British Journal of Mathematical and Statistical Psychology (2015), 68, 363–385

The British Psychological Society

www.wileyonlinelibrary.com

Prediction of true test scores from observed item scores and ancillary data

Shelby J. Haberman*, Lili Yao and Sandip Sinharay
ETS, Princeton, New Jersey, USA

In many educational tests which involve constructed responses, a traditional test score is obtained by adding together item scores obtained through holistic scoring by trained human raters. For example, this practice was used until 2008 in the case of GRE® General Analytical Writing and until 2009 in the case of TOEFL® iBT Writing. With use of natural language processing, it is possible to obtain additional information concerning item responses from computer programs such as e-rater®. In addition, available information relevant to examinee performance may include scores on related tests. We suggest application of standard results from classical test theory to the available data to obtain best linear predictors of true traditional test scores. In performing such analysis, we require estimation of variances and covariances of measurement errors, a task which can be quite difficult in the case of tests with limited numbers of items and with multiple measurements per item. As a consequence, a new estimation method is suggested based on samples of examinees who have taken an assessment more than once. Such samples are typically not random samples of the general population of examinees, so that we apply statistical adjustment methods to obtain the needed estimated variances and covariances of measurement errors. To examine practical implications of the suggested methods of analysis, applications are made to GRE General Analytical Writing and TOEFL iBT Writing. Results obtained indicate that substantial improvements are possible both in terms of reliability of scoring and in terms of assessment reliability.

1. Introduction

An increasingly common issue in educational testing is the efficient evaluation of a skill by use of multiple sources of information. A good example involves writing assessments in which two or more essays are scored by trained human raters. With use of natural language processing, additional information concerning item responses by computer programs such as electronic essay scorers can be achieved. Furthermore, it is easy to obtain such related information as the computer-generated features associated with the e-rater® program (Attali & Burstein, 2006; Burstein, Chodorow, & Leacock, 2004; Davey, 2009). Examples of computer-generated variables are number of discourse elements, average number of words per discourse element, and the square root of the number of grammatical errors detected per word. In addition to this information

*Correspondence should be addressed to Shelby J. Haberman, ETS, Princeton, NJ 08541, USA (email: shaberman@ets.org).
Sandip Sinharay has moved since this study was undertaken and is now based at CTB/McGraw-Hill, Monterey, CA 93940, USA.

DOI:10.1111/bmsp.12052

concerning writing proficiency, additional information concerning verbal ability is provided by other sections of a test.

The analysis suggested in this paper includes several methods that do not appear to be used simultaneously in the psychometric literature. Some of the methodology required is related to the augmentation literature (Haberman, 2008; Wainer, Sheehan, & Wang, 2000; Wainer et al., 2001); however, previous literature does not consider the correlated errors which can arise with essays scored by both humans and computers. Some methodology is also related to attempts to weight human and computer-derived essay scores (Haberman, 2011; Haberman & Qian, 2007); however, existing analyses either do not consider multiple constructed responses or do not consider covariates such as other test scores. Repeater analysis without adjustment for non-representative sampling has been considered for reliability estimation (Siegert & Guo, 2009; Zhang, 2008); however, adjustment methods based on minimum discriminant information (Haberman, 1984) have been employed in our paper to compensate for non-representative sampling.

To illustrate the issues involved, it is helpful to consider two examples, the TOEFL® iBT Writing test and the GRE® General Analytical Writing test. TOEFL is an assessment of English proficiency typically taken by examinees not from English-speaking countries who wish to study at an undergraduate, graduate, or professional level in the United States or Canada; however, the test is also employed for immigration purposes in some countries and for some forms of professional licensure. The four sections of the test are Listening, Reading, Speaking, and Writing. The Writing test includes two constructed-response items, an Integrated task and an Independent task, each of which requires that the examinee write an essay responsive to a prompt. The prompt for the Integrated task includes both oral and written content, so that the task to some extent tests the examinee's ability to both listen and read complementary information. In the Independent task an opinion must be supported in writing. Until 2009, each examinee response to each prompt was normally scored by at least two trained human raters drawn from a large pool of raters, although some exceptions existed which involved unsuitable essays or human raters who differed substantially in their scores. Each rater used a scoring rubric to provide an integer holistic score from 1 to 5 to the essay, with 5 the best score, although a score of 0 could be given for such cases as an essay not responsive to the prompt or a blank essay. Since that time, a transition has taken place in which a typical essay response is scored by one human rater and by e-rater (Attali & Burstein, 2006; Burstein et al., 2004; Davey, 2009), a computer program which uses natural language processing to provide a collection of real variables which describe the essay response. The e-rater score is a continuous value truncated to have a range from 0.5 to 5.5. In practice, some cases exist in which e-rater is not used due to unusual properties of the essay, and special rules leading to additional human raters are employed in cases where human and computer ratings are deemed excessively large. Use of e-rater began in 2009 in the case of the Independent task and in 2011 in the case of the Integrated task. In this paper, analysis relies on the e-rater features rather than on the e-rater score. The problem is to score the Writing test given the human scores, the computer-generated variables, and scores from the other sections of the test.

GRE General is a test typically used for admission to graduate schools in the United States and Canada. The test has undergone a recent revision, the revised GRE (rGRE), but the data in this paper are from the previous version. In both the earlier and current version of the test, the assessment has three parts, Quantitative Reasoning, Verbal Reasoning, and Analytical Writing. Analytical Writing consists of two items, an Issue essay and an

Argument essay. In the Issue task, the examinee is asked to discuss and express his/her perspective on a topic of general interest. In the Argument task, a brief passage is presented in which the author makes a case for some course of action or interpretation of events by presenting claims backed by reasons and evidence.

Since 2008, typical essays have been scored with one human rater and with essay features obtained from e-rater, although the e-rater score is not explicitly used in the scoring. Instead the e-rater score provides a confirmation of the human score, and additional human scores are only sought if unusual essay properties prevent use of e-rater or if the difference between the truncated e-rater score and the human score exceeds 0.5. In this paper, the computer-generated features associated with e-rater are employed rather than e-rater itself. Once again, the problem is to use the human ratings, the computer-generated variables, and the other section information to assess proficiency in analytical writing. The resulting approaches studied are different than those used by GRE.

In Section 2, the general methodology required for analysis is developed. In Section 2.1, the problem of estimation of a composite true score based on various linear predictors is described. Sections 2.2 and 2.3 provide the estimation of best linear predictors and corresponding measures of mean square error and proportional reduction in mean square error involves use of repeater data when item data are unavailable. The adjustment method is minimum discriminant information adjustment (MDIA), which is described in Sections 2.4 and in the Appendix. Its application to repeater data is discussed in Section 2.5. Applications of the methods of Sections 2.4 and 2.5 are then provided in Section 3. Conclusions are provided in Section 4.

2. Method

2.1. The general model

Both the TOEFL and GRE examples are special cases of a general problem that can be described in terms of classical test theory but differs somewhat from the traditional case. An examination results in an observed K-dimensional ($K \geq 1$) bounded random vector $\mathbf{X}$ with elements X_k, $1 \leq k \leq K$. For example, one might have $K = 2$, where X_1 equals the holistic score on the first prompt and X_2 equals the holistic score on the second prompt. The vector $\mathbf{X}$ has mean $E(\mathbf{X})$ and positive-definite covariance matrix $\mathrm{Cov}(\mathbf{X})$. The vector $\mathbf{X}$ may be decomposed into an unobserved true score component $\boldsymbol{\tau}$ with elements τ_k, $1 \leq k \leq K$, and an unobserved error component $\mathbf{e}$ with elements e_k, $1 \leq k \leq K$, so that $\mathbf{X} = \boldsymbol{\tau} + \mathbf{e}$. It is assumed that the error vector $\mathbf{e}$ has expectation $\mathbf{0}_K$, where $\mathbf{0}_K$ denotes the K-dimensional variable with all elements 0, and that $\mathbf{e}$ has a finite covariance matrix $\mathrm{Cov}(\mathbf{e})$. It is further assumed that the true score $\boldsymbol{\tau}$ and the error $\mathbf{e}$ are uncorrelated, so that τ_k and $e_{k'}$ are uncorrelated for $1 \leq k \leq K$ and $1 \leq k' \leq K$.

This paper considers the problem of estimation of a composite true score. For some K-dimensional vector $\mathbf{c}$ with elements c_k, $1 \leq k \leq K$, the composite true score $v = \mathbf{c}'\boldsymbol{\tau} = \sum_{k=1}^{K} c_k \tau_k$ is to be approximated by use of the observed vector $\mathbf{X}$. In linear prediction, v is approximated by a linear function $a + \mathbf{b}'\mathbf{X}$ of the observed vector $\mathbf{X}$, where a is a real constant and $\mathbf{b}$ is a K-dimensional constant vector. The mean square error of the prediction is

$$S(a, \mathbf{b}) = E([v - a - \mathbf{b}'\mathbf{X}]^2). \tag{1}$$

with best linear prediction, a real constant α and a K-dimensional vector constant $\boldsymbol{\beta}$ with elements β_k, $1 \leq k \leq K$, are selected so that

$$\mathrm{MSE} = S(\alpha, \boldsymbol{\beta}) \leq S(a, \mathbf{b}) \tag{2}$$

for all real a and K-dimensional vector constants $\mathbf{b}$. Unlike in classical test theory (Lord & Novick, 1968), the error variables e_k, $1 \leq k \leq K$, are not necessarily assumed to be uncorrelated. This change in assumptions requires an unconventional analysis in many typical cases.

To illustrate the issue involved, consider a case of an assessment with two prompts, each of which is scored by one human rater who provides a single holistic score and by a computer which generates a vector of eight feature variables. Let there also be two available scores from other related parts of the assessment. Then one might have $K = 20$, where X_1 equals the holistic score on the first prompt, X_2 equals the holistic score on the second prompt, $X_3,\ldots,X_{10}$ are the feature variables for the first prompt, $X_{11},\ldots,X_{18}$ are the feature variables for the second prompt, and X_{19} and X_{20} are scores on other parts of the assessment. If $\mathbf{c}$ is the 20-dimensional vector with elements $c_1 = c_2 = 1$ and $c_k = 0$ for $3 \leq k \leq K = 20$, then $\mathbf{c}'\boldsymbol{\tau}$ is the true score $\tau_1 + \tau_2$ for the sum $X_1 + X_2$ of the two holistic scores. Two variations on models for true scores and errors can be considered, depending on the desired inference. In the simplest case, the accuracy of human scoring is the focus. Different raters may provide different scores on the same examinee response to the same essay; however, for the electronic score raters, there are no scoring errors involved in this case and thus the variances $\mathrm{Var}(e_k)$ of the errors e_k are 0 for $3 \leq k \leq 18$. There are also no scoring errors for scores on other parts of the assessment; that is, the variances $\mathrm{Var}(e_k)$ of the errors e_k are 0 for $19 \leq k \leq K$. Thus the variance $\mathrm{Var}(e_k) = 0$ for $3 \leq k \leq K$. Because $[\mathrm{Cov}(e_k, e_{k'})]^2 \leq \mathrm{Var}(e_k)\mathrm{Var}(e_{k'})$, it follows that $\mathrm{Cov}(e_k, e_{k'}) = 0$ if either $3 \leq k \leq K$ or $3 \leq k' \leq K$. Presumably the variances $\mathrm{Var}(e_1)$ and $\mathrm{Var}(e_2)$ associated with scoring errors are positive; however, the raters can be assumed to be drawn independently for the two prompts, so that $\mathrm{Cov}(e_1,e_2) = 0$.

In a related but different case, the prompts themselves are regarded as drawn from pools of comparable prompts, so that the true score $\tau_1 + \tau_2$ can be regarded as the expected sum of human holistic scores among parallel tests. Here the focus is on the accuracy of the assessment. In this case, the feature scores also vary depending on the exact prompt to which the examinee responds. Thus the variances $\mathrm{Var}(e_k)$ can be assumed to be positive for $1 \leq k \leq K$. Within the same prompt, the essay feature variables may be correlated with one another, that is, the covariances $\mathrm{Cov}(e_k, e_{k'})$ may be non-zero if $3 \leq k < k' \leq 10$ or $11 \leq k < k' \leq 18$. Similarly, for human scores and essay features, $\mathrm{Cov}(e_1,e_k)$ may be non-zero for $3 \leq k \leq 10$ and $\mathrm{Cov}(e_2,e_k)$ may be non-zero for $11 \leq k' \leq 18$. Nonetheless, errors for different prompts are uncorrelated, so that $\mathrm{Cov}(e_k, e_{k'})$ is 0 if either k is 1 or $3 \leq k \leq 10$ and either k' is 2 or $11 \leq k' \leq 18$. The errors of the scores on other parts of the assessment are also uncorrelated with the errors for human raters or computer essay features. Therefore, $\mathrm{Cov}(e_k, e_{k'})$ is also 0 for $1 \leq k \leq 18$ and $19 \leq k' \leq K$.

2.2. *The best linear predictor*

Given the expectation $E(\mathbf{X}) = E(\boldsymbol{\tau})$ of the observed vector $\mathbf{X}$, the covariance matrix $\mathrm{Cov}(\boldsymbol{\tau})$ of the vector $\boldsymbol{\tau}$ of true scores, and the covariance matrix $\mathrm{Cov}(\mathbf{e})$ of the error vector $\mathbf{e}$, α and $\boldsymbol{\beta}$ can be thus found from (1) and (2). The expectation of v is

$$E(v) = \mathbf{c}'E(\mathbf{X}), \tag{3}$$

the covariance matrix of $\mathbf{X}$ satisfies

$$\text{Cov}(\mathbf{X}) = \text{Cov}(\boldsymbol{\tau}) + \text{Cov}(\mathbf{e}), \tag{4}$$

and the covariance vector

$$\text{Cov}(v, \mathbf{X}) = E([v - E(v)][\mathbf{X} - E(\mathbf{X})]) \tag{5}$$

is $\text{Cov}(\boldsymbol{\tau})\mathbf{c}$. The variance $\sigma^2(v)$ of v is $\mathbf{c}'\text{Cov}(\boldsymbol{\tau})\mathbf{c}$. Thus

$$\boldsymbol{\beta} = [\text{Cov}(\mathbf{X})]^{-1}\text{Cov}(\boldsymbol{\tau})\mathbf{c} \tag{6}$$

and

$$\alpha = (\mathbf{c} - \boldsymbol{\beta})'E(\mathbf{X}) = \mathbf{c}'[\text{Cov}(\mathbf{X})]^{-1}\text{Cov}(\mathbf{e})E(\mathbf{X}) \tag{7}$$

(Rao, 1973, p. 266). In addition, the mean square error is

$$\begin{aligned} \text{MSE} &= \sigma^2(v) - [\text{Cov}(v, \boldsymbol{\tau})]'[\text{Cov}(\mathbf{X})]^{-1}\text{Cov}(v, \mathbf{X}) \\ &= \mathbf{c}'\{\text{Cov}(\boldsymbol{\tau}) - \text{Cov}(\boldsymbol{\tau})[\text{Cov}(\mathbf{X})]^{-1}\text{Cov}(\boldsymbol{\tau})\}\mathbf{c}, \end{aligned} \tag{8}$$

and the coefficient of determination is

$$\rho^2 = 1 - \frac{\text{MSE}}{\sigma^2(v)} = \frac{\mathbf{c}'\text{Cov}(\boldsymbol{\tau})[\text{Cov}(\mathbf{X})]^{-1}\text{Cov}(\boldsymbol{\tau})\mathbf{c}}{\mathbf{c}'\text{Cov}(\boldsymbol{\tau})\mathbf{c}}. \tag{9}$$

In the literature on augmented scoring, ρ^2 is the proportional reduction in mean square error (PRMSE) from prediction of v by the linear predictor $\alpha + \boldsymbol{\beta}'\mathbf{X}$ relative to prediction of v by the constant $E(v)$ (Haberman, 2008).

In many cases, it is helpful to examine the effects of restrictions on the best linear predictor to examine which elements X_k of $\mathbf{X}$ are particularly important or unimportant in prediction of the composite true score v. For this purpose, one may consider an H by K matrix $\mathbf{M}$ of rank H for an integer $H \geq 1$. The predictor $\mathbf{MX}$ is then considered. For example, if $\mathbf{M}$ is the H by K matrix with elements M_{hk} equal to 1 for $h = k$ and to 0 for $h \neq k$, then $\mathbf{MX}$ is the vector with elements X_k, $1 \leq k \leq H$. The best linear predictor of v based on $\mathbf{MX}$ is then $\alpha_\mathbf{M} + \boldsymbol{\beta}_\mathbf{M}'\mathbf{MX}$, where

$$\boldsymbol{\beta}_\mathbf{M} = [\text{Cov}(\mathbf{MX})]^{-1}\mathbf{M}'\text{Cov}(\boldsymbol{\tau})\mathbf{c} \tag{10}$$

and

$$\alpha_\mathbf{M} = (\mathbf{c} - \mathbf{M}'\boldsymbol{\beta}_\mathbf{M})'E(\mathbf{X}). \tag{11}$$

In addition, the mean square error is

$$\begin{aligned} \text{MSE}_\mathbf{M} =& \sigma^2(v) - [\text{Cov}(v, \mathbf{M}\boldsymbol{\tau})]'[\text{Cov}(\mathbf{MX})]^{-1}\text{Cov}(v, \mathbf{M}\boldsymbol{\tau}) \\ =& \mathbf{c}'\{\text{Cov}(\boldsymbol{\tau}) - \text{Cov}(\boldsymbol{\tau})\mathbf{M}[\text{Cov}(\mathbf{MX})]^{-1}\mathbf{M}'\text{Cov}(\boldsymbol{\tau})\}\mathbf{c}, \end{aligned} \tag{12}$$

and the coefficient of determination is

$$\rho^2_{\mathbf{M}} = 1 - \frac{\mathrm{MSE}_{\mathbf{M}}}{\sigma^2(v)} = \frac{\mathbf{c}'\mathrm{Cov}(\tau)\mathbf{M}[\mathrm{Cov}(\mathbf{MX})]^{-1}\mathbf{M}'\mathrm{Cov}(\tau)\mathbf{c}}{\mathbf{c}'\mathrm{Cov}(\tau)\mathbf{c}}. \tag{13}$$

Note that for $\mathbf{M}$ equal to the K by K identity matrix, $\alpha_{\mathbf{M}} = \alpha$, $\boldsymbol{\beta}_{\mathbf{M}} = \boldsymbol{\beta}$, $\mathrm{MSE}_{\mathbf{M}} = \mathrm{MSE}$, and $\rho^2_{\mathbf{M}} = \rho^2$.

In some cases, it is useful to consider the extent to which a change in $\mathbf{c}$ can improve $\rho^2_{\mathbf{M}}$. For this purpose, consider a G by K matrix $\mathbf{L}$ with elements L_{gk}, $1 \le g \le G$, $G \ge 1$, $1 \le k \le K$. Let $\mathbf{L}$ have rank G. One may then consider $\mathbf{c} = \mathbf{Ld}$ for G-dimensional vectors $\mathbf{d}$. The object is to maximize the ratio

$$\frac{\mathbf{d}'\mathbf{L}'\mathrm{Cov}(\tau)\mathbf{M}[\mathrm{Cov}\ (\mathbf{MX})]^{-1}\mathbf{M}'\mathrm{Cov}(\tau)\mathbf{Ld}}{\mathbf{d}'\mathbf{L}'\mathrm{Cov}(\tau)\mathbf{Ld}}$$

over non-zero vectors $\mathbf{d}$. The maximum may be denoted by $\rho_{\mathbf{LM}}$, and a corresponding value $\mathbf{d}$ such that the sum of the elements of $\mathbf{d}$ is 1 and the first non-zero element of $\mathbf{d}$ is positive may be denoted by $\mathbf{d}_{\mathbf{LM}}$. This maximization corresponds to the standard problem of computation of a maximum relative eigenvalue (Rao, 1973; p. 74). The maximum coefficient of determination $\rho^2_{\mathbf{LM}}$ is the largest real number λ such that

$$\mathbf{L}'\mathrm{Cov}(\tau)\mathbf{M}[\mathrm{Cov}(\mathbf{MX})]^{-1}\mathbf{M}'\mathrm{Cov}(\tau)\mathbf{L} - \lambda\mathbf{L}'\mathrm{Cov}(\tau)\mathbf{L}$$

is singular, and $\mathbf{d}_{\mathbf{LM}}$ satisfies the relative eigenvector equation

$$\mathbf{L}'\mathrm{Cov}(\tau)\mathbf{M}[\mathrm{Cov}(\mathbf{MX})]^{-1}\mathbf{M}'\mathrm{Cov}(\tau)\mathbf{Ld}_{\mathbf{LM}} = \rho^2_{\mathbf{LM}}\mathbf{L}'\mathrm{Cov}(\tau)\mathbf{Ld}_{\mathbf{LM}}. \tag{14}$$

For example, if $\mathbf{c}$ is only of interest if $c_k = 0$ for $k > 2$, then $\mathbf{L}$ can be chosen to be the $2 \times K$ matrix with elements L_{gk} equal to 0 for $g \ne k$ and to 1 for $g = k$, $1 \le g \le 2$, $1 \le k \le K$.

2.3. Estimation of expectations and covariance matrices

In practice, the challenge in application of (6–9) is estimation from available data. To avoid singularity of the estimated covariance matrix, let the sample size n exceed the dimension K of $\mathbf{X}$. Consider K-dimensional independent and identically distributed observations $\mathbf{X}_i$, $1 \le i \le n$, with the same distribution as $\mathbf{X}$. Then estimate the expectation $E(\mathbf{X})$ by the sample mean

$$\bar{\mathbf{X}} = n^{-1}\sum_{i=1}^{n}\mathbf{X}_i \tag{15}$$

and the covariance matrix $\mathrm{Cov}(\mathbf{X})$ by the sample covariance matrix[1]

$$\overline{\mathrm{Cov}}(\mathbf{X}) = n^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'. \tag{16}$$

[1] Given the very large sample sizes in analysis and given the later use of weights, division by n in (16) appears to be appropriate.

However, more information is typically required to estimate the covariance matrix $\text{Cov}(\boldsymbol{\tau})$ for the vector $\boldsymbol{\tau}$ of true scores. In common applications of score augmentation (Haberman, 2008; Wainer et al., 2001), the elements X_k, $1 \leq k \leq K$, are separate test scores, and the errors e_k, $1 \leq k \leq K$, are uncorrelated. Item-level data are sufficient to estimate the variances $\sigma^2(e_k)$ by use of Cronbach's α. Thus $\text{Cov}(\mathbf{e})$ can be estimated. The desired estimation of $\text{Cov}(\boldsymbol{\tau})$ then follows from the estimates of $\text{Cov}(\mathbf{X})$ and $\text{Cov}(\mathbf{e})$. This assumption of uncorrelated errors does not typically apply to the cases under study in this paper. In addition, use of Cronbach's α in the problems discussed in this paper may be affected by use of single items to measure constructs that are distinct. For example, in TOEFL iBT, the Integrated task requires the ability to discuss specific material obtained from both written and oral material, whereas the Independent task requires the ability to handle a somewhat simpler prompt with less specific expectations for the response.

In the case of scoring accuracy in a test with $U \leq K$ constructed responses X_k, $1 \leq k \leq U$, each of which is typically scored by a single rater, a sample of essays may be used in which more than one rater is used to score the same essay. For prompt $k \leq U$, let $I(k)$ be a randomly selected subset with $r_k > 0$ members obtained from the integers $1,\ldots,n$. For i in $I(k)$, let X'_{ik} be an additional score for examinee i on prompt k obtained by a different rater than the rater for the score X_{ik}. Then the rater variability measure

$$\bar{V}(e_k) = (2r_k)^{-1} \sum_{i \in I(k)} (X_{ik} - X'_{ik})^2 \tag{17}$$

provides an estimate of $V(e_k)$. Under the assumption that $\text{Cov}(e_k, e_{k'})$ is 0 for $k \neq k'$ and $V(e_k) = 0$ for any $k > U$, best linear prediction becomes feasible. Let $\overline{\text{Cov}}(\mathbf{e})$ be the diagonal matrix with kth diagonal element $\bar{V}(e_k)$ for $1 \leq k \leq U$ and other diagonal elements 0. Then $\text{Cov}(\boldsymbol{\tau})$ has estimate

$$\overline{\text{Cov}}(\boldsymbol{\tau}) = \overline{\text{Cov}}(\mathbf{X}) - \overline{\text{Cov}}(\mathbf{e}). \tag{18}$$

Estimates for α, $\boldsymbol{\beta}$, MSE, ρ^2, $\alpha_{\mathbf{M}}$, $\boldsymbol{\beta}_{\mathbf{M}}$, $\text{MSE}_{\mathbf{M}}$, $\rho^2_{\mathbf{M}}$, $\rho^2_{\mathbf{LM}}$, and $\mathbf{d}_{\mathbf{LM}}$ are then obtained by substitution of $\bar{\mathbf{X}}$ for $E(\mathbf{X})$, $\overline{\text{Cov}}(\mathbf{X})$ for $\text{Cov}(\mathbf{X})$, $\overline{\text{Cov}}(\boldsymbol{\tau})$ for $\text{Cov}(\boldsymbol{\tau})$, and $\overline{\text{Cov}}(\mathbf{MX}) = \mathbf{M}'\overline{\text{Cov}}(\mathbf{X})\mathbf{M}$ for $\text{Cov}(\mathbf{MX})$.

This case has been considered with $U = 1$ and $K > 1$ (Haberman & Qian, 2007). Here scoring of an item is considered rather than scoring of a full test composed of constructed responses. In addition, a case of prediction of the sum of human item scores on a test consisting of two constructed-response items has been considered in the case of TOEFL (Haberman, 2011). In this example, for each item, available data include a human score and a single estimated human score generated by e-rater.

In typical cases involving the criterion of assessment accuracy of a test consisting of a quite limited number of constructed responses and a number of computer-generated features, it is necessary to employ repeater data to estimate the covariance matrix. One such approach was employed with TOEFL data (Haberman, 2011); however, the methodology does not readily generalize to more general use of computer-generated features. It was applied to two human scores and two computer-generated essay scores. The interest in this paper is in examining the case in which, for each item, both a human score and a number of computer-generated features are employed.

2.4. Adjustment by minimum discriminant information

In general, use of repeater data derived from operational testing programs involves the challenge of samples that are not necessarily representative of the population of examinees who take the test under study. Typically, examinees need not repeat an assessment, and they are likely to do so only if they have both reason to expect that they might achieve a higher score by repeating and reason to believe that a higher score will be of significant benefit. As a consequence, it is likely that repeaters are less proficient on average than are examinees randomly selected from the general population. In this paper, adjustment by minimum discriminant information (Haberman, 1984) is employed to weight a sample of repeater examinees to make the characteristics of the weighted sample more similar to the characteristics of a general sample of examinees.

Adjustment by minimum discriminant information (Haberman, 1984) obtains a weighted sample which satisfies a finite set of constraints on weighted averages. Let $\mathbf{Y}$ and $\mathbf{Z}$ be two random vectors of dimension m, and let u be a Bernoulli random variable such that $u = 1$ with positive probability. Let the conditional covariance matrix of $\mathbf{Y}$ given $u = 1$ be finite and positive definite, and let $\mathbf{Z}$ have a finite covariance matrix. The problem is to minimize the conditional discriminant information $E(w \log (w)|u = 1) = E(w \log (w/u)|u = 1)$ for comparison of the random weight variables w and u subject to the constraints that $w > 0$ if $u = 1$, $w = 0$ if $u = 0$,

$$E(w|u = 1) = 1, \tag{19}$$

and

$$E(w\mathbf{Y}|u = 1) = E(\mathbf{Z}) \tag{20}$$

(Csiszár, 1975; Haberman, 1984; Kullback & Leibler, 1951). A random variable w that minimizes the discriminant information subject to the given constraints is a MDIA weight. If any random variable w exists such that (19) and (20) hold, $w > 0$ if $u = 1$, and $w = 0$ if $u = 0$, then a unique real $c > 0$ and a unique m-dimensional vector $\boldsymbol{\beta}$ exist such that

$$w = cu \exp(\boldsymbol{\beta}'\mathbf{Y}) \tag{21}$$

is an MDIA weight. Conversely, if (19–20) hold for any $c > 0$ and m-dimensional $\boldsymbol{\beta}$, then w is an MDIA weight. If w' is also an MDI weight, then $w = w'$ with probability 1. In typical applications, one considers a real function g on the set R^m of m-dimensional vectors such that $wg(\mathbf{Y})$ has a finite conditional expectation $E(wg(\mathbf{Y})|u = 1)$. When the conditional expectation of $wg(\mathbf{Y})$ given $u = 1$ and the unconditional expectation of $g(\mathbf{Y})$ are both finite, then $E(wg(\mathbf{Y}|u = 1)$ is used to approximate $E(g(\mathbf{Y}))$. This approximation is of practical interest when $\mathbf{Y}$ is not fully observed if $u = 0$, $\mathbf{Y}$ is observed if $u = 1$, $\mathbf{Z}$ is always observed, and it is known that $\mathbf{Z}$ and $\mathbf{Y}$ have the same expectation. By (20), if g is a linear function, then

$$E(wg(\mathbf{Y})|u = 1) = E(g(\mathbf{Y})). \tag{22}$$

If the selection variable u and the random vector $\mathbf{Y}$ are independent, then (22) holds for any real function g on R^m such that $g(\mathbf{Y})$ has a finite expectation, for $c = 1$, $\boldsymbol{\beta}$ equals the m-

dimensional vector $\mathbf{0}_m$ with all elements 0, and $w = 1$. A more general case may be based on the conditional probability $P(u = 1|\mathbf{Y})$ that $u = 1$ given $\mathbf{Y}$. Assume that $P(u = 1|\mathbf{Y})$ is positive. Let

$$v = \frac{uP(u = 1)}{P(u = 1|\mathbf{Y})}. \tag{23}$$

If $vg(\mathbf{Y})$ has a finite conditional expectation given $u = 1$, then $g(\mathbf{Y})$ has a finite expectation and

$$E(vg(\mathbf{Y})|u = 1) = \frac{E(vg(\mathbf{Y}))}{P(u = 1)} = \frac{E(vg(\mathbf{Y})P(u = 1|\mathbf{Y}))}{P(u = 1)} = E(g(\mathbf{Y})) \tag{24}$$

(Rao, 1973; pp. 96–97). Conversely, if $g(\mathbf{Y})$ has a finite expectation, then $vg(\mathbf{Y})$ has a finite conditional expectation given $u = 1$ and (24) holds. In particular, $E(v\mathbf{Y}|u = 1) = E(\mathbf{Y}) = E(\mathbf{Z})$. If $\log P(u = 1|\mathbf{Y})$ is a linear function $a - \boldsymbol{\beta}'\mathbf{Y}$ of $\mathbf{Y}$ for some unknown real a and m-dimensional vector $\boldsymbol{\beta}$ and if $c = P(u = 1) \exp(-a)$, then $w = v$, which satisfies (19–21), is an MDIA weight.

Examples of use of adjustments by minimum discriminant information are relatively common in the literature, as evident from the references in Haberman (1984). A few simple cases are reviewed in the Appendix.

To estimate the weight w, consider a simple random sample of size n with independent and identically distributed observations $(\mathbf{Y}_i,\mathbf{Z}_i,u_i)$, $1 \leq i \leq n$, with the same distribution as $(\mathbf{Y},\mathbf{Z},u)$. Let R be the set of i with $u_i = 1$, and let n_R be the number of elements of R. If any positive real numbers w_i, $i \in R$, exist such that

$$n_R^{-1} \sum_{i \in R} w_i = 1 \tag{25}$$

and

$$n_R^{-1} \sum_{i \in R} w_i \mathbf{Y}_i = \bar{\mathbf{Z}} = n^{-1} \sum_{i=1}^{n} \mathbf{Z}_i, \tag{26}$$

then the sample discriminant information $-n_R^{-1} \sum_{i \in R} w_i \log(w_i)$ is minimized subject to (25) and (26) by unique positive MDIA sample weights w_i which satisfy

$$w_i = u_i \hat{c} \exp(\hat{\boldsymbol{\beta}}'\mathbf{Y}_i), \tag{27}$$

for some real $\hat{c} > 0$ and m-dimensional vector $\hat{\boldsymbol{\beta}}$. If the sample covariance matrix of the $\mathbf{Y}_i$, $i \in R$, is positive definite, then $\hat{c}$ and $\hat{\boldsymbol{\beta}}$ are uniquely determined (Haberman, 1984).

MDIA sample weights have both consistency and asymptotic normality properties. Let the conditional expectation $E(w \exp(\boldsymbol{\gamma}'\mathbf{Y})|u = 1)$ be finite for all m-dimensional vectors $\boldsymbol{\gamma}$ sufficiently close to $\mathbf{0}_m$. This condition always holds if the random vector $\mathbf{Y}$ is bounded. As the sample size n increases, the probability approaches 1 that the estimates $\hat{c}$ and $\hat{\boldsymbol{\beta}}$ exist and are uniquely determined. In addition, $\hat{c}$ converges to c with probability 1 and $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}$ with probability 1. If $E(w|g(\mathbf{Y})| \exp(\boldsymbol{\gamma}'\mathbf{Y}))$ is finite for all $\boldsymbol{\gamma}$ sufficiently

close to $\mathbf{0}_m$, then the average $\hat{E}(wg(\mathbf{Y})|u = 1)$ of the $w_{ig}(\mathbf{Y}_i)$, $i \in R$, converges with probability 1 to the conditional expectation $E(wg(\mathbf{Y})|u = 1)$. Note that the regularity conditions hold if $\mathbf{Y}$ and $g(\mathbf{Y})$ are bounded. In addition, if the conditional covariance matrix of $w\mathbf{Y}$ given $u = 1$ is finite and if the conditional variance of $wg(\mathbf{Y})$ given $u = 1$ is finite, then $n^{1/2}[\hat{E}(wg(\mathbf{Y})|u = 1) - E(wg(\mathbf{Y})|u = 1)]$ converges in distribution to a normal random variable as n approaches ∞ (Haberman, 1984). Once again, the regularity conditions hold if $g(\mathbf{Y})$ and $\mathbf{Y}$ are bounded.

2.5. Use of repeater data

To apply repeater data, in addition to the K-dimensional random vector $\mathbf{X}$ of test results, let $\mathbf{X}_*$ be a bounded K-dimensional vector of repeat test results with elements X_{k*}, $1 \leq k \leq K$. In the ideal case, $\mathbf{X}_* = \mathbf{T} + \mathbf{e}_*$, where the error $\mathbf{e}_*$ on the repeat test has the same mean and covariance matrix as the error $\mathbf{e}$ in the original test. In addition, $\mathbf{e}_*$ is uncorrelated with both the common true score τ and the error vector $\mathbf{e}$. Let $\text{Cov}(\mathbf{X},\mathbf{X}_*)$ be the K by K matrix with row k and column k' equal to the covariance of X_k and X_{k*}, and let $\text{Cov}(\mathbf{X}_*,\mathbf{X})$ be the transpose of $\text{Cov}(\mathbf{X},\mathbf{X}_*)$. Then $\text{Cov}(\mathbf{X},\mathbf{X}_*)$ and $\text{Cov}(\mathbf{X}_*,\mathbf{X})$ are both equal to the covariance matrix $\text{Cov}(\tau)$. If $(\mathbf{X}_i,\mathbf{X}_{i*})$, $1 \leq i \leq n$, are mutually independent and have common distribution $(\mathbf{X},\mathbf{X}_*)$, then a simple estimate of $\text{Cov}(\tau)$ is

$$\overline{\text{Cov}}(\tau) = \frac{1}{2}[\overline{\text{Cov}}(\mathbf{X}, \mathbf{X}_*) + \overline{\text{Cov}}(\mathbf{X}_*, \mathbf{X})], \tag{28}$$

where

$$\overline{\text{Cov}}(\mathbf{X}, \mathbf{X}_*) = n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_{i*} - \bar{\mathbf{X}}_*)', \tag{29}$$

$\overline{\text{Cov}}(\mathbf{X}_*, \mathbf{X})$ is the transpose of $\overline{\text{Cov}}(\mathbf{X}, \mathbf{X}_*)$, and

$$\bar{\mathbf{X}}_* = n^{-1} \sum_{i=1}^{n} \mathbf{X}_{i*}. \tag{30}$$

Use of (28) is needed even though $\text{Cov}(\mathbf{X},\mathbf{X}_*) = \text{Cov}(\mathbf{X}_*,\mathbf{X})$ because the sample estimates $\overline{\text{Cov}}(\mathbf{X}, \mathbf{X}_*)$ and $\overline{\text{Cov}}(\mathbf{X}_*, \mathbf{X})$ are not necessarily equal.

Unfortunately, in typical cases involving repeaters, $\mathbf{X}_{i*}$ is not always observable, and the availability of $\mathbf{X}_{i*}$ may be related to the value of $\mathbf{X}_i$. In addition, even if available, $\mathbf{X}_{i*}$ may differ from $\mathbf{X}$ in distribution due to learning of the subject being studied or due to greater familiarity with the test. To treat the situation, consider the added Bernoulli variable u with positive probability that $u = 1$, and consider the triple $(\mathbf{X},u\mathbf{X}_*,u)$. With this arrangement, $u = 1$ corresponds to the case of $\mathbf{X}_*$ observed, and $u = 0$ corresponds to $\mathbf{X}_*$ not observed. In the Appendix, $\mathbf{X}$, $\mathbf{X}_*$, and u are used to construct bounded m-dimensional vectors $\mathbf{Y}$ and $\mathbf{Z}$, and a weight function w is defined as in Section 2.4 so that $E(w|u = 1) = 1$ and $E(w\mathbf{Y}|u = 1) = E(\mathbf{Z})$. The definition of $\mathbf{Y}$ and $\mathbf{Z}$ ensures that the conditional expectations $E(w\mathbf{X}|u = 1)$ and $E(w\mathbf{X}_*|u = 1)$ are both equal to the unconditional expectation $E(\mathbf{X})$, and the weighted conditional covariance matrices

$$\mathrm{Cov}_w(X) = E(w[\mathbf{X} - E(w\mathbf{X}|u=1)][\mathbf{X} - E(w\mathbf{X}|u=1)]'|u=1) \tag{31}$$

and

$$\mathrm{Cov}_w(\mathbf{X}_*) = E(w[\mathbf{X}_* - E(w\mathbf{X}_*|u=1)][\mathbf{X}_* - E(w\mathbf{X}_*|u=1)]'|u=1) \tag{32}$$

are equal to the unconditional covariance matrix Cov(**X**) of **X**. In addition, the weighted conditional covariance matrices

$$\mathrm{Cov}_w(\mathbf{X}, \mathbf{X}_*) = E(w[\mathbf{X} - E(w\mathbf{X}|u=1)][\mathbf{X}_* - E(w\mathbf{X}_*|u=1)]'|u=1) \tag{33}$$

and

$$\mathrm{Cov}_w(\mathbf{X}_*, \mathbf{X}) = E(w[\mathbf{X}_* - E(w\mathbf{X}_*|u=1)][\mathbf{X} - E(w\mathbf{X}|u=1)]'|u=1) \tag{34}$$

are set equal. Thus given $u = 1$, the weighted conditional means and covariances associated with responses **X** and $\mathbf{X}_*$ are compatible with the ideal case. If u is independent of **X** and $\mathbf{X}_*$ and the ideal assumptions apply to these random vectors, then $w = 1$. The equation $\mathrm{Cov}_w(\mathbf{X}_*,\mathbf{X}) = \mathrm{Cov}(\mathbf{X}_*,\mathbf{X})$ holds if the conditional probability $P(u = 1|\mathbf{X},\mathbf{X}_*)$ that $u = 1$ given **X** and $\mathbf{X}_*$ is positive and

$$\log[P(u=1|\mathbf{X},\mathbf{X}_*)] = \eta + \boldsymbol{\eta}'\mathbf{X} + \boldsymbol{\eta}_*'\mathbf{X}_* + \mathbf{X}'\mathbf{A}\mathbf{X} + \mathbf{X}_*'\mathbf{A}_*\mathbf{X}_* + \mathbf{X}'\mathbf{B}\mathbf{X}_* \tag{35}$$

for some real constant η, m-dimensional vectors $\boldsymbol{\eta}$ and $\boldsymbol{\eta}_*$, and m by m matrices **A**, $\mathbf{A}_*$, and **B**, where $\mathbf{B} = -\mathbf{B}'$.

With a random sample $(\mathbf{X}_i,u_i\mathbf{X}_{i*},u_i)$, $1 \le i \le n$, with the distribution of $(\mathbf{X},u\mathbf{X}_*,u)$, one then obtains the sample $(\mathbf{Y}_i,\underline{\mathbf{Z}_i})$, $1 \le i \le n$, as in the Appendix. The resulting weights w_i lead to estimates $\bar{\mathbf{X}}$ of $E(\mathbf{X})$, $\underline{\mathrm{Cov}}(\mathbf{X})$ of Cov(**X**), and

$$\widehat{\underline{\mathrm{Cov}}}(\tau) = n_R^{-1}\sum_{i\in R} w_i[\mathbf{X}_i - \bar{\mathbf{X}}][\mathbf{X}_{i*} - \bar{\mathbf{X}}]' \tag{36}$$

of Cov(τ), where R denotes the set of i with $u_i = 1$.

Estimates for α, $\boldsymbol{\beta}$, MSE, ρ^2, $\alpha_{\mathbf{M}}$, $\boldsymbol{\beta}_{\mathbf{M}}$, $\mathrm{MSE}_{\mathbf{M}}$, $\rho^2_{\mathbf{M}}$, $\rho^2_{\mathbf{LM}}$, and $\mathbf{d}_{\mathbf{LM}}$ are then obtained by substitution of $\bar{\mathbf{X}}$ for $E(\mathbf{X})$, $\underline{\mathrm{Cov}}(\mathbf{X})$ for Cov(**X**), $\widehat{\underline{\mathrm{Cov}}}(\tau)$ for Cov(τ), and $\underline{\mathrm{Cov}}(\mathbf{MX})$ for Cov(**MX**).

3. Applications

3.1. TOEFL iBT: One representative sample

Here data include essay scores, feature scores, and scores on the TOEFL iBT Listening, Reading, and Speaking sections of 161,648 examinees in seven administrations of TOEFL iBT in 2010. The vector **X** contains $K = 13$ elements. The first element is the human holistic score for the Integrated prompt, the second element is the human holistic score for the Independent prompt, the third to tenth elements are eight features scores for the Independent prompt, and the last three elements are the scaled section scores for Reading, Listening, and Speaking. Variances due to rater error are estimable due to samples of essays in which two raters rate the same essay, and the use of two raters is not

related to any characteristics of the examinees. Examinees were removed from the sample if one of their essays did not conform to program requirements for use of e-rater. For example, an examinee was excluded if one of the essays contained fewer than 25 words or if an initial human rating was 0. In addition, examinees were excluded if no Listening or Speaking score was available. An outlier screen was performed by regression of an observed element of $\mathbf{X}$ on the remaining elements of $\mathbf{X}$. An observation was excluded if any of the resulting 13 studentized deleted residuals exceeded 4 in absolute value. After exclusions, there remained 158,163 examinees. Because no repeater data were present, analysis focuses on measurement of scoring accuracy. For each prompt, a sample of essay responses scored by two raters was available in which membership in the sample was not related to examinee performance on essays or other portions of the test.

The following features of an essay were used (Attali & Burstein, 2006; Burstein et al., 2004; Haberman & Sinharay, 2010):

1. Development: The logarithm of the average number of words per discourse element;
2. Organization: The logarithm of the number of discourse elements;
3. Grammar: Minus the square root of the number of grammatical errors detected per word;
4. Mechanics: Minus the square root of the number of mechanics errors detected per word;
5. Usage: Minus the square root of the number of usage errors detected per word;
6. Style: Minus the square root of the number of style errors detected per word;
7. Vocabulary level: Minus the median Standard Frequency Index of words in the essay for which the index can be evaluated (nwfmedian);
8. Word length: The average word length of words in the essay.

Note that a discourse element might be an introduction, a conclusion, a main point, or a supporting argument. The Standard Frequency Index is a measure of general relative word frequency in a large corpus of text.

Three basic vectors $\mathbf{c}$ were examined. In all cases, $c_k = 0$ for $k > 2$, c_1 and c_2 are positive, and $c_1 + c_2 = 1$, so that a weighted average of the true essay scores is estimated. The first choice used an arithmetic mean of human scores, so that $c_1 = c_2 = 0.5$. The second choice used

$$c_u = \frac{1/\hat{\sigma}(X_u)}{1/\hat{\sigma}(X_1) + 1/\hat{\sigma}(X_2)}$$

for $u = 1$ or 2, so that human rater scores are standardized by their observed standard deviations $\sigma(X_1)$ and $\sigma(X_2)$ for the two essay prompts. The third choice has

$$c_u = \frac{1/\hat{\sigma}(\tau_u)}{1/\hat{\sigma}(\tau_1) + 1/\hat{\sigma}(\tau_2)},$$

so that standardization of scores is based on the standard deviations of the true scores τ_1 and τ_2. These choices of c_u have been available in the literature for a considerable period of time (Gulliksen, 1950; ch. 20). As evident from Rudner (2001), among others, many other choices are possible. Nonetheless, it is quite often the case that the practical effect of different selections of c_1 and c_2 is limited (Wilks, 1938). For these data, the second value of c_1 was 0.42 and the third was 0.39. These two values are less than 0.5 because the estimated standard deviation of observed scores on the Integrated prompt is somewhat

larger than the corresponding estimated standard deviation of observed scores on the Independent prompt, and the estimated standard deviation of true scores on the Integrated prompt is somewhat larger than the corresponding estimated standard deviation of true scores on the Independent prompt.

In addition, $\hat{\rho}_{\mathbf{LM}}$ and $\mathbf{d}_{\mathbf{LM}}$ were found for the 2 by $K = 13$ matrix $\mathbf{L}$ with $L_{11} = L_{22} = 1$ and all other elements equal to 0, so that $\rho^2_{\mathbf{M}}$ was maximized over $\mathbf{c}$ such that $c_k = 0$ for $k > 2$, $c_1 + c_2 = 1$, and $c_1 > 0$ or $c_1 = 0$ and $c_2 > 0$. The e-rater predictor for the independent prompt is the estimated best linear predictor of X_2 based on $X_3,\ldots,X_{10}$. The estimated regression coefficients are in Table 1. This e-rater predictor is not exactly the same as the e-rater score actually used in the assessment; however, the sample correlation of .993 is very high. In addition, the sample size is sufficiently large that the e-rater score does not differ appreciably from the actually best linear predictor of X_2 by use of $X_3,\ldots,X_{10}$.

Table 2 shows the estimates of $\hat{\rho}^2_{\mathbf{M}}$ for the four values of the weight c_1 for the Integrated prompt for the following predictors, each of which corresponds to a matrix $\mathbf{M}$ with 13 columns:

- Model 1: The observed human scores X_1 and X_2;
- Model 2: The human scores X_1 and X_2 and the computer-generated essay features $X_3,\ldots,X_{10}$ for the independent prompt;
- Model 3: All 13 predictors;
- Model 4: The human scores X_1 and X_2 and the e-rater predictor for the independent prompt;
- Model 5: The human scores X_1 and X_2, the e-rater predictor for the independent prompt, and the scaled scores for the Listening, Reading, and Speaking sections.

For the five models, the estimated optimal values of c_1 are 0.67, 0.35, 0.32, 0.35, and 0.33, respectively.

Table 1. Regression slopes for e-rater scores for the Independent prompt: TOEFL iBT with one representative sample

Variable	Slope
Development	1.20
Organization	1.26
Grammar	3.48
Mechanics	2.18
Style	0.34
Usage	3.84
Vocabulary level	0.03
Word length	0.19

Table 2. Estimates $\hat{\rho}^2_{\mathbf{M}}$ for the five models: TOEFL iBT with one representative sample

Models	$c_1 = 0.5$	$c_1 = 0.42$	$c_1 = 0.39$	Optimal c_1
Model 1	.85	.84	.83	.85
Model 2	.89	.89	.89	.89
Model 3	.92	.92	.92	.93
Model 4	.89	.89	.90	.90
Model 5	.92	.92	.92	.92

For each model, the precise weighting selected by use of c_1 and c_2 has very little effect. Comparison of Model 2 to Model 1 shows that use of computer-generated features provides an appreciable increase in PRMSE. Comparison of Model 3 to Model 2 indicates that a further slightly smaller increase is provided by use of the other section scores, a rather striking result given that the computer-generated features are specific to one of the essays being rated and the other section scores have no direct connection to either essay. Use of the e-rater regression score rather than the eight individual features has little effect, for Model 4 leads to virtually the same results as Model 2 and Model 5 leads to virtually the same result as Model 3.

Tables 3 and 4 show the estimated regression coefficients $\hat{\beta}_k$ and the standardized partial regression coefficients $\hat{\sigma}(X_k)\hat{\beta}_k/\hat{\sigma}(v)$ for the five models for the case with $c_1 = 0.5$. Note that the coefficients for human scores are much higher for the Integrated prompt than for the Independent prompt and they differ appreciably for different models. For example, for Model 2, the weight on the human rating on the integrated prompt is 0.42 while that on the human rating on the independent prompt is 0.18. This difference is partly related to use of e-rater only for the Independent prompt; however, it is notable that even when only two human raters are used for prediction, the weights for human raters are 0.48 for the Integrated prompt and 0.34 for the Independent prompt.

The regression coefficients for each human rating are quite similar for Models 2 and 4 and for Models 3 and 5, a further indication that, in the case of scoring accuracy, use of the e-rater score for the Independent prompt is about as effective as use of all features for the Independent prompt. In addition, coefficients for a given human rating vary substantially from model to model. They are smallest in Model 3 and Model 5 due to use of other essay section scores in the prediction. It is worth noting that each of the Reading, Listening, and Speaking sections has a larger standardized regression coefficient in Model 3 than any computer-generated features other than organization and development. These results are particularly striking because scoring accuracy is involved. Nonetheless, in Model 5, the standardized regression weight for e-rater scores is a little larger than the individual weights for Reading, Listening, and Speaking scores. The computer-generated features

Table 3. Regression coefficients for the five models: TOEFL iBT with one representative sample

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Human rating: Integrated	0.47	0.42	0.33	0.42	0.33
Human rating: Independent	0.34	0.17	0.11	0.18	0.12
Development	–	0.48	0.46	–	–
Organization	–	0.51	0.49	–	–
Grammar	–	1.46	0.97	–	–
Mechanics	–	0.93	0.83	–	–
Style	–	0.16	−0.02	–	–
Usage	–	1.56	0.96	–	–
Vocabulary level	–	0.01	0.01	–	–
Word length	–	0.09	0.12	–	–
Listening	–	–	0.01	–	0.01
Reading	–	–	0.01	–	0.01
Speaking	–	–	0.02	–	0.02
e-rater	–	–	–	0.41	0.33

Note. Features and e-rater scores are for the Independent prompt.

Table 4. Standardized partial regression coefficients for the five models: TOEFL iBT with one representative sample

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Human rating: Integrated	0.69	0.61	0.48	0.60	0.48
Human rating: Independent	0.35	0.18	0.11	0.19	0.12
Development	–	0.21	0.20	–	–
Organization	–	0.22	0.21	–	–
Grammar	–	0.08	0.05	–	–
Mechanics	–	0.08	0.07	–	–
Style	–	0.02	−0.00	–	–
Usage	–	0.08	0.05	–	–
Vocabulary level	–	0.06	0.03	–	–
Word length	–	0.04	0.06	–	–
Listening	–	–	0.11	–	0.11
Reading	–	–	0.08	–	0.08
Speaking	–	–	0.13	–	0.12
e-rater	–	–	–	0.30	0.24

Note. Features and e-rater scores are for the Independent prompt.

and human scores are for the essays actually being scored, whereas the other section scores are not directly dependent on the examinee essays.

3.2. TOEFL iBT: One representative sample and one repeater sample

This analysis was based on human essay scores, computer-generated feature scores, and scores on the other TOEFL iBT sections of 49,469 examinees who took TOEFL iBT between July 2009 and June 2010 and also took the GRE exam during a similar period. It is important to emphasize that these results are more informative for the portion of examinees who take TOEFL iBT in order to apply for graduate study than for the population of TOEFL examinees. For each examinee, two human scores were available for the Integrated prompt and one human score was available for the Independent prompt. Because the data preceded regular use of e-rater for the Integrated prompt, the eight feature scores described in Section 3.1 were only available for the Independent prompt. Examinees were excluded if the essay for the Independent prompt contained fewer than 25 words, if any score from the human raters was 0, or if the Listening, Reading, or Speaking section score was not available. For these data, it was possible to identify repeaters, so that accuracy of assessment is the focus.

For these data, $K = 13$, X_1 is the average of the two human ratings for the Integrated prompt, X_2 is the human rating for the Independent prompt, and $X_3, \ldots, X_{13}$ are defined as in Section 3.1. An outlier screen performed as in Section 3.1 reduced the sample size to 49,035.

Models 1–5 were defined as in Section 3.1; however, it should be noted that X_1 is now an average human score based on two raters rather than a human score from one rater. Repeaters were only used if on the repeat examination, the same conditions on essay scores and Listening, Reading, and Speaking section scores were satisfied as were required for inclusion of the initial examination. In addition, a further outlier screen was performed in which observations were removed if absolute values of studentized deleted residuals exceeded 4 in any linear regression of X_{k^*} in the repeat administration on the 13 elements

of $\mathbf{X}$ in the first administration and on the remaining 12 elements of $\mathbf{X}_*$. In all, 10,142 repeaters were used. The e-rater regression was computed from the initial Independent essays as in Section 3.1. The vectors $\mathbf{c}$ were again selected so that $c_k = 0$ for $k > 2$ and $c_1 + c_2 = 1$. The selections of c_1 were found as in Section 3.1, except that variances of true scores were based on assessment accuracy rather than on scoring accuracy. Results are summarized in Tables 5–8. Optimal c_1 values for the five models are 0.49, 0.08, 0.49, 0.12, and 0.51, respectively.

Table 5 shows that there is a noticeable gain when features of the Independent prompt are added in the model (that is, the PRMSE $\hat{\rho}^2_{\mathbf{M}}$ of Model 2 is larger than the corresponding measure of Model 1). Note that the PRMSE of Model 3 is substantially larger than that of Model 2 when the other section scores are added. The values of the PRMSEs of Model 4 or 5 are nearly as large as those of Model 2 or 3, respectively, so that e-rater scores are quite competitive with a more general use of feature scores. The values of the PRMSE in Table 5 are considerably smaller than those in Table 2 because assessment reliability is measured rather than scoring accuracy.

Tables 7 and 8 show the regression coefficients and the standardized partial regression coefficients for the five models for the case with $c_1 = 0.5$. Compared to the study of scoring accuracy in TOEFL, the relative weight of the Independent prompt is increased for each model and the relative weight of the Integrated prompt is reduced. When e-rater is used in a model, it receives increased weight. When used, other section scores receive sharply increased weight. When all eight e-rater features are used, organization and development receive much less weight and other features receive more weight. Rather strikingly, in Model 5, the two lowest standardized regression coefficient are associated with the human ratings.

Table 5. Estimates $\hat{\rho}^2_M$ for the five models: TOEFL iBT with one representative sample and a repeater sample

Models	$c_1 = 0.5$	$c_1 = 0.44$	$c_1 = 0.43$	Optimal c_1
Model 1	.69	.69	.69	.69
Model 2	.76	.77	.77	.78
Model 3	.88	.88	.88	.88
Model 4	.75	.76	.76	.77
Model 5	.88	.88	.88	.88

Table 6. Estimated regression coefficients for e-rater score developed from Independent prompt: TOEFL iBT with one representative sample and a repeater sample

Variable	Coefficient
Development	1.20
Organization	1.26
Grammar	4.00
Mechanics	2.68
Style	0.30
Usage	3.71
Vocabulary level	0.05
Word length	0.19

Table 7. Regression coefficients for the five models: TOEFL iBT with one representative sample and a repeater sample

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Average human rating: Integrated	0.34	0.28	0.10	0.28	0.10
Human rating: Independent	0.35	0.19	0.08	0.19	0.08
Development	–	0.25	0.25	–	–
Organization	–	0.27	0.28	–	–
Grammar	–	2.32	1.42	–	–
Mechanics	–	1.30	1.15	–	–
Style	–	0.31	0.10	–	–
Usage	–	1.98	1.12	–	–
Vocabulary level	–	0.03	0.01	–	–
Word length	–	0.05	0.12	–	–
Reading	–	–	0.03	–	0.03
Listening	–	–	0.02	–	0.02
Speaking	–	–	0.04	–	0.03
e-rater	–	–	–	0.43	0.30

Note. Features and e-rater scores are for the Independent prompt.

Table 8. Standardized partial regression coefficients for the five models: TOEFL iBT with one representative sample and a repeater sample

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Average human rating: Integrated	0.52	0.43	0.16	0.43	0.16
Human rating: Independent	0.43	0.23	0.10	0.23	0.10
Development	–	0.12	0.11	–	–
Organization	–	0.12	0.12	–	–
Grammar	–	0.12	0.07	–	–
Mechanics	–	0.12	0.10	–	–
Style	–	0.05	0.02	–	–
Usage	–	0.11	0.06	–	–
Vocabulary level	–	0.11	0.04	–	–
Word length	–	0.02	0.05	–	–
Reading	–	–	0.27	–	0.27
Listening	–	–	0.17	–	0.17
Speaking	–	–	0.22	–	0.21
e-rater	–	–	–	0.36	0.25

Note. Features and e-rater scores are for the Independent prompt.

3.3. GRE Analytical Writing: One representative sample and one repeater sample

Here the available data included human essay scores, computer-generated feature scores, and scores on the Verbal and Quantitative sections of the GRE General test for about 58,000 examinees who took the test in 2009 or 2010 and also took the TOEFL iBT test during a similar period. Thus the data are informative for examinees interested in graduate study who are not from English-speaking countries rather than for the general population of GRE examinees. There were $U = 2$ prompts, an Issue prompt and an Argument prompt. For each prompt, a human holistic score and the eight essay features were available. The essay features were the same as in the TOEFL cases. Thus $K = 20$, X_1 was the human

holistic score for the Issue prompt, X_2 was the human holistic score for the Argument prompt, $X_3,\ldots,X_{10}$ were e-rater features for the Issue prompt, $X_{11},\ldots,X_{18}$ were e-rater features for the Argument prompt, X_{19} was the GRE Quantitative scaled score, and X_{20} was the GRE Verbal scaled score. As in Section 3.2, our focus is on the accuracy of scoring. After exclusions, 52,611 examinees were used for the initial administration. Of these examinees, 7,359 were used for the initial two administrations. Once again c_1 and c_2 had sum 1 and $c_k = 0$ for $k > 2$. The choices of c_1 were made as in Section 3.2. The same rules for excluding essays and examinees were applied as were applied in Section 3.2. The choice $c_1 = 0.52$ was based on estimated standard deviations of X_1 and X_2. The choice $c_1 = 0.53$ was based on estimated standard deviations of τ_1 and τ_2. The following models were used:

- Model 1: The observed human scores X_1 and X_2;
- Model 2: The human scores X_1 and X_2 and the computer-generated essay features $X_3,\ldots,X_{18}$ for the two prompts;
- Model 3: All 20 predictors;
- Model 4: The human scores X_1 and X_2 and the e-rater predictors;
- Model 5: The human scores X_1 and X_2, the e-rater predictors, and the scaled scores X_{19} and X_{20}.

Table 9 shows the values of the PRMSE statistics for the five models. The e-rater regression coefficients are in Table 10. Optimal c_1 values for the five models are 0.49, 0.51, 0.44, 0.50, and 0.44, respectively.

Table 9 shows that there is a substantial gain when essay features are employed individually or via e-rater, and the gain from other section scores is much smaller than in Section 3.2. The values of the PRMSE in Table 9 are roughly comparable to those in Table 5 for all models that do not employ section scores. The gain from other section

Table 9. Estimates $\hat{\rho}^2_{\mathbf{M}}$ for the five models: GRE Analytical Writing

Models	$c_1 = 0.5$	$c_1 = 0.52$	$c_1 = 0.41$	Optimal c_1
Model 1	.69	.69	.69	.69
Model 2	.80	.80	.80	.80
Model 3	.83	.83	.83	.83
Model 4	.80	.80	.80	.80
Model 5	.83	.83	.83	.83

Table 10. Regression coefficients for e-rater scores developed from two prompts: GRE Analytical Writing

Variable	Issue	Argument
Development	0.53	0.60
Organization	0.63	0.79
Grammar	3.27	3.82
Mechanics	2.57	2.19
Style	0.39	0.34
Usage	3.79	2.72
Vocabulary level	0.03	0.03
Word length	0.08	0.11

Table 11. Regression coefficients for the five models: GRE Analytical Writing

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Human rating: Issue	0.34	0.19	0.16	0.19	0.17
Human rating: Argument	0.35	0.20	0.17	0.20	0.17
Development: Issue	–	0.09	0.08	–	–
Organization: Issue	–	0.11	0.10	–	–
Grammar: Issue	–	1.24	1.11	–	–
Mechanics: Issue	–	0.90	0.84	–	–
Style: Issue	–	0.18	0.14	–	–
Usage: Issue	–	1.86	1.70	–	–
Vocabulary level: Issue	–	0.02	0.01	–	–
Word length: Issue	–	−0.01	−0.01	–	–
Development: Argument	–	0.13	0.10	–	–
Organization: Argument	–	0.20	0.16	–	–
Grammar: Argument	–	1.27	1.14	–	–
Mechanics: Argument	–	0.37	0.31	–	–
Style: Argument	–	0.14	0.10	–	–
Usage: Argument	–	0.63	0.54	–	–
Vocabulary level: Argument	–	0.01	0.00	–	–
Word length: Argument	–	0.02	0.00	–	–
Quantitative	–	–	0.00	–	0.00
Verbal	–	–	0.00	–	0.00
e-rater: Issue	–	–	–	0.37	0.32
e-rater: Argument	–	–	–	0.22	0.18

Table 12. Standardized partial regression coefficients for the five models: GRE Analytical Writing

Variable	Model 1	Model 2	Model 3	Model 4	Model 5
Human rating: Issue	0.45	0.25	0.22	0.25	0.22
Human rating: Argument	0.51	0.28	0.24	0.29	0.25
Development: Issue	–	0.05	0.05	–	–
Organization: Issue	–	0.06	0.06	–	–
Grammar: Issue	–	0.08	0.07	–	–
Mechanics: Issue	–	0.10	0.09	–	–
Style: Issue	–	0.03	0.03	–	–
Usage: Issue	–	0.13	0.11	–	–
Vocabulary level: Issue	–	0.08	0.06	–	–
Word length: Issue	–	−0.00	−0.01	–	–
Development: Argument	–	0.08	0.06	–	–
Organization: Argument	–	0.14	0.11	–	–
Grammar: Argument	–	0.09	0.08	–	–
Mechanics: Argument	–	0.04	0.04	–	–
Style: Argument	–	0.03	0.02	–	–
Usage: Argument	–	0.05	0.04	–	–
Vocabulary level: Argument	–	0.03	0.02	–	–
Word length: Argument	–	0.01	0.00	–	–
Quantitative	–	–	0.01	–	0.00
Verbal	–	–	0.21	–	0.21
e-rater: Issue	–	–	–	0.31	0.27
e-rater: Argument	–	–	–	0.21	0.17

scores is somewhat lower in this example than in TOEFL, as evident from the relatively modest differences between results for Models 2 and 4 and for Models 3 and 5. The result partly reflects the limited relevance of the Quantitative test to writing proficiency. Once again, the difference between use of e-rater scores and use of more general weighting of features is quite limited.

Tables 11 and 12 show the regression coefficients and the standardized partial regression coefficients for the five models for the case with $c_1 = 0.5$. In contrast to TOEFL iBT cases, the weights for the human ratings on the prompts are similar to each other and vary modestly from model to model except for Model 1. Not surprisingly, virtually all contributions of other section data are associated with the Verbal test.

4. Conclusions

The results of this paper have consequences both for the specific field of writing assessment and for the more general problem of use of multiple sources of information in assessment. The results for writing assessment are, at least to some degree, specific to individuals for whom English is not a native language and who are interested in study or professional work in an English-speaking country. Within these restrictions, results imply that use of other portions of a test and use of electronic scoring can substantially improve scoring of a writing assessment beyond what is possible from just use of human holistic scoring. Improvements involve both prediction of the expected score from use of human holistic scoring given the examinee responses (scoring accuracy) and prediction of the expected score from use of holistic scoring if prompts are regarded as randomly selected from pools of prompts (assessment accuracy). These results support score augmentation (Haberman, 2008; Wainer et al., 2001) even in applications in which errors are correlated. Under either the criterion of scoring accuracy or assessment accuracy, use of computer-derived essay features leads to a significant improvement in quality of prediction of true composite scores over prediction only from human scores. The use of the other section scores (Listening, Speaking, and Reading for TOEFL iBT Writing, and Quantitative and Verbal for GRE Analytical Writing) in the prediction model leads to further improvement in prediction quality, especially for assessment accuracy in TOEFL iBT. Validity issues can be raised concerning use of other test scores in writing assessment. This subject is somewhat outside the scope of this paper, but it should be noted that the criterion for assessment accuracy is prediction of writing scores on similar assessments.

One curious result is that the relative weights of features under the e-rater approach were substantially different under the criterion of assessment accuracy than the optimal relative weights of features. Nonetheless, use of optimal weights had little impact on the quality of predictions of true composite scores. Similarly, weighting of composite components had little impact on measures of quality of prediction. Results may simply reflect the use of positively correlated variables with positive weights (Wilks, 1938).

The problem of use of repeater data is a widespread one in educational assessment. In many cases, adequate estimates of test reliability are not available from internal test data due to limited numbers of items and due to items that measure slightly different constructs. This issue arises in TOEFL in the study of assessment accuracy due to the somewhat different skills required for the Integrated prompt and for the Independent prompt (Haberman, 2011). For the data under study, the repeater analysis suggested a correlation of τ_1 and τ_2 of .87, a value noticeably less than 1. The corresponding GRE

correlation is higher but still only .91. The complication with use of repeater data is the non-representative nature of the available samples for voluntary testing programs. Use of MDIA provides an approach to render the non-representative repeater sample more representative of the complete sample of examinees. In the approach, background variables may be used if desired in adjustment to supplement use of observed test scores. This methodology is quite widely applicable to the study of assessment data by use of repeaters. Because weights for different observations can vary considerably when repeater samples differ substantially from general samples, applications are typically most suitable for the large samples examined in this paper.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater v.2. *Journal of Technology, Learning and Assessment*, *4*(3), 1–29. doi:10.1017/s1351324906004189

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, *25*(3), 27–36. doi:10.1609/aimag.v25i3.1774

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*, 146–158. doi:10.1214/aop/1176996454

Davey, T. (2009, April). *Considering fairness and validity in evaluating automated scoring*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley.

Haberman, S. J. (1974). *The analysis of frequency data*. Chicago, IL, University of Chicago Press.

Haberman, S. J. (1979). *Analysis of qualitative data, volume II: New developments*. New York, NY: Academic Press.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*, 971–988. doi:10.1214/aos/1176346715

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229. doi:10.3102/1076998607302636

Haberman, S. J. (2011). *Use of e-rater in scoring of TOEFL iBT writing test* (ETS Research Report No. RR-11-25). Princeton, NJ: ETS.

Haberman, S. J., & Qian, J. (2007). Linear prediction of a true score from a direct estimate and several derived estimates. *Journal of Educational and Behavioral Statistics*, *32*, 6–23. doi:10.3102/1076998606298036

Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics*, *35*, 586–602. doi:10.3102/1076998610375839

Kullback, S. (1959). *Information theory and statistics*. New York, NY: John Wiley.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86. doi:10.1214/aoms/1177729694

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA, Addison-Wesley.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley.

Rudner, L. M. (2001). Informed test component weighting. *Educational Measurement: Issues and Practice*, *20*(1), 16–19. doi:10.1111/j.1745-3992.2001.tb00054.x

Siegert, K. O., & Guo, F. M. (2009). *Assessing the reliability of GMATR® Analytical Writing Assessment* (GMAC Research Report No. RR-09-02). McLean, VA: Graduation Management Admission Council.

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140. doi:10.1111/j.1745-3984.2000.tb01079.x

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Swygert, K. A., & Thissen, D. (2001). Augmented scores – 'borrowing strength' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Erlbaum.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*, 23–40. doi:10.1007/BF02287917

Zhang, Y. (2008). *Repeater analyses for TOEFL iBT®* (ETS Research Memorandum No. RM-08-05). Princeton, NJ: ETS.

Received 4 June 2013; revised version received 2 February 2015

Appendix :

Examples of adjustment by minimum discriminant information

A variety of examples of adjustment by minimum discriminant information can be found in the literature (Haberman, 1974, 1979, 1984; Kullback, 1959). For illustration, let A be a random variable with integer values from 0 to $m \geq 1$. Let $P(A = a)$ and $P(A = a \,|\, u = 1)$ be positive for each integer a from 0 to m. Let $\mathbf{Y} = \mathbf{Z}$ be the m-dimensional random vector with elements Y_a, $1 \leq a \leq m$, such that $Y_a = 1$ if $A = a$ and $Y_a = 0$ if $A \neq a$. For $0 \leq a \leq m$, let $r(a)$ be the ratio $P(A = a)/P(A = a \,|\, u = 1)$. Then $c = r(0)$, element β_a of $\boldsymbol{\beta}$ is $\log\,[r(a)/r(0)]$ for $1 \leq a \leq m$, and $w = r(A)$ is just a proportional adjustment. If h is a real function on the integers $0,1,\ldots,m$, then

$$E(wh(A)|u = 1) = \sum_{a=0}^{m} h(a)r(a)P(A = a|u = 1) = \sum_{a=0}^{m} h(a)P(A = a) = E(h(A)).$$

For another example, consider the case of a random variable A with an unconditional normal distribution and a conditional normal distribution given $u = 1.$ and a random variable B with a normal distribution. Let $m = 2$, and let $\mathbf{Y} = \mathbf{Z}$ have elements A and A^2. Let $\sigma(A)$ be the standard deviation of A, and let $\sigma(A \,|\, u = 1)$ be the conditional standard deviation of A given $u = 1$. In this case, $\boldsymbol{\beta}$ has elements

$$\beta_1 = \frac{E(A)}{V(A)} - \frac{E(A|u = 1)}{V(A|u = 1)},$$

and

$$\beta_2 = -\frac{1}{2V(A)} + \frac{1}{2V(A|u = 1)},$$

and

$$c = \frac{\sigma(A|u = 1)}{\sigma(A)} \exp\left\{\frac{-[E(A)]^2}{2V(A)} + \frac{[E(A|u = 1)]^2}{2V(A|u = 1)}\right\}.$$

It follows that

$$w = \frac{\sigma(A|u=1)}{\sigma(A)} \exp\left\{ -\frac{[A-E(A)]^2}{2V(A)} + \frac{[A-E(A|u=1)]^2}{2V(A|u=1)} \right\}.$$

For real function h on the real line such that $E(h(A))$ is finite, $E(wh(A)|u=1)$ is finite and equal to $E(h(A))$.

Use of repeater data

Let $\mathbf{Y}$ and $\mathbf{Z}$ be $m = K(3K+5)/2$-dimensional vectors with respective elements Y_a and Z_a, $1 \le a \le m$, defined in the following manner:

$$Y_a = \begin{cases} X_a, & 1 \le a \le K, \\ X_{k*}, & a = K+k, 1 \le k \le K, \\ X_k X_{k'}, & a = 2K + k' + k(k-1)/2, 1 \le k' \le k \le K, \\ X_{k*} X_{k'*}, & a = k' + [k(k-1) + K(K+5)]/2, 1 \le k' \le k \le K, \\ X_k X_{k'*} - X_{k'} X_{k*}, & a = K(2K+3) + k' + (k-1)(k-2)/2, 1 \le k' < k \le K, \end{cases} \tag{37}$$

$$Z_a = \begin{cases} X_a, & 1 \le a \le K, \\ X_k, & a = K+k, 1 \le k \le K, \\ X_k X_{k'}, & a = 2K + k' + k(k-1)/2, 1 \le k' \le k \le K, \\ X_k X_{k'}, & a = k' + [k(k-1) + K(K+5)]/2, 1 \le k' \le k \le K, \\ 0, & a = K(2K+3) + k' + (k-1)(k-2)/2, 1 \le k' < k \le K. \end{cases} \tag{38}$$

With this definition,

$$E(w\mathbf{X}|u=1) = E(w\mathbf{X}_*|u=1) = E(\mathbf{X}),$$

$$E(w\mathbf{X}\mathbf{X}'|u=1) = E(\mathbf{X}\mathbf{X}'),$$

$$E(w\mathbf{X}_*\mathbf{X}_*'|u=1) = E(\mathbf{X}\mathbf{X}'),$$

and

$$E(w\mathbf{X}\mathbf{X}_*'|u=1) = E(w\mathbf{X}_*\mathbf{X}'|u=1).$$

The last equation corresponds to the symmetry condition that $E(\mathbf{X}\mathbf{X}_*')$ is equal to $E(\mathbf{X}_*\mathbf{X}')$. It then follows that (31–34) hold.