to day and corresponds to the so-called reader by day interaction. Table 5.2 shows the estimated reader effects for each of the four days for essays A and C. The readers are listed in descending order of their estimated deviations from Table 5.1. The entries are the differences between the estimated reader grade level based on the readings of a given day and the average grade on that day. If the reader by day interaction is small then we would expect to see similar estimated effects across the four days for each reader. In fact, some readers exhibit considerable variability. For example, Reader 2 on Essay A has the largest positive deviation on day one, but nearly the smallest deviation on day three. Of course, some of the variability is the result of sampling fluctuations since each reader grades only eight papers each day. Nonetheless, these findings suggest that we should explore the possibility of calibrating readers separately each day rather than once overall.

Table 5.3 contains the ordinary "fixed-effects" analysis of variance (ANOVA) for the data from the three essays. For this analysis, two interactions have been added to model (1). As expected, the mean square for days is large only for C, while the reader mean square is sizeable for A, B and C. There is a hint that the reader by day interaction is significant.

5.2. Estimating Reliabilities from Variance Components

To compute reliabilities, however, we need to estimate components of variance. One difficulty is that the theory of variance component estimation is not well developed for incomplete factorials. However, in the present instance there is sufficient symmetry to facilitate the derivation of estimates without undue effort.

We consider the model:

$$y_{erd} = k + u_e + v_r + w_d + \text{error} \qquad (2)$$

where the meaning of the terms is the same as in (1). We have eliminated the time-of-day term since it was uniformly small over the three essays. Interactions have been absorbed into the error term. We let S_u^2, S_v^2 and S_w^2 and t^2 denote the variances of the $\{u_e\}$, $\{v_r\}$, $\{w_d\}$ and error, respectively. Estimates of the variance components are presented in Table 5.4. (See Appendix 2 for derivations.) The variance component for days is not negligible only for essay C. Note that for all three essays $\hat{S}_v^2$, the estimated variance component for readers is about twenty percent of $\hat{t}^2$, the error variance component.

The estimated variance components can be used to compute estimated reading reliabilities. The single reading reliability is estimated by:

$$\hat{r}_1 = \frac{\hat{S}_u^2}{\hat{S}_u^2 + \hat{S}_v^2 + \hat{S}_w^2 + \hat{t}^2}, \qquad (3)$$

where r_1 represents the correlation between pairs of readings of a set of essays; It is assumed that for each essay the readings are carried out by different readers on different days. An idea of the reading reliability of the calibrated scores can be obtained by setting $\hat{S}_v^2$ and $\hat{S}_w^2$ to zero, corresponding to perfect calibration:

$$\hat{r}_{C1} = \frac{\hat{S}_u^2}{\hat{S}_u^2 + \hat{t}^2} \quad . \qquad (4)$$

Thus, $\hat{r}_{C1}$ represents an upper bound on the reliability of a single reading, after calibration. The improvement in reliability actually obtained by calibration will be less than that indicated by the difference $\hat{r}_{C1} - \hat{r}_1$ which is too optimistic since the adjustments are estimated from the same essays on which the calibration will be carried out.

In practice, the adjustments to most essays would be made on the basis of an experiment involving a small number of essays. The estimated reader effects derived from this experiment will in general not be the same as those that would have been obtained had the estimation been based on the entire universe of essays. Thus, a more realistic estimate of the improvement in reliability resulting from this type of calibration can be computed by replacing $\hat{S}_v^2$ and $\hat{S}_w^2$ in (3) not by zero as in (4), but by quantities that reflect the variance remaining in reader averages and day averages after the adjustment. Estimates of these variances, denoted by $\tilde{S}_v^2$ and $\tilde{S}_w^2$, are provided by the SAS program VARCOMP. Consequently, we can calculate something akin to a cross-validated version of $\hat{r}_{C1}$:

$$\hat{r}_{XC1} = \frac{\hat{S}_u^2}{\hat{S}_u^2 + \tilde{S}_v^2 + \tilde{S}_w^2 + \hat{t}^2} \quad . \qquad (5)$$

Note that there is no need to introduce a covariance term because the estimates of the reader and day effects are orthogonal. The estimated reliabilities $\hat{r}_1$, $\hat{r}_{C1}$, $\hat{r}_{XC1}$ are contained in Table 5.5.

The raw (single-reading) reliability of essay C is considerably lower than that of the other two essays, but for all three essays the cross-validated reliability of the calibrated scores is substantially higher than the raw reliability. Thus, calibration seems to result in a large improvement in reliability, even though for essays A and B the day factor has no detrimental effect on reliability.

5.3. Estimating Reliabilities through Sampling

To supplement this analysis, a small sampling experiment was carried out. Three 32 x 12 matrices (representing examinees by readers) were constructed for each essay. The first matrix contained the raw scores, the second contained the calibrated scores and the third contained scores calibrated in a manner to simulate the proposed operational setting of the procedure.

Specifically, this latter calibration was executed by successively eliminating one examinee, estimating reader and day effects from the remaining thirty-one examinees and using those estimates to calibrate the scores assigned to the eliminated examinee. Thus, all the (adjusted) scores for each examinee in the third matrix are calibrated without using the data for that examinee.

To estimate the reliability, two elements are chosen at random from each row of the matrix, corresponding to two readings of each essay. The actual randomization procedure involves choosing two days at random without replacement from among the four days and one reading at random from among the three readings carried out on each day. The same elements are selected from all three matrices in each trial for a particular essay, so that comparisons among the different reliabilities are not confounded with differences between random selections. (A different randomization is chosen for each essay, however, so that a proper estimate of the reliability of the total essay score may be obtained.) The correlation between the thirty-two pairs of scores taken from each matrix estimates the corresponding single-reading reliability and constitutes a single trial of the sampling experiment. Fifty trials were performed and the results averaged. They are presented in Table 5.6 and may be compared to those in Table 5.5.

The agreement is quite close, certainly within the sampling error of the experiment. However, the improvement of the cross-validated calibrated scores over the raw scores is not quite as large in the sampling experiment as it is in the components of variance analysis. We will therefore carry out the calculations below using the results of the sampling experiment in order to be conservative concerning the potential benefits of the procedure.

One useful way to assess the effect of calibration is to compare the gain in reliability with the gain obtained with a full double-reading. The latter can be estimated from the raw single-reading reliability by use of the Spearman-Brown formula (Gulliksen, 1950):

$$r_2 = 2r_1 (1+r_1)^{-1} \quad .$$

The comparison can be made conveniently in terms of the gain ratio,

$$g = \frac{\hat{r}_{XC1} - \hat{r}_1}{\hat{r}_2 - \hat{r}_1} \times 100\% \quad .$$

Table 5.6 also displays the quantities $\hat{r}_2$ and g. The gain ratios are substantial, particularly in view of the fact that the calibration experiment involves a five to seven percent increase in the reading load over the single-reading, while double-reading obviously requires a one hundred percent increase.

A sampling experiment was also carried out to estimate the reliability of scores calibrated by making different adjustments for readers on different days; i.e. taking full advantage of the

PBIB design employed on each day. These results are presented in the last line of Table 5.6 labelled $\hat{r}_{C*1}$. Note that for essays B and C the results actually are poorer than those for $\hat{r}_{C*1}$. Presumably, the sampling variability in estimates of reader effects based on only eight readings in a PBIB design overwhelms the between day variability within readers. A cross-validated version of $\hat{r}_{C*1}$ was not run.

6. Analysis of American History

For this analysis we employ the model:

$$y_{erd} = k + u_e + v_r + w_d + (uw)_{ed} + (vw)_{rd} + \text{error} \qquad (6)$$

where $\{u_e\}$, $\{v_r\}$ and $\{w_d\}$ are defined as before and $\{(uw)\}_{ed}$ represent the deviations due to examinee-by-day interactions while $\{(vw)_{rd}\}$ represent the deviations due to reader-by-day interactions. Again, the observed data do not represent a complete replicate of the three factor design so that certain interactions are not estimable. (Grades are on a scale of zero to fifteen.)

Table 6.1 presents estimates of the main effects in (6) for essays D, E, X and Y. (Recall that D and E refer to the same question - the DBQ.) The estimated day effects for the DBQ are substantial, as they are for X. Those for essay Y are somewhat smaller. Estimated reader effects are comparatively larger here than in English Literature and Composition. Figure 6.1 displays stem-and-leafs for the four batches and should be compared with Figure 5.1. From one-third to one-half of the readers have deviations at least 0.5 points away from zero.

Since the PBIB design was employed for this essay as well, it is possible to estimate unbiasedly reader effects on each of the three days. To illustrate, Table 6.2 contains these estimates for essays E and Y while Figure 6.2 graphs the results for Essay E. Essays D and X display even more variability across days, suggesting that a calibration employing within-day reader effects may be more efficacious here.

The "fixed-effects" ANOVA for the four essays is presented in Table 6.3, which should be compared with Table 5.3. All the main effects are strongly significant here, even though the mean square for error tends to be two to three times larger than in the English examinations. It is especially interesting that for the DBQ the day-by-reader interactions are significant as well. Note that the results for D and E closely resemble each other as they should since they are replicates of the same experiment.

Using the methods and notation of Section 5, equation (2) and following, we obtain estimates of the variance components S_u^2, S_v^2, S_w^2, S_t^2. These are contained in Table 6.4. The estimated variance component for readers, $\hat{S}_v^2$, is less that 20 percent of $\hat{t}^2$, the estimated variance component for error. The estimated variance components for days, $\hat{S}_w^2$, are small but not negligible, except perhaps for essay Y.

The corresponding estimated reliabilities are found in Table 6.5. The calculations parallel those described in Section 5. Interestingly, the reliability of the DBQ appears to be substantially below that of the essays A, B and C in the English examinations while the reliabilities of the optional questions, X and Y, equal or exceed those of A, B and C. As before, a small sampling experiment of fifty runs was carried out to supplement the results obtained from the components of variance analysis. The results are displayed in Table 6.6. The estimated gain ratios are respectable but somewhat smaller than those for the English essays. The small proportional improvement for D+X, however, is somewhat disturbing. While it may be due in part to sampling fluctuations, it may also be due to elevated levels in some interactions.

Accordingly, a second set of calibrations was carried out. For this set, scores were adjusted using the estimated reader effects obtained from the PBIB designs. Thus, a different set of estimated reader effects was used for each day. The results are displayed on the last line in Table 6.6 and are denoted by $\hat{r}_{C*i}$. The gains are substantial for the DBQ, but a cross-validated version was not run because of cost considerations.

7. Comparisons

The analyses described in the previous sections indicate that experiments embedded in operational grading can provide useful information. How useful the information will prove in an operational setting depends on the extent to which the experiment truly represents a microcosm of the operational setup. This question of validity is particularly important in view of the remarks made by many of the readers that they treated the essays graded experimentally somewhat differently than they normally would.

One possibility is to compare the raw reliabilities obtained here with those obtained in earlier reader reliability studies. For example, Modu and Bleistein (1982) estimated the single-reading reliabilities of the three English Literature and Composition essays to be .47, .54 and .49. These figures are similar to ours. Bleistein, et al. (1980) estimated the single-reading reliability of the DBQ in the American History examination to be .51. This is substantially larger than our estimates. For the five optional questions, the reliabilities varied from .47 to .67 and our own estimates agree reasonably well. We conclude that, except for the DBQ, our estimated reliabilities are in line with previous findings.

Another approach to the question of validity is to compare our readers' characteristics as revealed in the experiment with the various summary statistics available from the operational readings. First we compare the mean scores assigned by the readers to the essays in the experiment with the mean scores they assigned essays operationally (Table 7.1). Differences, of course, could occur because the essays graded experimentally were not representative of those graded operationally.

For essays A and B in English Literature and Composition, the operational means are higher than experimental means, while for essay C they are equal. For essays D, E, X and Y, the experimental means are considerably higher than the operational means. No simple explanation is apparent, although sampling fluctuations are unlikely to provide a complete explanation. For this characteristic, our experimental results do not accord well with practice.

In calibration, however, it is not the overall level that matters, but the differences between readers. Our estimated reader effects can be compared with the differences between mean scores assigned by readers operationally. The latter differences, however, could be confounded with differences in the overall quality of the papers read by the various readers and the proportions of papers read on each day. Nonetheless, the comparisons are instructive and can be most vividly presented by plots of experimental reader effects against operational reader effects. These are presented in Figures 7.1a through 7.1g, together with the correlations. We note that with the exception of essay E, the correlations range between 0.71 and 0.90 with most around 0.80. The low value of 0.48 is due to a single outlier grading E. The variances of the estimated effects obtained through the two methods are comparable for all of the essays as well. These data tend to support the claim that the experiment produced valid representations of operational characteristics of readers.

The third and final comparison is based on calibrating the essays in the experiment with estimated reader effects derived from the operational readings and obtaining, through sampling experiments, estimates of the resulting reliabilities. Actually, if we were assured that each reader was assigned a random sample from the universe of essays for scoring, the differences between average grades assigned by the readers would be unbiased estimates of the true differences between readers. These estimated differences would be the basis of what might be termed an operational calibration, as opposed to the experimental calibration we have discussed so far. In practice, the essay booklets go through several stages of haphazard shuffling before final allocation so that readers receive what may be termed a quasi-random sample of essays. One purpose of carrying out a designed experiment along the lines described in this paper is to provide a baseline for judging the utility of an operational calibration.

For purposes of this comparison, however, we reversed the direction of the argument and reason that if the experiment is valid, this operational calibration should work reasonably well and, at the very least, not degrade the reliability. One problem is that we cannot operationally calibrate the readers for each day separately and cannot obtain estimates of overall trends from day to day because the relevant information was not preserved.

Accordingly, we employ estimated day effects obtained from the experiment and overall estimated reader effects obtained operationally. The results are presented in Table 7.2 and are rather interesting. The operational calibration appears to work about as well as the cross-validated experimental calibration.

These findings support the contention that the experimental outcomes are indicative of differences observed operationally. More important, it appears that the quasi-random assignment of examinees to readers works well enough and the samples are large enough that calibration of readers through adjustments based on operational readings should work well in practice. We can expect that operational calibration of readers using different adjustments for each day could yield even better results since the number of papers read on each day is quite large so that sampling fluctuations would be quite small. This is of great interest since these calibration constants can be obtained at minimal additional cost. This approach must be tested by carrying out traditional reliability studies for several examinations and investigating the improvements in reliability produced by operational calibration.

8. Conclusions

Our analyses in the previous sections suggest that the experiments carried out do provide useful information about essay scoring and that this information can be used for calibration or for other studies. (Some of these other studies will be described in a companion report.) In particular, the components of variance analyses enable us to estimate an upper bound to the reliability that can be achieved through calibration. Such calculations are not possible with data from an operational setting. However, we have also learned that calibration based directly on operational scoring is a strong possibility and may prove to be the most cost-effective way of improving reliability. This approach needs to be tested thoroughly in the forthcoming year. In particular, the logistics of obtaining the calibration constants in a timely fashion must be worked out.

What sorts of improvement in reliability can we expect? If our present results are any indication, then calibration can bring an improvement in reliability for the AP English examination equal to that produced by double reading one of the three essays. For example, in the 1982 reader-reliability study of AP English, the reliability of the total essay score was .69. Rereading of only one of the three essays increased the reliability of the total essay score to .71, .72 or .73. These values are comparable to the gains presented in Table 5.6.

While such gains are substantial, they are not overwhelming. This is testimony that the procedures employed by Advanced Placement to maintain standards throughout the grading process are reasonably effective. Certainly systematic differences do occur, but most of the variability that reduces reliability appears to be idiosyncratic and is not susceptible to calibration in this setting. Although our present results need to be replicated, it appears that some radical new approach will be required to produce new substantial improvements in reliability.

In fact, analysis of the estimated components of variance for the various essays suggests that if operationally readers receive approximately random samples of essays from the candidate population, then operational calibration should do a better job than experimental calibration in reducing the variability among readers. The latter approach, however, already reduces the between-reader variance to a negligible proportion of the error variance. Hence, further improvements can only result in very small increases in reliability. This scenario accords well with our empirical findings.

The preceding discussion has focused entirely on the import of this study for the AP program. What can other programs learn from it? It may very well be that AP presents a "worst-case" for this approach in the sense that AP readers are already quite well calibrated. In other programs, where training and standards are perhaps not as demanding, systematic differences may constitute a larger component of the total error and, consequently, improvement in reliability effected by calibration may be quite large. Certainly the results of the present study should encourage other programs to investigate both experimental and operational calibrations.

Appendix 1

Appropriate PBIB designs for the various experiments were chosen from a collection of designs enumerated in Bose, et al. (1954) They employ the following notation:

b = # Llocks

k = # units/blocks

v = # treatments

r = # replicates/treatment

In the present setting, the essays play the role of blocks and readers the role of treatments. Then k is the number of times the essay is read on a given day while r is the number of essays scored by a reader on a given day.

The designs employed are classified as semi-regular group divisible designs with two associate classes. The symbol λ_i is used to denote the number of times two treatments (readers) that are ith associates, occur together in the same block (read the same essay). Obviously, for these designs v/k is the number of days that would be required to obtain all essay-reader combinations.

Below we present a summary of the characteristics of the designs.

Essays	Design	b	k	v	r	λ_1	λ_2
A,B,C	SR27	32	3	12	8	0	2
D,X	SR60	27	7	21	9	0	3
E,Y	SR29	27	4	12	9	0	3

Appendix 2

Since the experimental design employed is an incomplete three factor design, estimation of the variance components is not straightforward even if all interactions are assumed negligible. The approach taken here is somewhat ad hoc, but is based on the symmetry inherent in the design.

The objective is to estimate the components of variance corresponding to equation (2). First, by pooling over days, a complete two factor design (essays and readers) is obtained. The SAS program VARCOMP can be employed directly to obtain estimates of the variance components for essays, readers and errors. It remains only to estimate the variance component for days.

The second step is to fill in the missing values in the original three factor design in accordance with the assumed model. For example, suppose the combination of essay e' and r' occurs on day d' but not on day d". Then we set

$$\hat{y}_{e'r'd''} = \hat{k} + \hat{u}_{e'} + \hat{v}_{r'} + (\hat{w}_{d''} - \hat{w}_{d'}) ,$$

where the estimated quantities on the right hand side are all derived from the standard fixed effects analysis of the original design. Before the variance component for days can be computed, however, an appropriate adjustment to the mean squares must be made owing to the fact that some of the data has been manufactured.

A reasonable adjustment seems to be the following: Let $(MS)*_w$ denote the mean square for days in the completed three factor design and $\hat{t}^2$ the estimated variance component for error from the (pooled) two factor design. Further, let E be the number of examinees, R the number of readers and D the number of days. Than the estimated variance component for days is

$$\hat{S}_w{}^2 = \left[(MS)*_w - D\hat{t}^2\right] / ER.$$

(Note: If the full design had actually been carried out, the above formula with D=1 would be correct.)

As a check, the above method was employed to estimate the variance components for examinees and readers $S_u{}^2$ and $S_v{}^2$. For example, the analagous estimate of $S_v{}^2$, would be

$$\hat{S}_v{}^2 = \left[(MS)*_v - D\hat{t}^2\right] / ED,$$

where $(MS)*_v$ is the mean square for readers in the completed three factor design. This agrees exactly with the result obtained from VARCOMP. Of course, this agreement does not constitute a mathematical proof that the estimates obtained are correct.

One such derivation would involve using the EM algorithm (Dempster, et al., 1977) to obtain "estimates" of the sufficient statistics required for the estimation of the variance components. The maximum likelihood estimates of the variance components could be calculated by substituting these values for the sufficient statistics into the appropriate formulas. This approach is currently being explored and will be reported on elsewhere.

As a further check, Paul Rosenbaum has suggested eliminating reader effects and pooling the data over readers. For the American History examination, this results in a two factor design (examinees x days) with four observations in each cell. Analysis by VARCOMP leads to variance component estimates which are essentially identical to those derived from the first method of pooling. The same holds true when this procedure is applied to the English Literature and Composition examination.

References

Bejar, I. (1985). A preliminary study of raters for the test of spoken English. Research Report 85-5. Princeton, NJ: Educational Testing Service.

Bose, R. C., Clatworthy, W. H., Shrikhande, S. S. (1954). Tables of partially balanced designs with two associate classes. North Carolina Agricultural Experimental Station. Technical Bulletin No. 107.

Bleistein, C. A., Levy, J., and Modu C. C. (1981). Internal Memorandum, Princeton, N.J.: Educational Testing Service.

Blok, H. (1985). Estimating the reliability, validity and invalidity of essay ratings. Journal of Educational Measurement, 22, 41-52.

Breland, H. M. (1983). The direct assessment of writing skill: A measurement review. College Board Report No. 83-6, New York: College Entrance Examination Board.

Coffman, W. E. (1971). Essay examinations.(in) Educational measurement. Second edition. R.L. Thorndike, editor. Washington, D.C.: American Council on Education.

Cronbach, L. J., Gleser, G. C., Nanda, H. and Rajaratnam, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles, New York: John Wiley.

de Gruijter, D. N. M. (1984). Two simple models for rater effects. Applied Psychological Measurement, 8, 213-218.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum Likelihood from incomplete data via the EM algorithm. The Journal of the Royal Statistical Society B, 39 (1), 1-38.

Ebel, R. L. (1951) Estimation of the Reliability of Ratings. Psychometrika, 16, 407-424.

Finlayson, D. S. (1951). The reliability of the marking of essays. British Journal of Educational Psychology, 21, 126-134.

Fleiss, J. L. (1981). Balanced incomplete block designs for inter-rater reliability studies. Applied Psychological Measurement, 5, 105-112.

Godshalk, F. I., Swineford, F., and Coffman, W. E. (1966). The measurement of writing ability. New York: College Entrance Examination Board.

John, P. W. M. (1971). Statistical design and analysis of experiments. New York: The Macmillan Co.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillside, NJ: Lawrence Erlbaum Associates.

Modu, C. C. and Bleistein, C. A. (1982). College Board 1982 Advanced Placement examination [in] English Literature and Composition reader reliability study. Report No. 82-85. Princeton, N.J.: Educational Testing Service.

Paul, S. R. (1981). Bayesian methods for calibration of examiners. British Journal of Mathematical and Statistical Psychology, 34, 213-223.

Snedecor, G. W. and Cochran, W. G. (1980). Statistical methods. (seventh edition) Ames, Iowa: The Iowa State University Press.

Stanley, J. C. (1961). Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. Psychometrika, 16, 205-219.

Stanley, J. C. (1962). Analysis-of-variance principles applied to the grading of essay tests. Journal of Experimental Education, 30, 279-283.

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

Table 5.1

Estimates of Main Effects in Model 1 for Essays A, B and C

Essay	A	B	C
Grand Mean	4.828	4.589	4.982
Day Deviations			
1	.0678	-.1302	-.1901
2	-.0366	.0155	-.5235
3	-.0470	.0990	.2786
4	.0158	.0157	.4349
Time of Day Deviations			
AM	-.060	-.079	-.010
PM	+.060	+.079	+.010
Reader Deviations			
1	-.515	-.714	-.919
2	.828	-.026	-.419
3	.235	-.683	-.419
4	.360	.006	.393
5	.016	-.370	-.169
6	-.515	-.276	-.044
7	-.109	.756	-.075
8	-.078	.693	.862
9	.610	.256	-.138
10	-.578	-.245	.831
11	.047	.068	.050
12	-.297	.537	.049

Table 5.2a

Reader Deviations by Day from Day Mean: Essay A

Reader No.	Reader Average Deviation	Day	1	2	3	4
8	.86		.16	1.74	.50	- .61
10	.83		-.67	1.46	.63	.96
4	.39		.27	.34	.50	.09
11	.05		.29	-.45	.63	.39
12	.04		-.50	-.15	.11	- .27
6	-.04		-.50	.10	-1.20	.54
7	-.07		.15	-.54	-.31	- .41
9	-.13		.26	-.09	.61	-1.08
5	-.16		.35	-.32	.32	.14
2	-.41		1.16	-.30	-1.00	.07
3	-.41		.94	-.71	.05	.04
1	-.91		-.67	-.98	- .87	.21

Table 5.2b

Reader Deviations by Day from Day Mean: Essay C

Reader No	Reader Average Deviation	Day	1	2	3	4
2	.82		1.06	.79	.63	.62
9	.61		.15	.68	.75	.96
4	.36		[illegible]	.54	.13	.46
3	.23		.84	-.45	.25	-.22
11	.04		.18	.48	.00	.18
5	.01		.24	-1.02	.00	.56
8	-.07		.05	-.02	-.12	-.38
7	-.10		.04	-.59	-.68	.40
12	-.29		-.60	.30	-.25	-.79
6	-.51		-.60	.05	-.25	-.72
1	-.51		-.77	-.52	-.19	-.97
10	-.58		-.77	-.21	-.25	-.16

Table 5.3

Analysis of Variance for Essays A, B and C

	Essay	A		B		C	
Source	df	SS	MS	SS	MS	SS	MS
Days	3	.80	.27	2.61	.87	55.38	18.46
Time of Day	1	.04	.04	.26	.26	.21	.21
Readers	11	70.90	6.45	84.99	7.73	91.03	8.28
Examinees	31	609.37	19.66	591.48	19.08	442.78	14.28
Day X Readers	33	48.08	1.46	72.57	2.20	95.54	2.90
Day X Examinee	93	70.63	.76	85.48	.92	102.55	1.10
Error	211	168.83	.80	195.60	.93	259.36	1.23
Total	383						

Table 5.4

Estimated Components of Variance for Essays A, B and C

Essay	A	B	C
Source			
Readers	.175	.209	.217
Examinees	1.570	1.506	1.080
Days	.000	.000	.178
Error	.843	1.037	1.342

Table 5.5

Estimated Reliabilities from Components of Variance Analysis for Essays A, B and C

Essay	A	B	C
$\hat{r}_1$ (raw)	.607	.547	.383
$\hat{r}_{C1}$ (calibrated)	.651	.592	.446
$\hat{r}_{XC1}$ (calibrated and cross-validated)	.642	.583	.436

Table 5.6

Estimated Reliabilities from Sampling Experiment for Essays A, B and C and Total Essay Score

Essay	A	B	C	Total
$\hat{r}_1$ (raw)	.57	.56	.41	.68
$\hat{r}_{C1}$ (calibrated)	.61	.62	.48	.74
$\hat{r}_{XC1}$ (calibrated and cross-validated)	.61	.59	.46	.72
$\hat{r}_2$ (double reading)	.726	.718	.581	.810
g (gain ratio)	26%	19%	29%	31%
$\hat{r}_{C*1}$ (calibrated by PBIB)	.67	.61	.47	

Table 6.1

Estimates of Main Effects in Model 6 for Essays D, E, X and Y

Essay	D	E	X	Y
Grand Mean	7.365	7.361	7.201	7.767
Day Deviations				
1	.169	.389	.254	-.147
2	.180	-.148	.000	-.036
3	-.349	-.241	-.254	.187
Reader Deviations				
1	.11	1.83	.17	1.25
2	.08	-1.02	.39	-.31
3	-.42	.24	-.24	-.50
4	-1.28	.77	-.42	.86
5	-.28	-.40	.73	-.37
6	-.96	-1.03	-.94	.03
7	-.17	-.70	.28	.03
8	.03	-.59	.17	-1.12
9	.52	-.03	-.13	.29
10	-.50	.56	-1.13	.03
11	-.08	-.08	-.16	-.29
12	-.11	.44	-.68	.14
13	.95		-.72	
14	1.565		.87	
15	-.16		.87	
16	1.28		1.84	
17	-.37		-.05	
18	1.30		1.16	
19	-1.31		-.27	
20	-.50		-1.01	
21	.29		-.81	

Table 6.2a

Reader Deviations by Day from Day Mean: Essay E

Reader No.	Reader Average Deviation	Day 1	2	3
1	1.83	1.35	1.69	2.46
4	.77	1.20	.81	.30
10	.56	2.18	-.42	-.07
12	.44	1.24	.74	-.66
3	.24	-.76	.44	1.05
9	-.03	.09	-.64	.46
11	-.08	-.10	.18	-.32
5	-.40	-.91	-.31	-.49
8	-.59	-.24	-.85	-.66
7	-.70	-.95	-1.00	-.54
2	-1.02	-1.35	-.20	-1.52
6	-1.03	-1.75	-1.24	-.11

Table 6.2b

Reader Deviations by Day from Day Mean: Essay Y

Reader No.	Reader Average Deviation	Day 1	2	3
1	1.25	.82	1.52	1.41
4	.86	1.11	.97	.51
9	.29	.34	.59	-.07
12	.14	.70	-.54	.25
10	.03	.43	-.54	.21
7	.03	.05	.36	-.31
6	-.03	.13	-.06	-.16
11	-.29	-.58	-.57	.29
2	-.31	-.76	-.14	-.05
5	-.37	-.58	-.41	-.11
3	-.50	-.55	-.31	-.64
8	-1.12	-1.11	-.84	-1.42

Table 6.3

Analysis of Variance for Essays D, X, E and Y

	Essay	D		X	
Source	df	SS	MS	SS	MS
Day	2	34.58	17.29	24.38	12.19
Reader	20	335.28	16.76	316.71	15.84
Examinee	26	940.19	36.16	1893.84	72.84
Day X Reader	40	328.72	8.22	150.83	3.77
Day X Examinee	52	132.44	2.55	163.67	3.15
Error	426	1066.22	2.50	951.65	2.23
Total	566				

	Essay	E		Y	
Source	df	SS	MS	SS	MS
Day	2	24.96	12.48	6.22	3.11
Reader	11	192.38	17.49	132.38	12.03
Examinee	26	934.17	35.93	1620.89	62.34
Day X Reader	22	113.35	5.15	26.25	1.19
Day X Examinee	52	164.49	3.16	93.81	1.80
Error	210	525.40	2.50	380.09	1.81
Total	323				

Table 6.4

Estimated Components of Variance for Essays D, E, X and Y

Essay	D	E	X	Y
Source				
Readers	.512	.543	.496	.381
Examiners	1.582	2.760	3.353	5.049
Days	.076	.090	.052	.013
Error	2.937	2.808	2.435	1.749

Table 6.5

Estimated Reliabilities from Components of Variance Analysis for Essays D, E, X and Y

Essay	D	E	X	Y
$\hat{r}_1$ (raw)	.310	.445	.529	.702
$\hat{r}_{C1}$ (calibrated)	.350	.496	.579	.743
$\hat{r}_{XC1}$ (calibrated and cross-validated)	.341	.484	.569	.732

Table 6.6

Estimated Reliabilities from Sampling Experiment for Essays D, E, X, Y, D+X, E+Y

Essay	D	E	X	Y	D+X	E+Y
$\hat{r}_1$ (raw)	.26	.42	.53	.71	.56	.70
$\hat{r}_{C1}$ (calibrated)	.31	.48	.58	.75	.59	.75
$\hat{r}_{XC1}$ (calibrated and cross-validated)	.29	.45	.56	.74	.57	.73
$\hat{r}_2$ (double reading)	.41	.59	.69	.83	.72	.82
g (gain ratio)	20%	18%	19%	25%	6%	25%
$\hat{r}_{C*1}$ (calibrated by PBIB)	.42	.51	.59	.74		

Table 7.1

Mean Scores Assigned to Essays Experimentally and Operationally

	A	B	C	D	E	X	Y
Experimental Mean	4.83	4.59	4.98	7.36	7.36	7.20	7.77
Operational Mean	5.22	4.88	4.97	6.77	6.98	6.85	7.05

Table 7.2

Essay Reliabilities: Comparison of Operational and Experimental Calibrations

	A	B	C	A+B +C	D	X	D+ X	E	Y	E+ Y
Operational Calibration	.64	.60	.47	.72	.28	.57	.58	.43	.73	.71
Experimental Cross-validated Calibration	.61	.59	.46	.72	.29	.56	.57	.45	.74	.73

Figure 1.1

Timing of Experimental Gradings Within Operational Framework

American History		D E	D E			X Y		X Y	X Y	E	D			
English Literature & Composition							A B C	A B C	A B C	A B C	A B C	A B C	A B C	A B C
Date	am	pm	am	pm	am	pm	am	pm	am	pm	am	pm	am	pm
	6		7		8		9		10		11		12	

Figure 3.1

Allocation Plan For Partially Balanced Incomplete Block Design for Essay E

	Readers											
	1	2	3	4	5	6	7	8	9	10	11	12
Essay 1	X	X	X	X								
2					X	X	X	X				
3									X	X	X	X
4	X	X	X					X				
5					X	X	X					X
6				X					X	X	X	
7	X	X	X									X
8				X	X	X	X					
9								X	X	X	X	
10	X			X		X					X	
11			X		X			X		X		
12		X					X		X			X
13	X					X		X			X	
14			X		X					X		X
15		X		X			X		X			
16	X					X					X	X
17			X	X	X					X		
18		X					X	X	X			
19	X			X			X			X		
20		X			X			X			X	
21			X			X			X			X
22	X						X	X		X		
23		X			X						X	X
24			X	X		X			X			
25	X						X			X		X
26		X		X	X						X	
27			X			X		X	X			

Figure 3.2a

Essay E. Field Plan for Day 1

Essays Read

Reader	1	1	4	7	10	13	16	19	22	25
	2	1	4	7	12	15	18	20	23	26
	3	1	4	7	11	14	17	21	24	27
	4	1	6	8	10	15	17	19	24	26
	5	2	5	8	11	14	17	20	23	26
	6	2	5	8	10	13	16	21	24	27
	7	2	5	8	12	15	18	19	22	25
	8	2	4	9	11	13	18	20	22	27
	9	3	6	9	12	15	18	21	24	27
	10	3	6	9	11	14	17	19	22	25
	11	3	6	9	10	13	16	20	23	26
	12	3	5	7	12	14	16	21	23	25

Figure 3.2b

Essay E. Field Plan for Day 2

Essays Read

Reader	1	2	5	8	11	14	17	20	23	26
	2	2	5	8	10	13	16	21	24	27
	3	2	5	8	12	15	18	19	22	25
	4	2	4	9	11	13	18	20	22	27
	5	3	6	9	12	15	18	21	24	27
	6	3	6	9	11	14	17	19	22	25
	7	3	6	9	10	13	16	20	23	26
	8	3	5	7	12	14	16	21	23	25
	9	1	4	7	10	13	16	19	22	25
	10	1	4	7	12	15	18	20	23	26
	11	1	4	7	11	14	17	21	24	27
	12	1	6	8	10	15	17	19	24	26

Figure 3.2c

Essay E. Field Plan for Day 3

Essays Read

Reader	1	3	6	9	12	15	18	21	24	27
	2	3	6	9	11	14	17	19	22	25
	3	3	6	9	10	13	16	20	23	26
	4	3	5	7	12	14	16	21	23	25
	5	1	4	7	10	13	16	19	22	25
	6	1	4	7	12	15	18	20	23	26
	7	1	4	7	11	14	17	21	24	27
	8	1	6	8	10	15	17	19	24	26
	9	2	5	8	11	14	17	20	23	26
	10	2	5	8	10	13	16	21	24	27
	11	2	5	8	12	15	18	19	22	25
	12	2	4	9	11	13	18	20	22	27

Figure 5.1

Stem-and-Leaf Display of Estimated Reader Effects

Essay

	A	B	C
+10*			
9			
8	2		6 3
7		5	
6	1	9	
5*	3		3
4			
3	6		9
2	3	5	
1			
+0	1 4	0 6	5 4
-0	7	2	4 7
-1	0		6 3
-2	9	7 4	
-3	7		7
-4		1 1	11
-5*	117		
-6	8		8
-7	1		1
-8			
-9		1	1
-10*			

units = .01 points

Figure 6.1

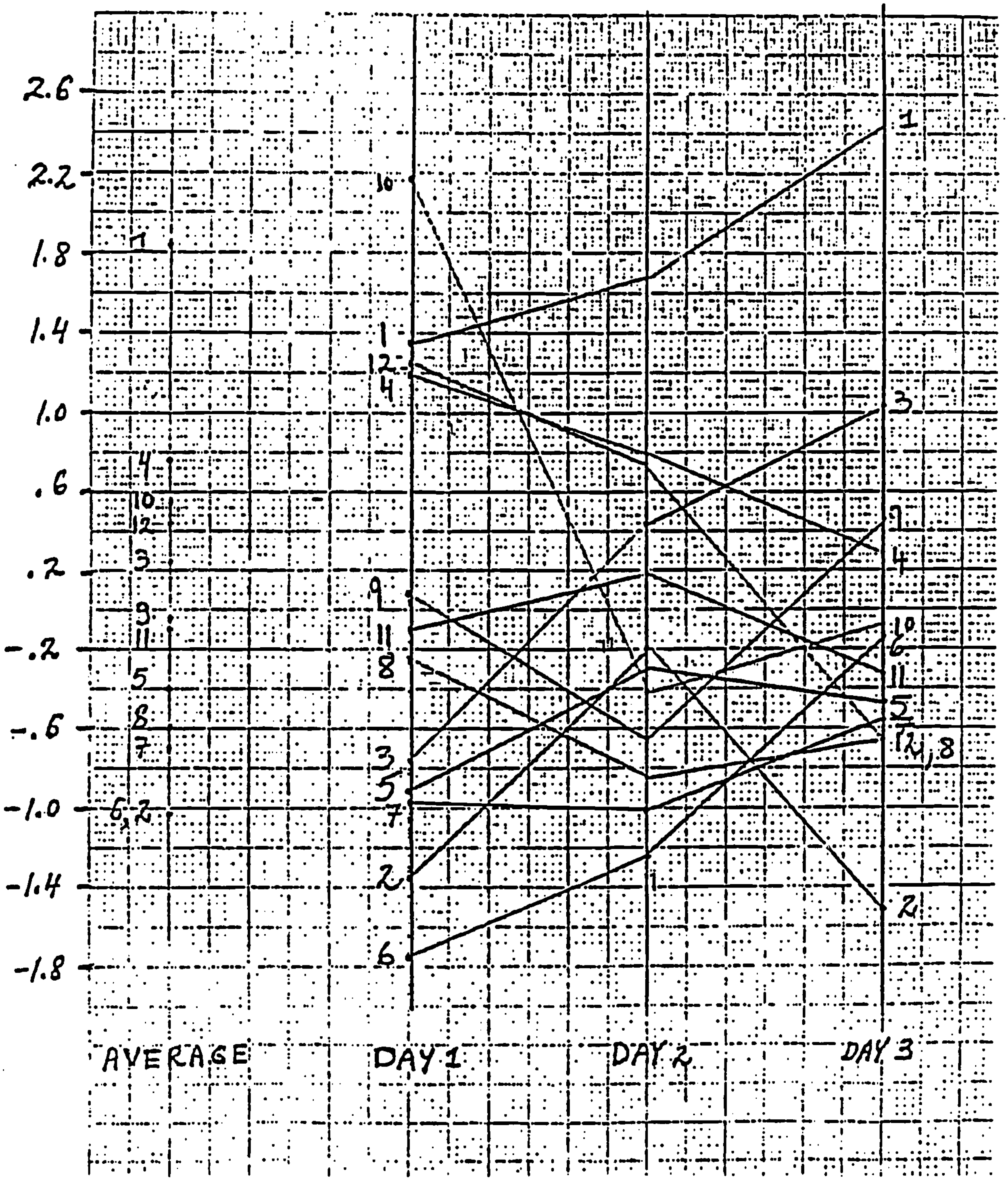
Stem-and-Leaf Display of Estimates Reader Effects

	Essay			
	D	E	X	Y
+18*		3	4	
17				
16				
15*	6			
14				
13	0			
12	8			5
11			6	
10*				
9	5			
8			7 7	6
7		7	3	
6				
5*	2	6		
4		4		
3			9	
2	9	4	8	9
1	1		7 7	4
+0*	8 3			3 3
-0*	8	3 8	5	3
-1	7 1 6		3 6	
-2	8		4 7	9
-3	7			1 7
-4	2	0	2	
-5*	0 0	9		0
-6			8	
-7		0	2	
-8			1	
-9	6		4	
-10*		2 3	1	
-11			3	2
-12	8			
-13	3			

units = .01 points

Figure 6.2

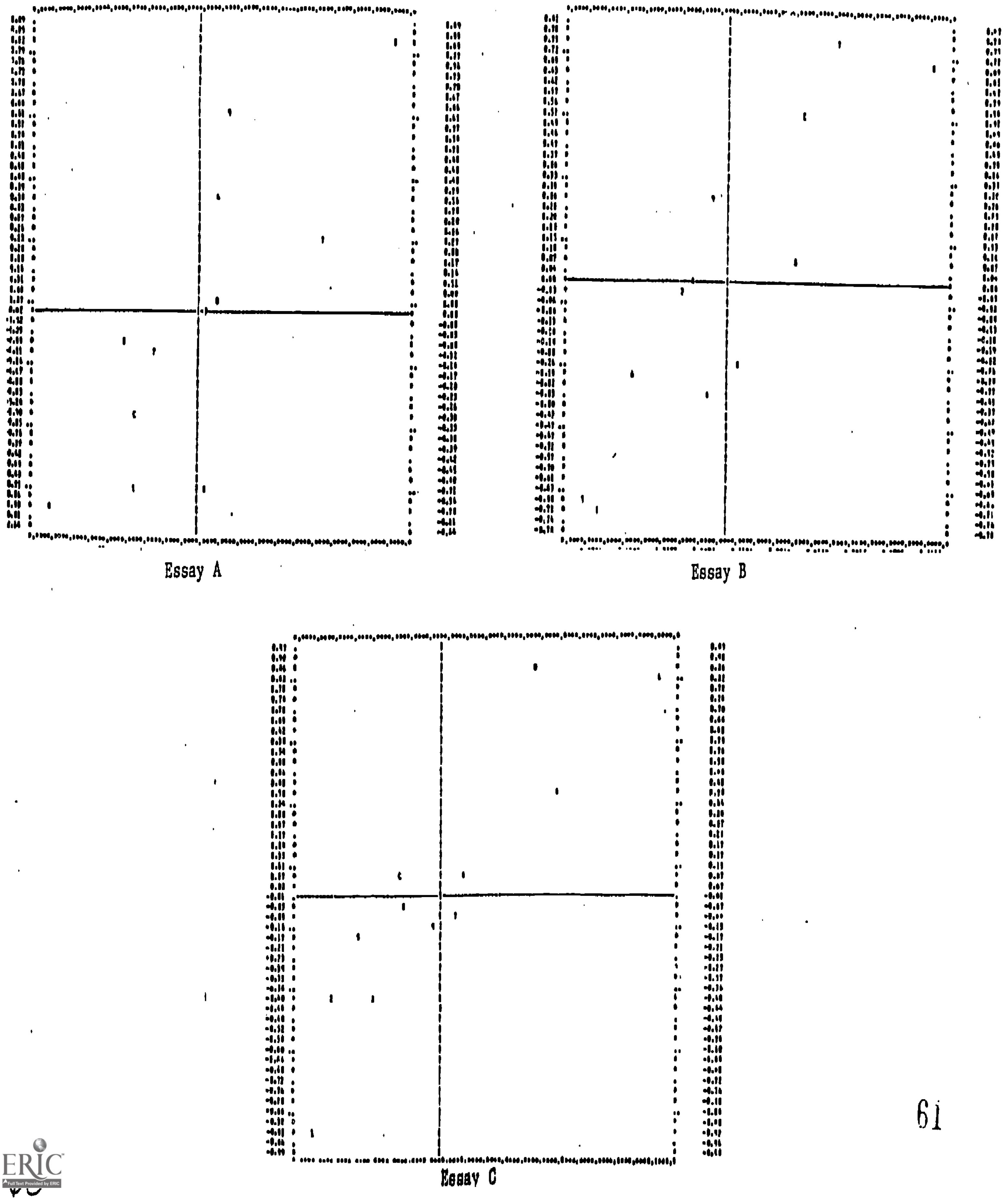
Average Reader Deviations and
Reader Deviations by Day from Day Mean: Essay E



BEST COPY AVAILABLE

Figure 7.1a,b,c

Experimental(Y) vs. Operational(X) Estimated Reader Effects



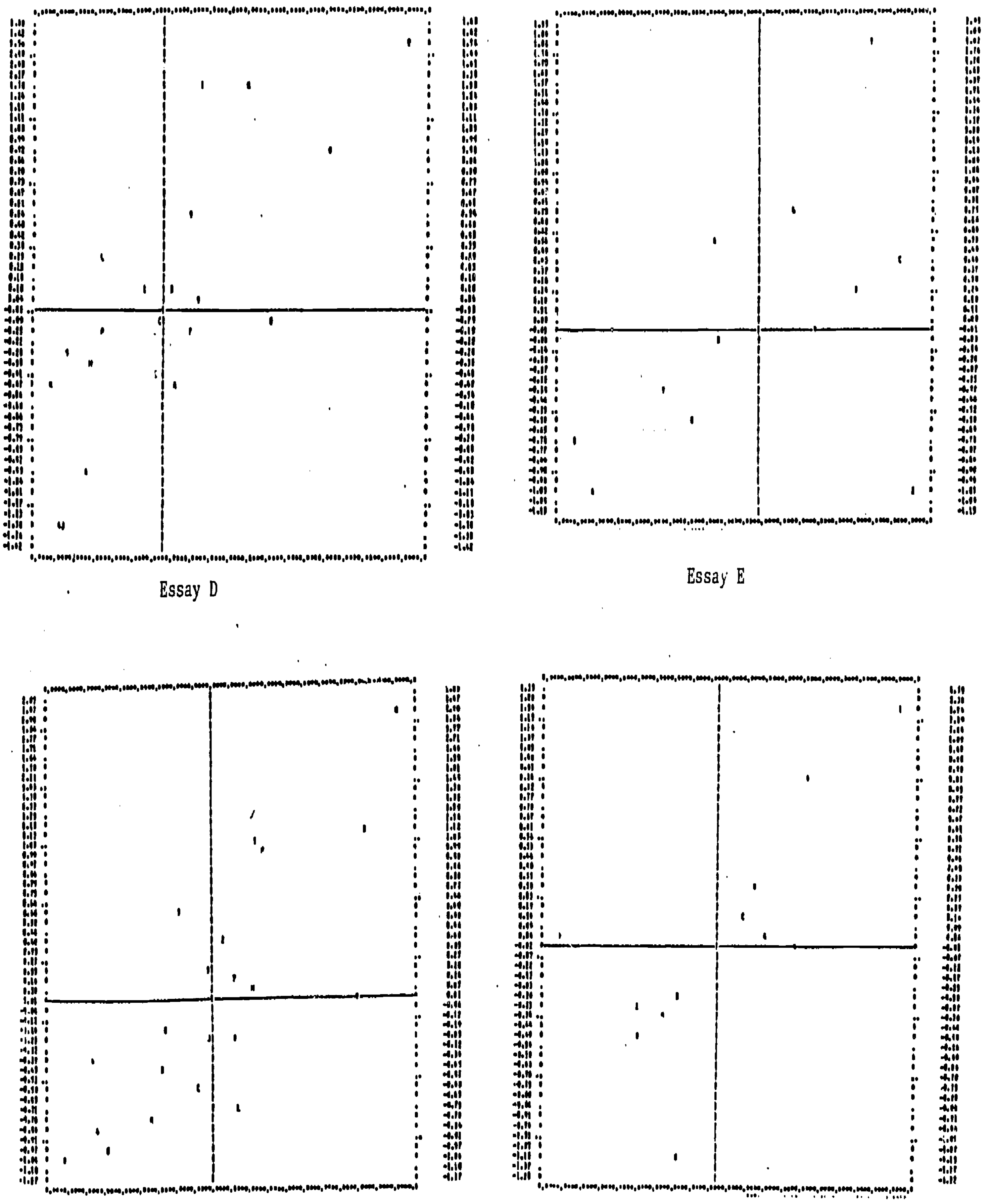
Essay A

Essay B

Essay C

Figure 7.1d,e,f,g

Experimental(Y) vs. Operational(X) Estimated Reader Effects



Essay D

Essay E

Essay X

Essay Y