DOCUMENT RESUME

ED 383 722 TM 023 143

AUTHOR Longford, Nicholas T.
TITLE Reliability of Essay Rating and Score Adjustment. Program Statistics Research Technical Report No. 93-36.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-93-52
PUB DATE Oct 93
NOTE 52p.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Educational Diagnosis; Error of Measurement; *Essays; Estimation (Mathematics); Evaluation Methods; *Interrater Reliability; Models; *Rating Scales; Standardized Tests; *True Scores; *Writing Evaluation

ABSTRACT

A model-based approach to rater reliability for essays read by multiple readers is presented. Variation of rater severity (between-rater variation) and rater inconsistency (within-rater variation) is considered in the presence of between-examinee variation. An additive variance component model is posited and the method of moments for its estimation described. The model involves no distributional assumptions. Minimum mean squared error estimators of examinees' true scores and readers' severities are derived. Model diagnostic procedures are an integral component of the approach. The methods are illustrated on data from standardized educational tests. Six tables and seven figures are included. (Contains eight references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Marilynn Halpern
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

RR-93-52

ED 383 722

Reliabililty of Essay Rating and Score Adjustment

Nicholas T. Longford
Educational Testing Service

PROGRAM STATISTICS RESEARCH

Technical Report No. 93-36

TM023143

Educational Testing Service
Princeton, New Jersey 08541

BEST COPY AVAILABLE

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).

Reliability of Essay Rating and Score Adjustment

Nicholas T. Longford
Educational Testing Service

Program Statistics Research
Technical Report No. 93-36

Research Report No. 93-52

Edcational Testing Service
Princeton, New Jersey 08541

October 1993

Copyright © 1993 by Educational Testing Service. All rights reserved.

Reliability of essay rating and score adjustment

N. T. Longford

Educational Testing Service, Princeton, NJ

September 20, 1993

Abstract

A model-based approach to rater reliability for essays read by multiple readers is presented. Variation of rater severity (between-rater variation) and rater inconsistency (within-rater variation) is considered in presence of between-examinee variation. An additive variance component model is posited and the method of moments for its estimation described. The model involves no distributional assumptions. Minimum mean squared error estimators of examinees' true scores and readers' severities are derived. Model diagnostic procedures are an integral component of the approach. The methods are illustrated on data from standardized educational tests.

Some key words: mean squared error, reliability, shrinkage estimators, variance components.

Acknowledgements

The research leading to this paper was motivated by the Internal Audit at the Educational Testing Service in November 1992. Samuel Livingston, Nancy Petersen, and Rebecca Zwick made useful suggestions on an earlier version of this paper. David Anderson, Jane Faggen, Lee Jones, Behroz Maneckshana, and Rick Morgan provided the datasets analyzed in the paper. Support of the Program Research Planning Council at ETS is acknowledged.

Introduction

Reliability of the scoring process for examinees' responses graded by expert raters (readers) is an important concern in educational testing. Growing reliance on item-types with non-standard format, such as constructed response items and portfolios, requires development of methods for analysis of between-rater differences and for adjustment of scores. Traditionally, these two problems have been treated without their integration; one method was applied to estimate between-rater differences, and another method was used to adjust the scores based on these estimated differences. This paper presents a method in which the two problems are treated integrally. In particular, schemes for adjustment of examinee scores are proposed which are not intermediated by the estimates of characteristics of the individual readers.

Motivated by the generalizability theory (Shavelson and Webb, 1991), we focus on a *population* of readers, rather than the specific readers that happen to have been recruited, and use variance parameters to summarize differences among the readers. An important rationale for this is to make inferences from one set of readers, examinees, and test forms applicable to other settings which can be regarded as draws from the same population. The adjustment schemes are based on shrinkage estimators (Morris, 1983) of the true scores. They incorporate information about the readers and take account of the uncertainty about their characteristics. The main advantages of this approach are in model parsimony, ability to pool information across administrations of tests, and applicability to any noninformative assignment design (of readers to assays).

The approach presented here is similar to that of Braun (1988), and extends it in some aspects, in particular, for multivariate scores and for estimation of true scores. For brevity, the term 'true score' of a response is used for the expected score of the response over the readers in the population from which they have been drawn. Braun (1988) gives a comprehensive review of the literature on scoring reliability. Linacre (1988) and Lunz, Wright, and Linacre (1990) consider

reader severity as an additional facet of the Rasch model for polytomously scored items, and estimate severity of each reader. Rater reliability as a source of measurement error has been extensively studied in medical applications; see, Landis and Koch (1980), Tanner and Young (1985) and Uebersax (1993).

The approach taken here is applicable to both continuous and ordinal categorical scales, and it enables direct estimation of both reader characteristics and examinees' true scores. Approaches based on the classical analysis of variance (ANOVA) or the ordinary regression are oriented toward estimation of the effects associated with the engaged ('realized') readers and examinees, as opposed to the hypothesized populations from which they are drawn. Thus, no inference can be made about a future administration of the same or a similar test form to examinees drawn from the same population and using readers from the same population. Also, ANOVA estimation of the effects associated with the readers and responses is very inefficient when there are a large number of units (readers and/or examinees) because a considerable amount of information provided by the other units is not used. The approach implemented in Linacre (1988) is an adaptation of ANOVA for binary data, and it shares the problems of its classical counterpart. Uebersax (1993) gives a comprehensive review of other approaches based on models akin to item response theory.

In a typical situation an item is administered to I examinees of varying ability, and each response, constructed or performed by an examinee, is scored K times, at most once by each of J readers. An item is an instruction to write an essay, solve a problem, perform a task, or the like. An examinee's response may be documented on paper, computer, videotape, or the like, or observed during the performance. The scoring scale (the range of possible scores) and rubric (the correspondence of the ability, skill, or knowledge to the scale scores) are important components of the item definition. The readers are experts in the subject area and have received extensive training and instruction about the rating process. Rating of recorded responses is organized into sessions; each

session consists of one rating of each response.

Usually, the mean rating given by the K readers who were assigned a given response is adopted as its score. In a simplistic approach the sample intercorrelations among the K readings (sessions) are used as a measure of agreement of the readers. In this paper, an approach to reader reliability based on a variance component model is presented. To motivate the development and to highlight the deficiencies of some of the established approaches, consider the following two extreme assignment schemes:

- Each reader is assigned all the essays in a session ($J = K$)
- Each of the IK readings is done by a different reader ($J = IK$).

A pair of readers is said to be *consistent* if the differences in the scores they give to the responses that both of them have rated are constant. Formally, declare two readers who have rated not more than one response in common as consistent, also. A set of readers is said to be consistent if every pair in the set is consistent.

For a set of consistent readers the sample correlations for the pairs of sessions (i.e., readers) in the first assignment scheme are all equal to unity, but in the second scheme these correlations may be much smaller. This is disconcerting; the sample correlation depends on the assignment design, even though it is supposed to be a characteristic of the rating process.

Two distinct ways in which readers may differ can be readily recognized. The readers may vary in their *severity*; some tend to give higher scores while others tend to give lower scores. Further, readers may disagree on the relative merits of the responses; reader A may rate response x higher than response y, disagreeing with reader B who rates response y higher than response x. Such a disagreement is referred to as *inconsistency*, or reader-by-examinee interaction. Unlike severity, which, in principle, can be corrected by adjusting (*calibrating*) the scores, there is no way of adjusting for inconsistency.

It will be shown that the standard approach to score adjustment based on estimation of reader severity is deficient, because the optimal adjustment depends not only on the estimated severity but also on the amount of information about severity; that is, when severity is poorly estimated it should be given small weight in the adjustment. A direct method for estimation of true scores, which does not rely on estimates of severity, will be presented.

A conceptually useful way of studying the problem of reader reliability is to consider the $I \times J$ matrix $\mathbf{Y} = \{y_{ij}\}$ of scores given to response i by reader j. Usually, most entries of this matrix are not observed (e.g., when each response is rated twice, there are $J-2$ missing observations in each row of $\mathbf{Y}$). Consistency corresponds to constant differences between any two columns of $\mathbf{Y}$. In that case, the ordering of the scores is the same in each column. Departure of the scores from this pattern corresponds to inconsistency.

Each response need not be rated the same number of times. Let κ_i be the set of sessions in which response i was rated, and K_i be the number of these sessions ($1 \leq K_i \leq K$). For instance, when each response is rated once in each of two sessions, $K_i = 2$ and $\kappa_i = \{1, 2\}$ for all responses i.

The readers may be assigned unequal numbers of responses both within and across the K sessions. It will be assumed that the process by which responses are assigned to readers is non-informative in each session, as are the sets $\{\kappa_i\}$, so that the process can be regarded as randomized, subject to the constraint that no essay be read twice by the same reader. Often there are no systematic differences among the sessions, and so they can be regarded as interchangeable. An example to the contrary is likely to arise when, say, the readers are instructed between two sessions to be more lenient.

The paper is organized as follows. The next section describes a variance component model for readers' scores. The between-examinee, within-reader, and between-reader variances are identified as descriptors of the rating process. In the following section the moment method for estimation of these variances is

described. Then, the rater reliability model is expanded to account for varying behaviour of the readers across sessions, for consistent differences among the readers and for multivariate scores. The following two sections present equations for calibration of the readers and for near-optimal estimation of the examinees' true scores using shrinkage estimators. Diagnostic procedures are described in the next section. A section contains several examples in which performance of the proposed adjustment schemes is evaluated. A simulation study is used for generating standard errors and for assessing the impact of uncertainty about the variances on score adjustment. The paper is concluded with a summary.

Variance component model

For the realized scores y_{ij} consider the additive model

$$y_{i,j_{ik}} = \alpha_i + \beta_{j_{ik}} + \varepsilon_{i,j_{ik}}, \tag{1}$$

where j_{ik} is the index for the reader who graded the response of examinee $i = 1, \ldots, I$ in session $k \in \kappa_i$; α_i is the *true score* of examinee i, β_j is the severity of reader $j = 1, \ldots, J$, and ε_{ij} is the residual term interpretable as a reader-by-examinee interaction.

Different (disjoint), overlapping, or identical pools of readers may be used in the sessions. Let J_k be the number of readers used in session k, n_{jk} the number of responses graded by reader j in session k, and n_j the total number of responses graded by reader j, that is, $n_j = \sum_{k=1}^{K} n_{jk}$. The total number of readers used (the size of the reader pool) is denoted by J. Note that $\sum_{k=1}^{K} J_k \geq J \geq \max_k J_k$. Further, let I_k be the number of responses rated in session k, and N the total number of ratings, so that $N = \sum_{k=1}^{K} I_k = \sum_{i=1}^{I} K_i$. A session in which every response is rated, $I_k = I$, is called a complete session. Rating of an essay usually consists of a small number ($K = 1$ or 2) of complete sessions.

It is assumed that $\{\alpha_i\}$, $\{\beta_j\}$, and $\{\varepsilon_{ij}\}$ in (1) are mutually independent random samples with respective means μ, 0, and 0, and variances σ_a^2, σ_b^2, and

σ_e^2. These variances represent variation of examinees' true scores, variation in reader severity, and reader inconsistency, respectively. Various departures from the model assumptions can be considered as a component of σ_e^2.

Given the model in (1), the scores given to the same response by two different readers have the correlation

$$r_1 = \mathrm{cor}(y_{i,j_{i1}}, y_{i,j_{i2}}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_e^2},$$

while a pair of hypothetical scores given to the same response from two independent ratings by the same reader have the correlation

$$r_2 = \mathrm{cor}(y_{i,j_{i1}}, y_{i,j_{i1}'}) = \frac{\sigma_a^2 + \sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_e^2}.$$

The correlation of the true score with the mean of K scores given for a response is

$$\begin{aligned} r_a = \mathrm{cor}\left(\alpha_i, K^{-1}\sum_{k=1}^{K} y_{i,j_{ik}}\right) &= \frac{\sigma_a^2}{\sqrt{\sigma_a^2(\sigma_a^2 + \sigma_b^2/K + \sigma_e^2/K)}} \\ &= \left(1 + \frac{\tau_b + \tau_e}{K}\right)^{-\frac{1}{2}}, \end{aligned} \tag{2}$$

where $\tau_b = \sigma_b^2/\sigma_a^2$ and $\tau_e = \sigma_e^2/\sigma_a^2$ are the *relative* variances of severity and inconsistency, respectively. The correlation of a pair of mean scores in independent replication of the rating process, equal to r_a^2, is often quoted as a measure of the quality of the rating process. For two complete sessions the sample correlation,

$$\hat{r} = \frac{\sum_i (y_{i,j_{i1}} - \bar{y}_1)(y_{i,j_{i2}} - \bar{y}_2)}{\sqrt{\sum_i (y_{i,j_{i1}} - \bar{y}_1)^2 \sum_i (y_{i,j_{i2}} - \bar{y}_2)^2}}$$

($\bar{y}_k$ is the sample mean of the I_k scores given in session k) is considered as an estimator of $r_1 = 1/(1 + \tau_b + \tau_e)$. Hence r_a could be estimated by a suitable transformation as

$$\hat{r}_a = \left(1 + \frac{1 - \hat{r}}{K\hat{r}}\right)^{-\frac{1}{2}}.$$

Figure 1 summarizes the relationship of these two correlations.

Note that the estimator $\hat{r}$ does not involve independent scores because sets of examinees share the same readers. Also, when the designation of the assignment of ratings to sessions is arbitrary, the correlation depends on the selected assignment. The correlation $\hat{r}$ is affected by the assignment design and compounds variation in severity and inconsistency. Identification of these components is essential for improved calibration of readers, estimation of examinees' true scores, and for informed choice of the assignment design.

Estimation

The variance components σ_a^2, σ_b^2, and σ_e^2 are estimated by matching certain sums of squares with their expectations, and these estimates are substituted for the true values in the appropriate expression for a correlation or another quantity.

Define the following statistics:

1. the *within-examinee* sum of squares,

$$S_E = \sum_{i=1}^{I} \sum_{k \in \kappa_i} (y_{i,j_{ik}} - y_{i,\cdot})^2,$$

where $y_{i,\cdot} = \sum_{k \in \kappa_i} y_{i,j_{ik}} / K_i$ is the mean score for response i;

2. the *within-reader* sum of squares for session k,

$$S_{R,k} = \sum_{i \in (k)} (y_{i,j_{ik}} - z_{j,k})^2,$$

where the summation is over all responses rated in session k, and $z_{j,k}$ is the mean score given by reader j in session k,

$$z_{j,k} = \frac{1}{n_{jk}} \sum_{(i;j_{ik}=j)} y_{i,j_{ik}};$$

3. the *total* sum of squares,

$$S_{T,k} = \sum_i (y_{i,j_i,k} - \bar{y}_k)^2$$

where $\bar{y}_k$ is the mean score in session k,

$$\bar{y}_k = \frac{1}{I_k} \sum_i y_{i,j_i,k} .$$

The expectations of these statistics, assuming the model in (1), are

$$\begin{aligned} \mathbf{E}(S_E) &= (N - I)(\sigma_b^2 + \sigma_e^2) \\ \mathbf{E}(S_{R,k}) &= (I_k - J_k)(\sigma_a^2 + \sigma_e^2) \\ \mathbf{E}(S_{T,k}) &= (I_k - 1)(\sigma_a^2 + \sigma_e^2) + \left(I_k - \frac{\sum_j n_{jk}^2}{I_k} \right) \sigma_b^2 . \end{aligned} \tag{3}$$

These expectations are linear functions of the variance components. This is particularly advantageous for the moment matching method described below. To avoid trivial cases, assume that $K \geq 2$, $J \geq 2$, and $N > I$, that is, at least one response is rated more than once. When $J_k = 1$ the expectations $\mathbf{E}(S_{R,k})$ and $\mathbf{E}(S_{T,k})$ coincide. There may be a session with a single reader, $J_k = 1$, but the total number of readers, J, has to be greater than one.

Matching the statistics of S_E, $S_R = \sum_k S_{R,k}$, and $S_T = \sum_k S_{T,k}$ with their expectations leads to a system of three linear equations which has the solution

$$\begin{aligned} \hat{\sigma}_b^2 &= \frac{S_T - S_R(N - K)/(N - \sum_k J_k)}{N - \sum_k \sum_j n_{jk}^2 / I_k} \\ \hat{\sigma}_e^2 &= \frac{S_E}{N - I} - \hat{\sigma}_b^2 \\ \hat{\sigma}_a^2 &= \frac{S_R}{N - \sum_k J_k} - \hat{\sigma}_e^2 . \end{aligned} \tag{4}$$

These variance estimates can, in principle, be negative. Such values are admissible when the variance components are interpreted as certain (conditional) covariances. In practice, it is often meaningful to replace them by zeros.

Extensions

The readers may conduct themselves differently in each session. For example, the model in (1) can be expanded to allow for session-specific severity of each reader. Consider the *average* severity β_j, and a *deviation* $\gamma_{j,k}$ of the reader's severity in session k from the average severity (reader-by-session interaction), so that each score conforms to the model

$$y_{i,j_{ik}} = \alpha_i + \beta_{j_{ik}} + \gamma_{j_{ik},k} + \varepsilon_{i,j_{ik}}, \tag{5}$$

where $\{\gamma_{j,k}\}$ is a random sample from a distribution with mean 0 and variance σ_g^2, independent from the other random variables (α, β, and ε). The variance σ_g^2 represents within-reader *between-session* variation. As a further extension, variances $\sigma_{g,k}^2$ specific to sessions can be considered. Also, it may be meaningful to consider session specific means $\mu_k = \mathbf{E}(y_{i,j_{ik}} \mid k)$, and/or inconsistency variances $\sigma_{e,k}^2$ varying from session to session, so as to accommodate, for instance, higher inconsistency in the first session.

For illustration, assume the variance σ_g^2 to be common to all sessions. In addition to the statistics S_E, $S_{R,k}$, and $S_{T,k}$ define the between-session sum of squares as

$$S_B = \sum_k \sum_j (z_{j,k} - z_j)^2,$$

where z_j is the mean of all the scores given by reader j;

$$z_j = \frac{1}{n_j} \sum_k \sum_{(i j_{ik}=j)} y_{i,j} = \frac{1}{n_j} \sum_k n_{jk} z_{j,k}.$$

The expectation of S_B, assuming the model in (5), is

$$\mathbf{E}(S_B) = \left(\sum_{k=1}^{K} J_k - J \right) \sigma_a^2 + (K-1) J \sigma_g^2 + \left(N - \sum_j \frac{1}{n_j} \sum_k n_{jk}^2 \right) \sigma_e^2. \tag{6}$$

The expectations of the other sums of squares, S_E, $S_{R,k}$, and $S_{T,k}$, are obtained from (3) by replacing σ_b^2 with $\sigma_b^2 + \sigma_g^2$ (the equation for $S_{R,k}$ is unchanged). The estimates of the variances σ_a^2, σ_b^2, σ_g^2, and σ_e^2 are obtained by solving the system of four linear equations that match the statistics S_E, S_R, S_B, and S_T with their (theoretical) expectations. The system of equations can be solved using (4), with σ_b^2 replaced by $\sigma_b^2 + \sigma_g^2$; the estimate of the sum of the variances $\sigma_b^2 + \sigma_g^2$ is then decomposed using (6).

Severity of the readers may depend on extraneous factors, such as the time of the day. Also, it may be desirable to relate severity to readers' characteristics and attributes, such as gender and experience. Such features can be accommodated by replacing the random terms β_j in (5) by a linear regression. If the factors of interest are categorical, the moment matching method of estimation can be supplemented by suitable contrasts of scores which facilitate identification of the effects.

Equations for other extensions of the proposed model, including models that accomodate session-specific severity variance and reader inconsistency, are derived analogously. Such models are likely to be useful only when essays are read a relatively large number of times (e.g., $K > 3$ times) by identical or highly overlapping pools of readers.

Multiple criteria

Readers sometimes score responses to an item for several aspects, such as technical skill, originality, and presentation. The models and the associated methods of estimation have straightforward extensions for multivariate scores; the equations (3) and (6) remain valid, with the variances σ_a^2, σ_b^2, σ_e^2, and σ_g^2 replaced by variance *matrices* $\boldsymbol{\Sigma}_a$, $\boldsymbol{\Sigma}_b$, $\boldsymbol{\Sigma}_e$, and $\boldsymbol{\Sigma}_g$. These matrices are of interest for exploring relationships among the component scores (between examinees, between readers, and within readers). There is no obvious extension for the correlations r_1, r_2 and r_a to the multivariate case, other than defining the correlations of

the component scores.

Adjustment for severity (calibration)

Standard approaches to estimating examinees' true scores, exemplified by Braun (1988), focus on estimation of readers' severity coefficients which would then be used for adjustment of the examinees' scores. The method described here estimates true scores without intermediation of the estimated readers' severities. The advantage of this method can be readily recognized by considering readers assigned extremely small or large workloads. The severity of a reader with a small workload is estimated subject to substantial sampling variation, and therefore adjustment by this 'noisy' quantity is not advisable. On the other hand, adjustment for severity is effective when the severity is well determined. Thus, the amount of *information about severity* of the readers should play an important role in efficient adjustment.

In the following two sections minimum mean squared error estimators of the readers' severities and examinees' true scores are derived. Although in this approach severity estimates are not necessary for estimation of true scores, they are still useful for identifying unusual readers.

Estimating severity

When the severity of reader j is realized on several ratings, the realization of β_j is estimable. The conditional expectation of β_j, given the model parameters and the data, is an obvious estimator of β_j. Evaluation of this expectation involves inversion of the variance matrix for all the ratings, a formidable task without taking advantage of the pattern of the matrix. Complex algebra is involved even if the pattern is appropriately exploited.

An alternative method can be motivated by shrinkage estimation. Consider the following two estimators of the realization of β_j: the trivial estimator identically equal to zero, and the difference of the mean of the ratings given by reader

j and the mean of all the ratings, $z_j - \bar{y}$. When $\sigma_b^2 = 0$, zero is the optimal estimator of the realization of β_j, because the readers do not differ in severity. When σ_b^2 is large, or the workloads n_j are large, $z_j - \bar{y}$ is a good estimator of β_j. The linear combination of these estimators,

$$\hat{\beta}_{j,s} = s_j(z_j - \bar{y}),$$

is adopted, with the reader-specific coefficient s_j which minimizes the the mean squared error (MSE), $\mathbf{E}(\hat{\beta}_{j,s} - \beta_j)^2$.

Denote by $[j,k]$ and $[j]$ the respective sets of responses rated by reader j in session k and in all the sessions; $[j] = \bigcup_k [j,k]$. The model equation (1) implies

$$\begin{aligned} z_j &= \frac{1}{n_j}\sum_{i\in[j]}\alpha_i + \beta_j + \frac{1}{n_j}\sum_{k=1}^{K}\sum_{i\in[j,k]}\varepsilon_{i,j,k} \\ \bar{y} &= \frac{1}{N}\left(\sum_{i=1}^{I}K_i\alpha_i + \sum_j n_j\beta_j + \sum_{i=1}^{I}\sum_{k\in\kappa_i}\varepsilon_{i,j_{ik}}\right), \end{aligned}$$

and elementary algebra yields

$$\begin{aligned} \mathbf{E}(\hat{\beta}_{j,s} - \beta_j)^2 &= s_j^2\sigma_a^2\left(\frac{1}{n_j} - \frac{2}{Nn_j}\sum_{i\in[j]}K_i + \sum_i\frac{K_i^2}{N^2}\right) \\ &+ \sigma_b^2\left\{(1-s_j)^2 + 2s_j(1-s_j)\frac{n_j}{N} + \frac{s_j^2}{N^2}\sum_{j'}n_{j'}^2\right\} + s_j^2\sigma_e^2\left(\frac{1}{n_j} - \frac{1}{N}\right) \\ &= C_{j,0} - 2C_{j,1}s_j + C_{j,2}s_j^2, \end{aligned} \tag{7}$$

where

$$\begin{aligned} C_{j,0} &= \sigma_b^2 \\ C_{j,1} &= \sigma_b^2\left(1 - \frac{n_j}{N}\right) \\ C_{j,2} &= \sigma_a^2\left(\frac{1}{n_j} - \frac{2}{Nn_j}\sum_{i\in[j]}K_i + \sum_i\frac{K_i^2}{N^2}\right) + \sigma_b^2\left(1 - \frac{2n_j}{N} + \frac{1}{N^2}\sum_{j'}n_{j'}^2\right) \\ &\quad + \sigma_e^2\left(\frac{1}{n_j} - \frac{1}{N}\right). \end{aligned}$$

Note that these equations contain totals over all examinees ($\sum_i$), and over all the examinees rated by reader j ($\sum_{i\in[j]}$). The former are common to all readers, but the latter, together with the workloads n_j, may vary among the readers, resulting in different optimal coefficients s_j.

The MSE in (7) has a unique minimum at $s_j^* = C_{j,1}/C_{j,2}$, and the attained minimum is $C_{j,0} - C_{j,1}^2/C_{j,2}$. The coefficient s_j^* can be interpreted as the optimal shrinkage of the deviation $z_j - \bar{y}$ towards zero. When the readers' workloads do not vary a great deal, $\max_j n_j < \sum_j n_j^2/N$, and hence $0 \leq C_{j,1} < C_{j,2}$. Then $0 \leq s_j^* \leq 1$. Values of s_j^* close to zero and unity are attained only in unusual scenarios; for instance $s_j^* = 0$ only if $\sigma_b^2 = 0$.

In practice, the variances are not known and their estimates are used instead. This is problematic in small samples, that is, when the number of readers and/or examinees is small. The simulations discussed below provide some insight into sample size issues. When the variances are known, severity of each reader is estimable even if each response is rated only once. Note how calibration depends on the reader's load n_j; in general, higher load n_j is associated with less shrinkage toward zero.

Estimating true scores

In practice, estimation of a reader's severity is of secondary concern to estimation of the examinees' true scores α_i, although estimation of the coefficients β_j can facilitate this. Simplistic schemes for adjustment of the 'raw' scores $\bar{y}_{i,\cdot} = K_i^{-1} \sum_{k\in\kappa_i} y_{i,j_{ik}}$, are based on various linear combinations of the mean score given by reader j, z_j, and the mean score for all the sessions, $\bar{y}$. Common examples of such adjustment schemes are

$$(\ \hat{\alpha}_i^{(z)} =)\ \ \bar{y}_{i,\cdot} - \frac{1}{K_i} \sum_{k\in\kappa_i} z_{j_{ik}} + \bar{y}$$

and

$$(\hat{\alpha}_i^{(b)} =) \ \bar{y}_{i,.} - \frac{1}{K_i} \sum_{k \in \kappa_i} \hat{\beta}_{j_{ik}} + \bar{y},$$

where $\hat{\beta}_j$ is an estimator of the realization of β_j.

In the literature on rater reliability, the discussion about scoring is often limited to the dilemma of whether to adjust (use $\hat{\alpha}_i^{(z)}$ or $\hat{\alpha}_i^{(b)}$), or not (use $\bar{y}_{i,.}$). A continuum of shrinkage estimators can be defined to fill in the void between these two extremes. For instance, the estimator

$$\hat{\alpha}_{i,u} \ = \ \bar{y}_{i,.} - u \left(\frac{1}{K_i} \sum_{k \in \kappa_i} z_{j_{ik}} - \bar{y} \right), \tag{8}$$

with a constant u, corresponds to no adjustment when $u = 0$, and to (full) adjustment by the mean of the K readers' means who rated the response i when $u = 1$. It is shown below that intermediate values of u yield more efficient estimators. Moreover, different coefficients u can be used for the examinees; more shrinkage is appropriate for responses rated by readers who had small workloads because the ratings contain less information about the severity of their readers.

Another class of estimators of α_i is given by the equation

$$\hat{\alpha}_{i,t} \ = \ (1-t)\bar{y}_{i,.} + \frac{t}{K_i} \sum_{k \in \kappa_i} z_{j_{ik}}, \tag{9}$$

where t is a constant. It does not have as appealing a motivation as (8); the estimator adjusts the scores by shifting them closer to the reader's means. Nevertheless, in several examples analyzed below $\hat{\alpha}_{i,t}$ performs better than $\hat{\alpha}_{i,u}$. The coefficients u in (8) and t in (9) can be set so as to minimize the expected squared error. Before determining these coefficients, consider a more general scheme based on the class of estimators formed as the linear combinations of the statistics $\bar{y}_{i,.}$, z_j, and $\bar{y}$:

$$\hat{\alpha}_i \;=\; v_{1i}\bar{y}_{i,\cdot} - \frac{v_{2i}}{K_i}\sum_{k\in\kappa_i} z_{j_{ik}} + v_{3i}\bar{y}, \tag{10}$$

where v_{hi}, $h = 1, 2, 3$, are the examinee-specific coefficients, such that

$$v_{1i} - v_{2i} + v_{3i} = 1. \tag{11}$$

This constraint is necessary to ensure unbiasedness, that is, $\mathrm{E}(\hat{\alpha}_i - \alpha_i) = 0$. The shrinkage coefficients v_{hi} are chosen so as to minimize the MSE for the true score α_i. Although the algebra involved appears tedious it is elementary.

Let $r_{ii',k}$ be equal to 1 if responses i and i' were graded by the same reader in session k, and equal to 0 otherwise. Further, let $n_i^+ = \sum_{k\in\kappa_i} n_{j_{ik}}/K_i$, $n_i^- = \sum_{k\in\kappa_i} n_{j_{ik}}^{-1}/K_i$, $n^{(2)} = \sum_j n_j^2/N$, and

$$R_i \;=\; \frac{1}{K_i}\sum_{i'=1}^{I}\left(\sum_{k\in\kappa_i}\frac{r_{ii',k}}{n_{j_{ik}}}\right)^2.$$

Note that $1 \le n_i^+ \le I$ and $1 \ge n_i^- \ge 1/I$. The MSE of $\hat{\alpha}_i$ is

$$\begin{aligned}\mathrm{E}(\hat{\alpha}_i - \alpha_i)^2 \;=\;& \\ &\left\{(1-v_{1i})^2 + 2(1-v_{1i})\left(v_{2i}n_i^- - v_{3i}\tfrac{K_i}{N}\right) + v_{2i}^2\tfrac{R_i}{K_i} - v_{2i}v_{3i}\tfrac{2K_i}{N} + v_{3i}^2\tfrac{K_i}{N}\right\}\sigma_a^2 \\ &+\left\{(v_{1i}-v_{2i})^2\tfrac{1}{K_i} + 2(v_{1i}-v_{2i})v_{3i}\tfrac{n_i^+}{N} + v_{3i}^2\tfrac{n^{(2)}}{N}\right\}\sigma_b^2 \\ &+\left\{v_{1i}^2\tfrac{1}{K_i} + 2(v_{1i}-v_{2i})v_{3i}\tfrac{1}{N} - (2v_{1i}-v_{2i})v_{2i}\tfrac{n_i^-}{K_i} + v_{3i}^2\tfrac{1}{N}\right\}\sigma_e^2. \end{aligned}\tag{12}$$

This is a quadratic function in the coefficients v_{hi}. Assuming non-negative variances σ_a^2, σ_b^2, and σ_e^2 the coefficients with the quadratic terms v_{hi}^2 are all positive. Therefore, (12) has either a unique minimum or a continuum of minima located on a straight line. Differentiation with respect to the coefficients v_{hi} yields the linear system of three equations

$$\begin{aligned} v_{1i}A_{11} - v_{2i}A_{12} + K_i v_{3i}A_{13}/N &= K_i\sigma_a^2 \\ v_{1i}A_{12} - v_{2i}A_{22} + K_i v_{3i}A_{13}/N &= K_i n_i^-\sigma_a^2 \\ v_{1i}A_{13} - v_{2i}A_{13} + v_{3i}A_{33} &= K_i\sigma_a^2, \end{aligned}\tag{13}$$

BEST COPY AVAILABLE

where

$$\begin{aligned} A_{11} &= K_i\sigma_a^2 + \sigma_b^2 + \sigma_e^2 \\ A_{12} &= K_i n_i^- \sigma_a^2 + \sigma_b^2 + n_i^- \sigma_e^2 \\ A_{13} &= K_i\sigma_a^2 + n_i^+\sigma_b^2 + \sigma_e^2 \\ A_{22} &= R_i\sigma_a^2 + \sigma_b^2 + n_i^-\sigma_e^2 \\ A_{33} &= (N\sigma_a^2 + K_i n^{(2)}\sigma_b^2 + K_i\sigma_e^2)/N. \end{aligned}$$

The constraint in (11) can be enforced either by substitution, or by application of the Lagrange multipliers. When the variance of $\bar{y}$ is negligible in comparison with that of z_j and $\bar{y}_i$, e.g., when there are few sessions and many readers, application of the constraint corresponds to adjustment of v_{3i}, with negligible adjustments of v_{1i} and v_{2i}. Thus, it suffices to solve the first two equations in (13) and then set $v_{3i} = 1 - v_{1i} + v_{2i}$. When a shrinkage estimator $\hat{\beta}_{j,s}$ is used instead of z_j in (10), these two adjustment schemes coincide because $\hat{\beta}_{j,s}$ is a linear function of z_j and $\bar{y}$.

Several problems with the solution of (13) are readily recognized: difficulty with interpretation of the optimal coefficients v_{hi}, lack of any insight into the dependence of the coefficients on the variance components, sampling variation of the coefficients, and the amount of reduction of the MSEs over simpler schemes. These issues are discussed below using several examples.

For the more restrictive schemes given by (8) and (9) the equations for the MSE are substantially simpler. For (8),

$$\begin{aligned} \mathbf{E}\left\{(\hat{a}_{i,u} - \alpha_i)^2\right\} &= u_i^2\left(\frac{R_i}{K_i} - \frac{2K_i}{N} + \frac{K_i^2}{N}\right)\sigma_a^2 \\ &\quad + \left\{(1-u_i)^2\frac{1}{K_i} + 2u_i(1-u_i)\frac{n_i^+}{N} + u_i^2\frac{n^{(2)}}{N}\right\}\sigma_b^2 \\ &\quad + \left\{\frac{1}{K_i} + u_i(2-u_i)\left(\frac{1}{N} - \frac{n_i^-}{K_i}\right)\right\}\sigma_e^2 \end{aligned}$$

$$= \quad D_{0,i} - 2D_{1,i}u_i + D_{2,i}u_i^2, \tag{14}$$

where

$$D_{0,i} = \frac{\sigma_b^2 + \sigma_e^2}{K_i}$$

$$D_{1,i} = \sigma_b^2\left(\frac{1}{K_i} - \frac{n_i^+}{N}\right) + \frac{\sigma_e^2}{K_i}\left(\frac{n_i^-}{K_i} - \frac{1}{N}\right)$$

$$D_{2,i} = \sigma_a^2\left(\frac{R_i}{K_i} - \frac{2K_i}{N} + \frac{1}{N^2}\sum_{i'=1}^{I} K_{i'}^2\right) + \sigma_b^2\left(\frac{1}{K_i} - 2\frac{n_i^+}{N} + \frac{n^{(2)}}{N}\right)$$
$$+\sigma_e^2\left(\frac{n_i^-}{K_i} - \frac{1}{N}\right).$$

When $D_{2,i} > 0$ the MSE in (14) has a unique minimum at $u_i^* = D_{1,i}/D_{2,i}$, and its minimum value is $D_{0,i} - D_{1,i}^2/D_{2,i}$. When the readers' workloads do not vary substantially, $D_{2,i} \geq D_{1,i} \geq 0$, and then $0 \leq u_i^* < 1$; u_i^* can be interpreted as a shrinkage coefficient. No adjustment ($u_i^* = 0$) is optimal when $n_{j_{ik}} = I$ for all k (one reader per session rating all the responses). When each reader rates every response, $D_{2,i} = D_{1,i} = 0$, and so the MSE in (14) is constant. In that case adjustment is superfluous.

Note that the examinee variance σ_a^2 is not involved in $D_{1,i}$. Since the coefficient of σ_a^2 in $D_{2,i}$ is positive, higher σ_a^2 (all else held equal) leads to smaller u_i^* (less shrinkage). On the other hand, higher between-reader (severity) variation and higher inconsistency are associated with more shrinkage.

In the administration of a typical large-scale testing program there are a large number of examinees and readers, and each response is rated a small number of times (once or twice). Each reader has a substantial workload n_j, and so the scores contain abundant information about his/her severity. If readers' severities are variable, adjustment with high u_i is likely to be better than no adjustment. Note that even when $\sigma_b^2 = 0$ the optimal adjustment is for $u_i > 0$. However, when the workloads n_j are much smaller than the number of examinees I, the coefficient of σ_b^2 in $D_{1,i}$ is much smaller than that in $D_{2,i}$ (the coefficients of σ_e^2

are equal). Therefore, adjustment is more important (u_i^* is larger) the larger the severity variance σ_b^2.

The estimators of the true scores α_i based on the coefficients u_i^* break down in some extreme scenarios. For example, when $\sigma_a^2 = 0$ the optimal estimator of each examinee's score α_i is the sample mean $\bar{y}$. When the readers are perfectly consistent, $\sigma_e^2 = 0$, and each pair of readers can be linked through responses (reader A is linked with reader B if there are readers C, D, E, ..., Z, such that there is at least one response each rated by the pairs of readers A and C, C and D, D and E, ..., and Z and B), then the true scores α_i can be determined exactly. In these cases the estimator $\hat{\alpha}_{i,u}$ is very inefficient.

The estimators given by (9) can be analyzed by the same approach. The mean squared error is

$$\begin{aligned} \mathbf{E}\left(\hat{\alpha}_{i,t} - \alpha_i\right)^2 &= \tfrac{\sigma_b^2+\sigma_e^2}{K_i} \\ -2t_i\tfrac{\sigma_e^2}{K_i}(1-n_i^-) + t_i^2\left\{\sigma_a^2\left(1-2n_i^- + \tfrac{R_i}{K_i}\right) + \tfrac{\sigma_e^2}{K_i}(1-n_i^-)\right\} \\ &= E_{0,i} - 2E_{1,i}t_i + E_{2,i}t_i^2\,, \end{aligned} \tag{15}$$

for implicitly defined coefficients $E_{k,i}$, and so its unique minimum is attained at

$$t_i^* = \frac{\sigma_a^2(1-n_i^-)}{\sigma_a^2(K_i - 2K_i n_i^- + R_i) + \sigma_e^2(1-n_i^-)}.$$

Clearly $0 \le t_i^* \le 1$. When the readers are very inconsistent, that is, σ_e^2 is much larger than σ_a^2, t_i^* is close to unity. When σ_e^2 is much smaller than σ_a^2, t_i^* is close to zero, and so the optimal adjustment is minute. This is in agreement with intuition. It is interesting, though, that the optimal coefficient t_i^* does not depend on σ_b^2. This is of some importance because σ_b^2 is usually the least precisely estimated variance component.

Diagnostic procedures

The readers may display any of a whole gamut of behaviours different from that assumed by the model in (1). In this section, informal procedures for detecting

some types of departure from the model in (1) are discussed.

Homogeneity of the inconsistency deviations ε_{ij} is an important assumption in (1). Define the sum of squares within reader j as the subtotal within $S_{R,k}$ corresponding to the reader j;

$$S_{R,jk} = \sum_{i \in (j,k)} (y_{i,j,k} - z_{j,k})^2 . \tag{16}$$

Assuming normality of the scores $y_{i,j}$, $S_{R,jk}/(\sigma_a^2 + \sigma_e^2)$ has the χ^2 distribution with $n_{jk} - 1$ degrees of freedom. For large n_{jk} its distribution is *approximately* $\chi^2_{n_{jk}-1}$, even when the scores are not normally distributed. The statistics $S_{R,jk}$ can be pooled over sessions k, but not over readers because the corresponding statistics are not independent. Thus, checking for variance homogeneity involves comparison of the statistics $S_{R,jk}$ or their combinations with the critical values of the appropriate χ^2 distributions. Unlike in other uses of the χ^2 distribution both very large and very small values of the statistics are evidence against the model. Small values are a sign that the reader is giving almost the same grade for every response.

A finer insight, relevant for large scale tests, is enabled by considering responses rated by the same pair of readers. The variance of the differences $y_{i1} - y_{i2}$ for such a set of responses is $2\sigma_e^2$. The corresponding (within-pair) sums of squares can be pooled, because they are independent, thus generating statistics which can be compared with the critical values of the corresponding χ^2 distributions. Readers who tend to disagree with their fellow-readers, or agree with them, more than would be expected can be identified from these statistics.

These two sets of statistics imply a general diagnostic method based on defining suitable subtotals of the sums of squares S_E, S_R, S_T, and others, if applicable, which have conditionally independent summands. If the number of terms in these totals is moderate to large, the totals are approximately χ^2 distributed. The sums of squares selected should reflect the principal concerns about model violations.

These diagnostic procedures rely heavily on the assumption of non-informative allocation of responses to readers. For instance, it is difficult to distinguish between a reader who was assigned lower quality responses from a reader with high severity, in particular when each response is rated a small number of times.

Reliability

In the approach presented here the mean squared error appears to be a more natural metric for assessment of the rating process than the traditionally applied correlation coefficients, such as r_a, and r_a^2. Counterparts of these correlations can be defined for adjusted scores, for instance, by replacing the unadjusted score in (2) with the adjusted score.

For the optimal shrinkage coefficients u^* and t^* these correlations are:

$$r_{a,iu} = \mathrm{cor}(\alpha_i, \hat{\alpha}_{i,u^*}) = \frac{\sqrt{\sigma_a^2}}{\sqrt{D_{0,i} - D_{1,i}^2/D_{2,i}}} \left\{1 - u^* \left(1 - \frac{K_i}{N}\right)\right\}$$

$$r_{a,it} = \mathrm{cor}(\alpha_i, \hat{\alpha}_{i,t^*}) = \frac{\sqrt{\sigma_a^2}}{\sqrt{E_{0,i} - E_{1,i}^2/E_{2,i}}}, \qquad (17)$$

where $E_{0,i}$, $E_{1,i}$, and $E_{2,i}$ are the absolute, linear, and quadratic coefficients of t_i in the right-hand side of (15). Note that unlike r_a, the correlations $r_{a,iu}$ and $r_{a,it}$ are not constant across the responses i, unless a balanced assignment design is employed. Equation (17) implies that for every response i the correlation $r_{a,it}$ is greater than r_a, but $r_{a,iu}$ may not be. The counterparts of r_a^2 are the correlations $r_{a,iu}^2$ and $r_{a,it}^2$. The corresponding equations for the general adjustment scheme based on (10) are not tractable.

Examples

Advanced Placement Biology test

The Advanced Placement Biology test contains a large number of multiple-choice items and four constructed response items (essays) that are rated by

expert readers. The dataset, drawn from an experimental administration in 1992, comprises scores on the multiple-choice items and the four essays for 297 examinees. Each essay was rated by two readers; a different set of readers was used for each essay. Most readers rated 24 or 25 essays in each of the two sessions. There are a few exceptions when a reader in the first session was apparently replaced by a different reader in the second session. The allocation of readers to essays is almost balanced; most pairs of readers rated five responses in common. The rating scale is 0 – 10 (integer scores).

The estimates of the variance components are given in the left-hand side of Table 1. For all four items reader inconsistency dominates variation in severity. In fact, for essay C the estimated severity variance is essentially zero. That limits the scope of adjustment. The reduction of the MSE due to adjustment is modest, though perceptible, for all examinees and all essays.

Note that all conclusions related to MSEs made in this section are contingent on the assumption that σ_a^2 (variance of the true scores), σ_b^2 (severity variance), and σ_e^2 (inconsistency variance) are known and equal to their estimates. This assumption is subjected to scrutiny in the next section.

The estimated mean squared errors are summarized in Table 1. For the unadjusted scores (column 'NAdj') the MSEs are constant across the examinees. For the adjustment schemes based on (8) ('uAdj'), (9) ('tAdj'), and (10) ('AAdj') the ranges of the MSEs are given for each item. For instance, using the adjustment scheme 'AAdj' the estimated mean squared errors for item C are in the range 0.361 – 0.367.

For essay A the MSE for the raw mean score (no adjustment) is 1.37, the MSEs using u_i^* are equal to 1.32 ($u_i^* \doteq 0.30$), and the MSEs using t_i^* are equal to 1.22. The adjustment scheme tAdj is better than uAdj for three of the essays, although the reduction in MSE is greater than 0.1 only for essay A. For essay A, the adjustment AAdj yields further reduction of the MSE by 0.07. Score adjustment for essays B and D is useful but neither restrictive adjustment

scheme (uAdj or tAdj) approaches the efficiency of the general scheme (AAdj). For essay C, gains by adjustment are marginal.

With incomplete (or no) information about the variances, the choice of the shrinkage factor is fraught with danger, as is the more discrete choice of whether to adjust the scores at all. For illustration of this problem Figures 2 and 3 contain plots of the MSEs for the adjusted scores using (8) and (9), as functions of the respective coefficients u_i and t_i. For essays A and D these functions are within a narrow band for all responses, and therefore only the function for an arbitrarily chosen response is plotted. For essays B and C there were responses rated by readers with small workloads, and so their MSEs as functions of the shrinkage coefficient differ from the rest of the responses. The MSEs for such responses, number 7 for essay B and number 102 for essay C, are plotted together with arbitrarily selected responses from the rest.

For essay A the choice of the shrinkage coefficient is not crucial; not much reduction of the MSE can be achieved, although full adjustment, $u = 1$, is clearly worse than no adjustment, $u = 0$. For a response to essay B rated by a reader with a small workload, full adjustment is extremely risky because the reader's mean z_j is based only on one response. There is an interesting contradiction; the minimum MSE for this response is smaller than the MSE for responses rated by readers with the usual workload. Better adjustment should be achieved when more information is available about the readers. Here this is not the case; this is another sign that the adjustment scheme is not fully efficient. The same phenomenon can be observed for essay C for which one reader with a small workload was engaged.

Figure 3 contains the corresponding plots for the adjusted scores using (9) as functions of the coefficient t_i. The contradiction observed for the scheme based on (8) arises here also. The choice of the shrinkage coefficient is somewhat more important; as in Figure 2, the largest meaningful adjustment, corresponding to $t_i = 0.5$, is detrimental.

The example of these four essays suggests that indiscriminate use of (full) adjustment is very risky. Two factors contributing to this outcome are relatively small sample size (numbers of readers and their workloads) and small variation of reader severity. Thus, without considering shrinkage estimators, no adjustment is preferable to full adjustment. The simpler adjustment schemes based on (8) and (9) improve estimation of the true scores somewhat, but the scheme based on (10) is clearly superior. Note that for the simpler schemes the mean squared error of the adjusted score is smaller than the raw score only in a narrow range of the coefficients u or t around the optimal coefficient u^* (or t^*).

Large differences among the MSEs for the essays suggest that combining the four (adjusted) scores into a single score should be done in a weighted fashion reflecting differential reliability of the scores.

Studio Art Portfolio Assessment

The Advanced Placement Studio Art test comprises a portfolio assessment in which artwork submitted by each examinee is rated on several criteria. Each of the six criteria, denoted A – F, is subjectively rated on the scale 0 – 4 by raters from the same pool. A score of zero is very rare; in most sessions it is received by less than one per cent of the 3756 examinees. The criterion A is rated by three different raters and the other criteria by two raters each. A rater may assess a portfolio on several criteria, but no criterion is assessed twice by the same rater. The workloads of the raters within criteria vary considerably; apart from a few raters who rated fewer than four portfolios the workloads are in the range 120 – 600.

Table 2 contains the variance estimates and a summary of the MSEs using the adjustment schemes uAdj, tAdj, and AAdj. Relative to variation of the true scores (σ_a^2), the reader inconsistency (σ_e^2) is very large. For the Advanced Placement Biology test the inconsistency variance σ_e^2 is less than 30 per cent of the total variance $\sigma_a^2 + \sigma_b^2 + \sigma_e^2$ for each essay; here the inconsistency variance is

around 40 per cent for all six criteria. Here, as in the Biology test, inconsistency variation dominates variation in severity (σ_b^2). Nevertheless, a considerable reduction of MSE is achieved by the adjustment schemes. There is little to choose between the schemes uAdj and tAdj, but AAdj yields a substantial additional improvement.

Reader inconsistency is the principal cause of low score reliability. In principle, the variation could be reduced by further training and instruction of the readers. It is instructive to consider two components of inconsistency: disagreement in the merit of the rated material and variation in the grades given by the same reader in a hypothetical independent replication of the rating. These components may be reduced by training and instruction, by allowing the raters more time, and the like. Since rating of a criterion by the same reader cannot be replicated, one can only speculate about the relative contributions of these two causes to inconsistency variation.

For illustration of model diagnostics the scores for criterion F are explored further. Of the 33 readers who took part, nine rated a total of 15 portfolios, and so no meaningful diagnostics for these readers can be generated. The remaining 24 readers had workloads of 143 – 533 responses. The standardized within-reader sample variances $S_{R,jk}/\{n_{jk}(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)\}$ are displayed in Table 3 (first two columns) together with their aggregates over the two sessions, $\sum_k S_{R,jk}/\{n_j(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)\}$ (third column). The expectations of these statistics are equal to unity, and their standard deviations to $\sqrt{2/\{n_{jk}(\sigma_a^2 + \sigma_e^2)\}}$ and $\sqrt{2/\{n_j(\sigma_a^2 + \sigma_e^2)\}}$, respectively. For orientation, $\sqrt{2/(\hat{\sigma}_a^2 + \hat{\sigma}_e^2)} \doteq 2.2$. Most of the statistics in Table 3 are within their theoretical standard deviations of unity. Also, the statistics for the two sessions are very close to one another for several readers. Two readers, 112 and 190, stand out; their statistics are large in the first session, but close to unity in the second. Reader 190 had a small workload in the first session (18 responses), and so $S_{R,j1} = 1.82$ does not present strong evidence that the reader is unusual. However, $S_{R,j1} = 1.85$ for reader 112 is a

strong indication of departure from the model because the standard deviation associated with $S_{R,j1}$ is equal to $2.2/\sqrt{183} = 0.163$. Reader 125 has high values of $S_{R,jk}$ in both sessions. There is little evidence that any of the readers give the same score to almost everyhody (small values of the χ^2 statistics); reader 117 has the smallest value of these statistics (0.53 in session 2).

The variation in the score differences for responses rated by the same pair of readers can be treated similarly. The within-pair statistics can be accumulated over the readers, thus generating a χ^2-type statistic for each reader. If these statistics are aggregated within sessions, the components are independent. If they are aggregated across sessions, they are correlated because an elementary statistic is included for both readers. However, the correlations are small because the severity variance is small. For the criterion F there are no outlying readers for either session, or for the aggregates across sessions; no reader can be identified who is in exceptional agreement/disagreement with the fellow readers. For brevity, details are omitted.

CAPA tests

The two examples analysed above suggest that inconsistency variation should be the principal concern with the readers. This is not a surprising conclusion; severity of the readers is 'anchored' by common expectations as well as by the scoring rubric. Being well-acquainted with the quality of the examinees, the readers make sure that they do not give extreme grades to too few or to too many examinees.

The English Language and Literature (ELL) and Social Science (SSc) tests are two components of the Content Area Performance Assessments (CAPA). CAPA were developed jointly by Educational Testing Service and the California Commission on Teacher Credentialing (CTC). CAPA in conjunction with a battery of multiple-choice NTE Specialty Area tests is used for teaching certification in California. The data analyzed here are from the November 1992

(operational) administration in which each test was taken by about 400 examinees.

The ELL and SSc tests contain two essays each, denoted ELL1, ELL2, SSc1, and SSc2, each rated by a pair of readers. The two tests use disjoint pools of readers, but the pools for the pair of essays within each test are identical. The scoring scale is 1 – 6 for each essay, and the 18 fully participating readers in each test rated between 36 and 82 responses. A small number of other readers rated not more than four responses each.

Results of the essay scoring analysis are summarized in Table 4. The estimated severity variances (σ_b^2) are about a sixth (ELL1) to a quarter (the other three essays) of the estimated inconsistency variances (σ_e^2). The reduction in MSE due to score adjustment is modest, though, because the readers' workloads are small.

The impact of the score adjustments can be assessed by summarizing the adjustments. Figure 4 contains the histograms of the adjustments $\hat{\alpha}_{i,u} - \bar{y}_{i,\cdot}$, $\hat{\alpha}_{i,t} - \bar{y}_{i,\cdot}$, and $\hat{\alpha}_i - \bar{y}_{i,\cdot}$ for the respective schemes uAdj, tAdj, and AAdj. The sample variances of these adjustments are 0.013, 0.016, and 0.027 (the sample means are within 0.01 of zero for each scheme); the scheme AAdj has the largest adjustments, uAdj the smallest. The adjustments for tAdj and AAdj are highly correlated ($\rho = 0.77$), while the adjustments for uAdj have lower correlations with both tAdj and AAdj ($\rho \doteq 0.50$).

The adjustments are in the range –0.40 – 0.50, with 92.5 per cent of the adjustments in the range –0.25 – 0.25; the coarseness of the rating scale remains transparent even after either of the three kinds of adjustments. For instance, most adjusted scores rounded to the nearest half-integer are equal to the raw scores. The coarseness of the adjusted scores is also readily observed in the plots of the adjusted scores against the subscores for the multiple-choice part of the test, drawn in Figure 5. The correlations of the scores from essay ELL1 with the multiple-choice score are in the range 0.460 – 0.468, lowest for the unadjusted

scores and highest for uAdj and AAdj. Such a small change in the correlation could not be interpreted as an improvement in validity even if the essay and the multiple-choice part of the test were known to have a common unidimensional underlying trait.

Standard errors

The variance components have an important role in estimation of the true scores α_i. In all three adjustment schemes the *estimated* variance components are used instead of the unknown parameter values. Therefore, it is important to establish the sampling variation of the variance components, and the dependence of the adjustments on the variance components. The latter is relatively straightforward for the two restrictive schemes (uAdj and tAdj), but not for AAdj. This reduces somewhat the efficiency of the scheme AAdj for small administrations in which the variances are estimated subject to a lot of uncertainty.

Although feasible, derivation of the sampling variance matrix of the statistics S_E, S_R, and S_T is extremely tedious, unless $\sigma_b^2 = 0$. In any case, the variance matrix, and therefore the distribution of the estimators of the variances σ_a^2, σ_b^2, and σ_e^2 depends on unknown kurtoses of the random terms α_i, β_j, and ε_{ij}.

Since estimation of the variances is relatively simple the standard errors for the variances can be estimated by simulation. For a given assignment design, say, that for essay B in the Advanced Placement Biology test, the observed scores are replaced by those generated by the fitted model (with randomly drawn 'examinees' and 'readers'). All the random variables are drawn from normal distributions with variances set to their estimated values from the real dataset ($\sigma_a^2 = 3.74$, $\sigma_b^2 = 0.45$, and $\sigma_e^2 = 1.40$). The effect of rounding and truncation (to scores 0, 1, ..., 10) can be explored in the same study by reestimating the variances with the generated (normal) scores rounded and truncated.

The results are summarized in Table 5. For each variance its true value and mean and the standard deviation of the 200 simulated estimates are given. The

estimates are evaluated for the normally generated data (left-hand side of the table) and for their rounded and truncated versions (right-hand side).

The results for the normal scores indicate that the estimators of the variances are unbiased. For the rounded scores the distribution of the estimators is somewhat different, although the moments of $\hat{\sigma}_e^2$ are only slightly affected by rounding. On average, about 13 scores (out of 297) were truncated in a session. These scores are likely to contribute to reduction of the means of the variance estimates. The results in Table 5 can in no way be generalized. For instance, when the scoring rubric is coarser, as in the CAPA tests, the variance $\hat{\sigma}_e^2$ is affected. Results of the same simulation study for essay ELL1 in CAPA are presented in Table 6, in the same format as Table 5. Now the rounding causes a slight inflation of the variance estimator σ_e^2. Thus, to get a rough idea of the sampling variation of the variance estimators in simulations rounding can be ignored. Figure 6 contains the histograms of the six sets of estimators; in the top row the histograms for the normal scores, and in the bottom row those for the rounded scores are drawn. The shapes of the sampling distributions of the variance estimators are only moderately skewed.

The simulated data contain abundant information about the examinee variance σ_a^2 and the inconsistency variance σ_e^2. However, for small administrations estimation of the reader severity variance σ_b^2 is clearly the Achilles heel of this approach. In both simulations the standard deviation of $\hat{\sigma}_b^2$ is equal to about half its mean. This is likely to erode the advantage of the adjustment schemes over no adjustment, but probably not to the extent that no adjustment would be preferable.

The scheme tAdj has a distinct advantage over uAdj in that it does not depend on (an estimate of) σ_b^2. Analytical discussion of the scheme AAdj is not feasible, but on a few examples the adjustments for tAdj and AAdj are highly correlated, especially when the estimated severity variance is small.

Considerable improvement in estimation of σ_b^2 may be achieved by pooling

information across multiple administrations of the same test or across similar forms of the same test when similar selection procedures, instruction, and training of the readers are conducted. Estimation of σ_b^2 could also be improved by embedding additional ratings in the assignment design, as proposed by Braun (1988), although this is of lesser importance since estimates of readers' severity coefficients are not required for true score estimation.

The simulations can also throw light on the efficiency of the sample correlation $\hat{r}$ as an estimator of the between-reader correlation $r_1 = (1 + \tau_b + \tau_e)^{-1}$. For the simulations based on the allocation design for essay B of the Advanced Placement Biology test the simulated value of r_1 is equal to 1/(1+1.85/3.74) = 0.669. The mean of the simulated estimates $(1 + \hat{\tau}_b + \hat{\tau}_e)^{-1}$ for the normal scores is 0.665 and the mean of the sample correlations is 0.664. The sampling standard deviations of these estimators are 0.044 and 0.045. For the rounded scores the correlations are only marginally reduced (to around 0.65), and the two estimators are almost equally efficient; in fact they are nearly identical. The simulated estimates of the pairs of correlations are plotted in Figure 7 for both normal and rounded scores. The sample correlation of the pairs of simulated estimates of the correlations is 0.998 for both sets of scores. Thus, the sample correlation is a good estimator of r_1.

The results of the simulations for essay ELL1 in the CAPA test lead to similar conclusions. Details are omitted to conserve space.

Summary

A method for decomposition of the variance of essay ratings was presented. It identified sources of variation due to examinees and readers (severity and inconsistency). Extensions of the proposed model take account of changes in severity and inconsistency of the readers across the sessions or due to extraneous factors, enable relating severity and inconsistency across items (multivariate models), and allow for unequal numbers of ratings of the essays. The gains in

efficiency of the adjusted scores over the unadjusted scores are modest, though perceptible, especially for administrations in which readers have large workloads.

Information about the variance components, and severity variation in particular, are important for adjustment of scores. Explicit equations were given which enable detailed discussion and near-optimal choice of adjustment of the scores. Standard errors for the estimated variances as well as for the correlations (reliabilities) can be obtained by simulations. Simulations can also be instrumental in deciding on the assignment design. In particular, use of readers with small workloads should be avoided. In the studied examples reader inconsistency dominates variation in reader severity, and therefore the MSE's of scores can be reduced by adjustment only moderately.

The computational procedures were implemented on a Sun/Unix workstation using the Splus3.0 software and the program codes developed can be obtained from the author upon request. In practice, most of the computation can be carried out interactively, with exception of the simulations. For instance, the simulations for Advanced Placement Biology and the CAPA test took about five and ten minutes of elapsed time, respectively.

References

Braun, H.I. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics* **13**, 1–18.

Landis, J.R. and Koch, G.G. (1980). The measurement of observer agreement for categorical data. *Biometrics* **33**, 159–174.

Linacre, J.M. (1988) *FACETS*. MESA Press, Chicago.

Lunz, M.E., Wright, B.D., and Linacre, J.M. (1990). Measuring the impact of judge severity on examination of scores. *Applied Measurement in Education* **3**, 331–345.

Morris, C.N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the Statistical Association* **78**, 47–80.

Shavelson, R.J. and Webb, N.M. (1991). *Generalizability Theory.* Sage Publications, Inc., Newbury Park, CA.

Tanner, M.A. and Young, M.A. (1985). Modeling agreement among raters. *Journal of the American Statistical Association* **80**, 175–180.

Uebersax, J.S. (1993). Statistical modeling of expert ratings on medical treatment appropriateness. *Journal of the American Statistical Association* **88**, 421–427.

Tables

Table 1: Estimates of the variance components and estimated mean squared errors for Advanced Placement Biology test.
The acronyms 'NAdj', 'uAdj', 'tAdj', 'AAdj' stand for the mean squared errors for unadjusted scores, and scores adjusted using (8), (9), and (10), respectively.

Advanced Placement Biology							
	Variances			Mean squared errors			
Item	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_e^2$	NAdj	uAdj	tAdj	AAdj
A	8.20	0.32	2.42	1.372	1.321–1.324	1.218–1.221	1.145–1.147
B	3.74	0.45	1.40	0.926	0.768–0.865	0.817–0.873	0.718–0.723
C	7.07	–0.01	0.78	0.392	0.369–0.386	0.367–0.377	0.361–0.367
D	4.32	0.72	1.72	1.222	1.127–1.128	1.081–1.081	0.924–0.924

Table 2: Estimates of the variance components for Advanced Placement Studio Art Portfolio Assessment.
The layout and notation are the same as in Table 1.

Advanced Placement Studio Art							
	Variances			Mean squared errors			
Criterion	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_e^2$	NAdj	uAdj	tAdj	AAdj
A	0.310	0.016	0.199	0.072	0.066–0.067	0.060–0.064	0.055–0.058
B	0.448	0.033	0.220	0.127	0.108–0.111	0.105–0.108	0.089–0.097
C	0.384	0.034	0.222	0.138	0.112–0.116	0.113–0.120	0.087–0.100
D	0.320	0.047	0.264	0.155	0.124–0.133	0.117–0.137	0.094–0.104
E	0.304	0.053	0.276	0.164	0.128–0.139	0.121–0.143	0.095–0.106
F	0.359	0.053	0.279	0.166	0.133–0.141	0.127–0.147	0.101–0.113

Table 3: Within-reader diagnostics for Advanced Placement Studio Art Portfolio Assessment.
For sessions 1 and 2 the standardized versions of the statistics $S_{R,jk}$ are given. The column 'Both sessions' contains the standardized versions (theoretical expectation equal to unity) of these statistics pooled across the sessions. Each statistic is accompanied by the workload on which it is based. Statistics mentioned in the text are printed in bold.

Reader diagnostics — Studio Art Portfolio						
	Session 1		Session 2		Both sessions	
Reader	χ^2/df	load	χ^2/df	load	χ^2/df	load
110	0.80	152	0.76	174	0.77	326
111	0.65	183	0.77	197	0.71	380
112	**1.85**	151	1.15	141	1.51	292
113	1.05	154	0.85	122	0.96	276
114	1.09	164	1.15	153	1.13	317
115	0.92	232	0.81	301	0.86	533
116	0.91	162	0.89	134	0.90	296
117	0.77	126	**0.53**	86	0.67	212
118	0.94	170	1.25	153	1.09	323
119	0.86	175	1.03	189	0.97	364
120	0.85	214	0.92	166	0.88	380
122	1.22	97	0.74	46	1.11	143
123	1.42	128	1.11	124	1.28	252
124	1.39	131	1.11	197	1.27	328
125	**1.86**	123	1.43	114	1.68	237
126	0.94	207	0.79	224	0.86	431
127	0.84	140	1.72	96	1.30	236
128	1.24	170	1.05	183	1.17	353
129	0.80	156	0.88	214	0.85	370
180	0.85	195	0.74	177	0.80	372
181	1.03	167	0.83	119	0.94	286
182	1.04	232	1.08	205	1.05	437
183	1.03	102	1.07	90	1.06	192
190	**1.82**	18	1.11	143	1.19	161

Table 4: Estimates of the variance components for CAPA tests.
The layout and notation are the same as in Table 1.

	Variances			Mean squared errors			
Item	$\hat{\sigma}_a^2$	$\hat{\sigma}_b^2$	$\hat{\sigma}_c^2$	NAdj	uAdj	tAdj	AAdj
ELL1	0.730	0.062	0.371	0.217	0.190–0.195	0.182–0.188	0.152–0.161
ELL2	0.762	0.056	0.240	0.148	0.124–0.127	0.132–0.133	0.108–0.112
SSc1	0.707	0.077	0.321	0.199	0.166–0.173	0.170–0.171	0.136–0.143
SSc2	1.358	0.064	0.264	0.169	0.142–0.144	0.152–0.153	0.130–0.132

Table 5: Summary of the simulations using the assignment design for essay B in Advanced Placement Biology test.
Two hundred simulations were generated. The scores were generated according to the model in (5), with variances equal to their estimates from the real dataset. The right-most column summarizes the numbers of scores that were smaller than zero or greater than 10 (out of a total of $2 \times 297 = 594$ scores).

	200 simulations, Essay B in AP Biology						
	Normal scores			Rounded scores			
	σ_a^2	σ_b^2	σ_e^2	σ_a^2	σ_b^2	σ_e^2	Trunc.
True value	3.740	0.450	1.400	3.740	0.450	1.400	n/a
Mean	3.756	0.455	1.407	3.440	0.418	1.408	26.370
St. dev.-n	0.421	0.261	0.194	0.371	0.241	0.188	6.788

Table 6: Summary of the simulations using the assignment design for essay ELL1 in the CAPA test.
Two hundred simulations were generated. The scores were generated according to the model in (5), with variances equal to their estimates from the real dataset. The right-most column summarizes the numbers of scores that were smaller than one or greater than 6 (out of a total of 2×419 = 838 scores).

	200 simulations, Essay ELL1 in the CAPA test.						
	Normal scores			Rounded scores			
	σ_a^2	σ_b^2	σ_e^2	σ_a^2	σ_b^2	σ_e^2	Trunc.
True value	0.730	0.062	0.371	0.730	0.062	0.371	n/a
Mean	0.749	0.077	0.350	0.722	0.074	0.421	19.955
St. dev.-n	0.067	0.038	0.040	0.065	0.038	0.041	5.075

Figures

1. Relationship of the correlations of two scores given to the same essay by different readers
2. The mean squared error as a function of the shrinkage coefficient u_i using the adjustment scheme based on (8). Advanced Placement Biology test.
3. The mean squared error as a function of the shrinkage coefficient t_i using the adjustment scheme based on (9). Advanced Placement Biology test.
4. Histograms of the score adjustments for essay ELL1. The CAPA test.
5. Plots of the adjusted scores for ELL1 against the scores from the multiple-choice part of the CAPA test.
6. Histograms of the simulated estimates of the variances using the assignment design from the ELL1 essay in the CAPA test.
7. Plots of the estimates of the correlation r_1 based on the variance estimates, and the sample correlation $\hat{r}$.

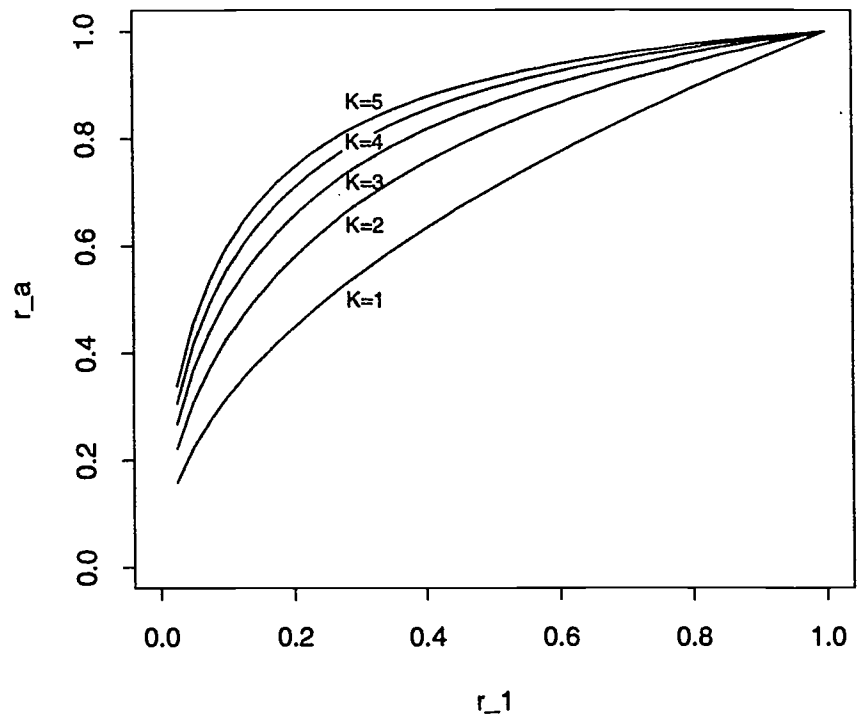

Figure 1: Relationship of the correlations of two scores given to the same essay by different readers, r_1, on the horizontal axis, and the correlation of the mean score (K readers) with the true score, r_a, on the vertical axis.
The number of scores contributing to the mean, K, is marked in the plot.

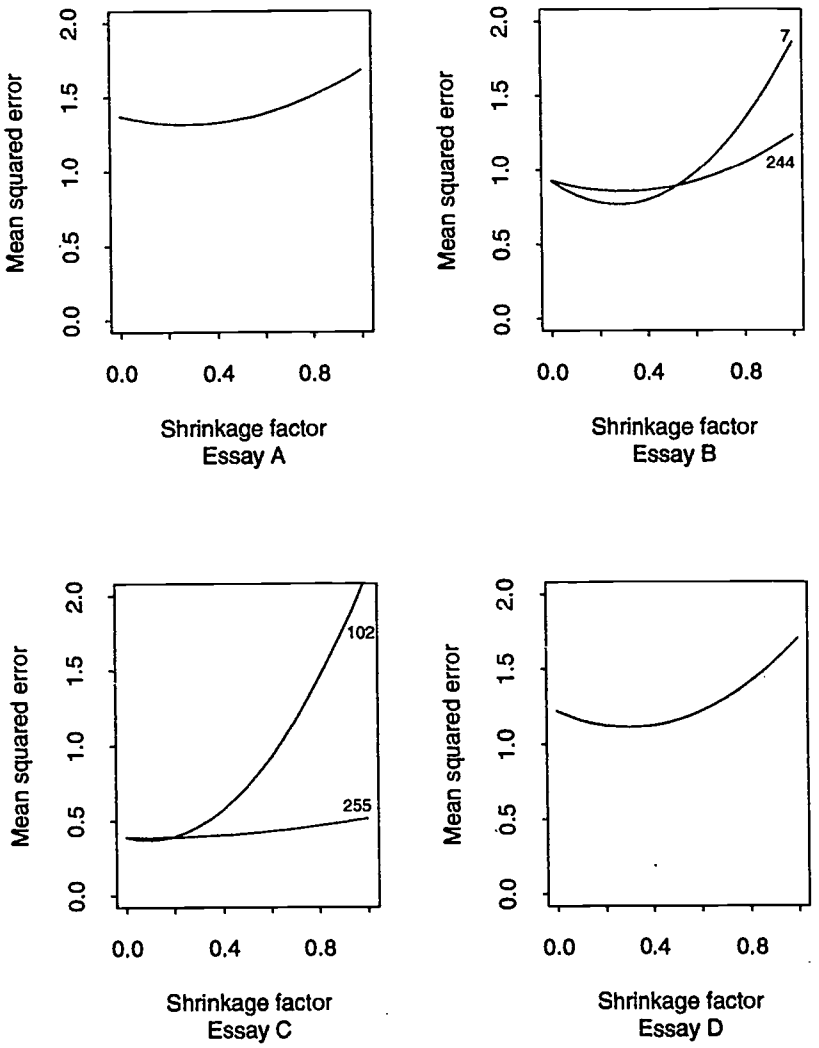

Figure 2: The mean squared error as a function of the shrinkage coefficient u_i using the adjustment scheme based on (8). Advanced Placement Biology test. Each essay is represented by a plot of the function for one or two responses. Response 7 for essay B and response 102 for essay C were rated by one reader each with small workload. The functions for responses with the usual workload almost coincide.

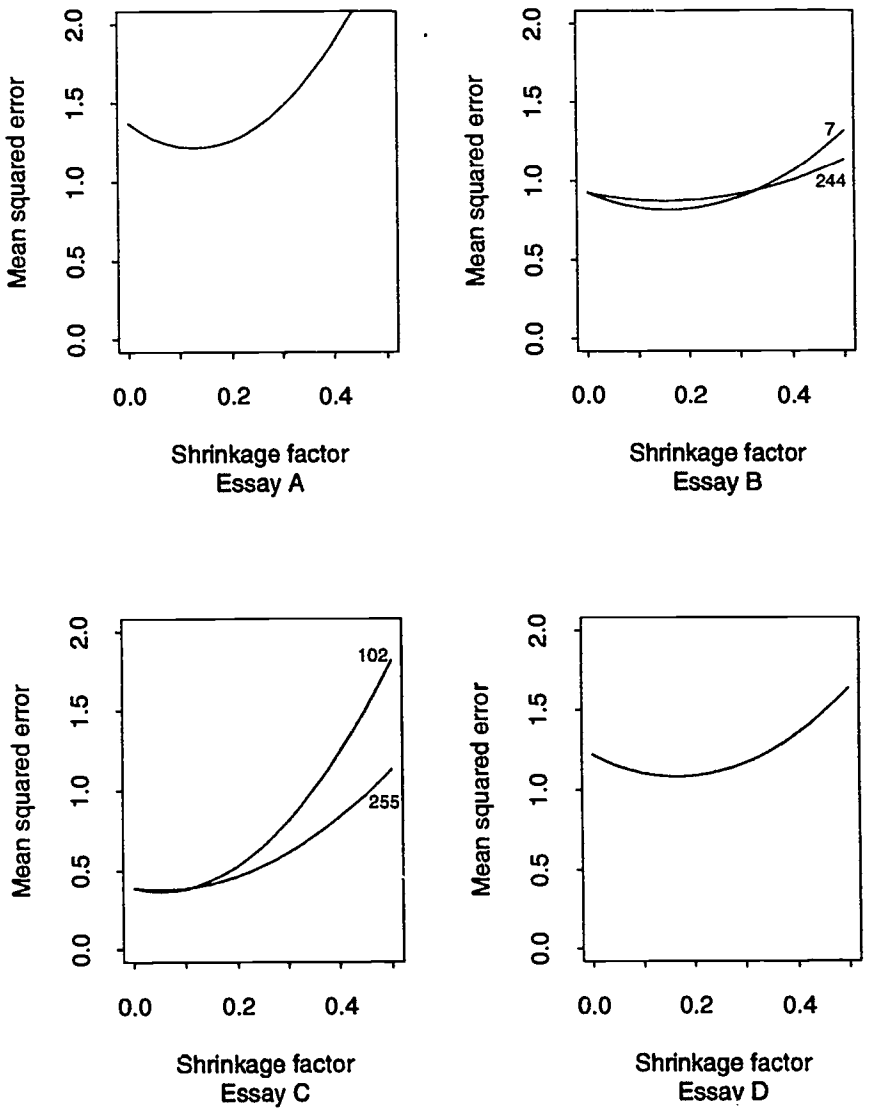

Figure 3: The mean squared error as a function of the shrinkage coefficient t_i using the adjustment scheme based on (9). Advanced Placement Biology test. The same layout as in Figure 2 is used, and the functions for the same responses are plotted.

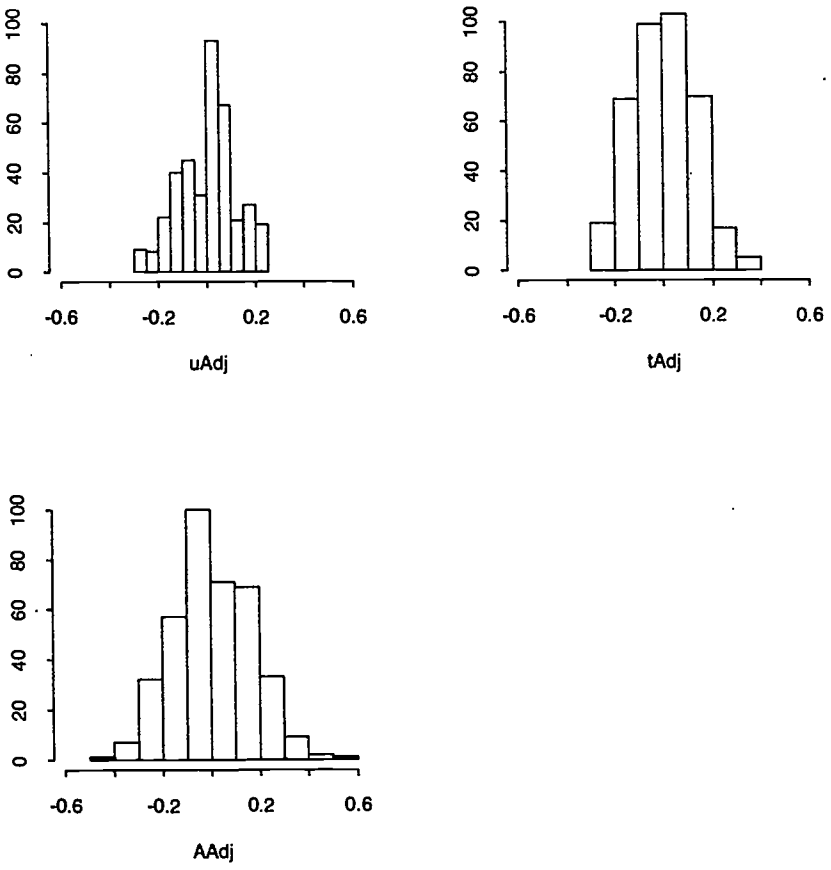

Figure 4: Histograms of the score adjustments for essay ELL1. The CAPA test.

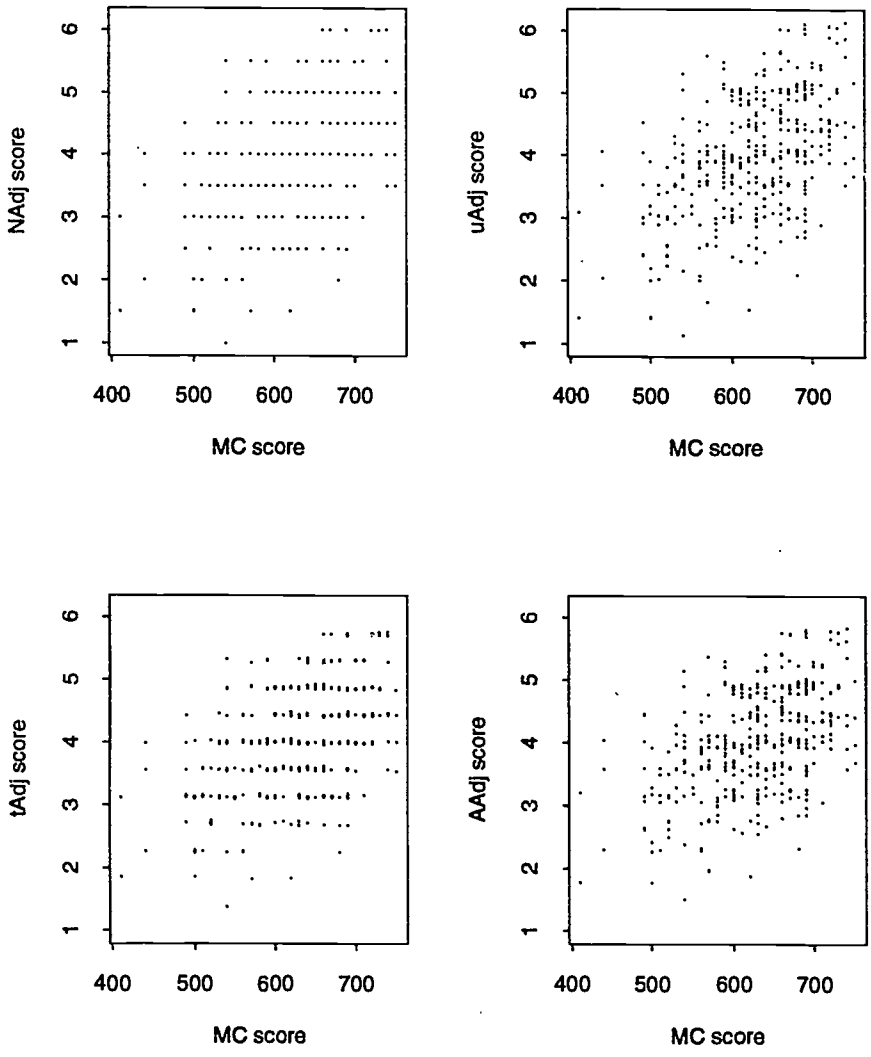

Figure 5: Plots of the unadjusted and adjusted scores for ELL1 against the scores from the multiple-choice part of the CAPA test.

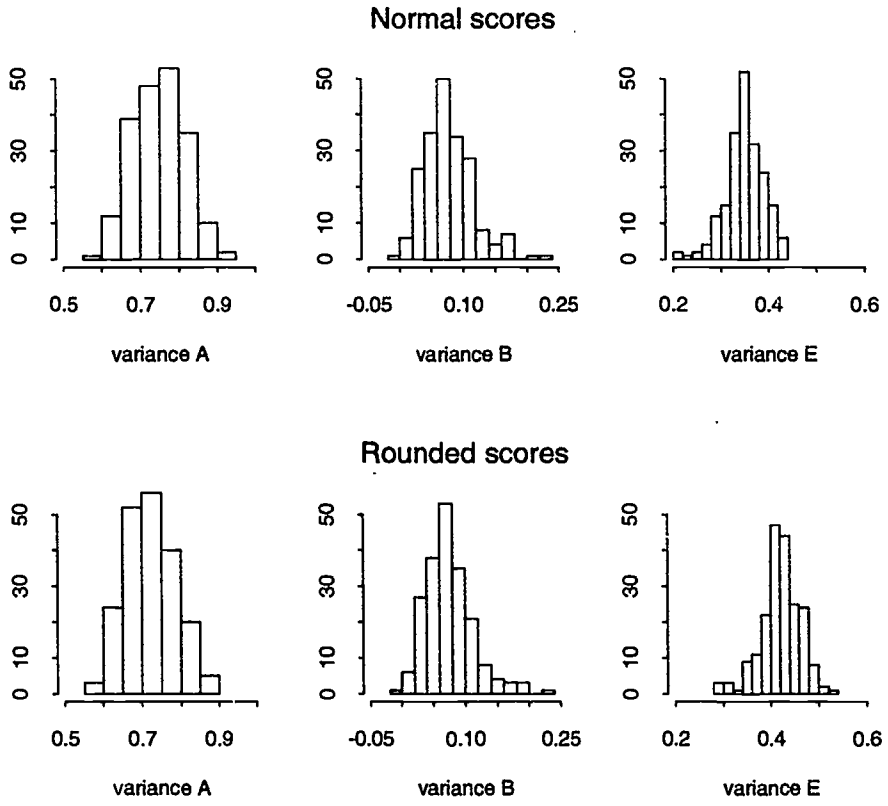

Figure 6: Histograms of the simulated estimates of the variances using the assignment design from the ELL1 essay in the CAPA test.
The first row refers to the normal scores, the second row to the rounded scores, the symbols (variances) 'A', 'B', and 'E' stand for σ_a^2, σ_b^2, and σ_e^2, respectively.

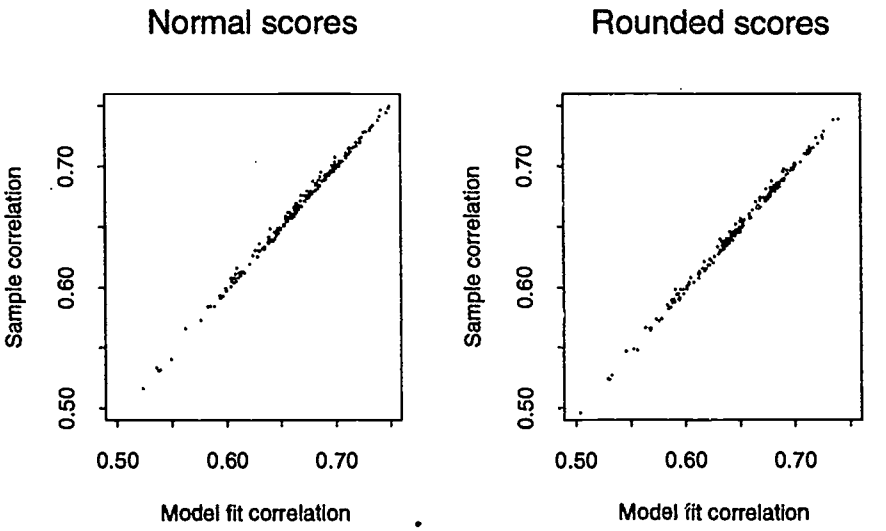

Figure 7: Plots of the estimates of the correlation r_1 based on the variance estimates (horizontal axis), and the sample correlation $\hat{r}$ (vertical axis). Simulated data based on the allocation design for essay B of the Advanced Placement Biology test.