Research Report

The Best Linear Predictor for True Score From a Direct Estimate and Several Derived Estimates

Shelby J. Haberman

Jiahe Qian

Research & Development

August 2004
RR-04-35

The Best Linear Predictor for True Score From a Direct Estimate and Several Derived Estimates

Shelby J. Haberman and Jiahe Qian

ETS, Princeton, NJ

August 2004

Abstract

Statistical prediction problems often involve both a direct estimate of a true score and covariates of this true score. Given the criterion of mean squared error, this study determines the best linear predictor of the true score given the direct estimate and the covariates. Results yield an extension of Kelley's formula for estimation of the true score to cases in which covariates are present. The best linear predictor is a weighted average of the direct estimate and of the linear regression of the direct estimate onto the covariates. The weights depends on the reliability of the direct estimate and on the multiple correlation of the true score with the covariates. One application of the best linear predictor is to approximate the human true score from the observed holistic score of an essay and from essay features derived from a computer analysis.

Key words: Covariates, direct estimation, essay assessment, Kelley's formula, statistical prediction, holistic scoring

Acknowledgements

The authors thank John Mazzeo and Ida Lawrence for their support and suggestions.

Introduction

Statistical prediction problems may involve both direct estimation of a true score and covariates related to the true score. For example, in the Graduate Management Admission Test®(GMAT™), a final essay score is based on a human holistic score, the direct estimate, and essay features such as number of words in the essay, error rates per word in grammar or usage, and numerical measures of word diversity. The essay features, the covariates, are determined by use of a computer analysis of the essay (Attali & Burstein, 2004). The current procedure in GMAT for essays that can be evaluated employs an holistic score that is an integer in the range 1 to 6 and an e-rater® score, an integer from 1 to 6, generated from the computer analysis. Normally, the reported score is the average of the human holistic score and of the e-rater score; however, an additional reader is employed if the human and e-rater scores differ by more than 1.

The approach used in GMAT is not necessarily an optimal approach to assignment of a final score to an essay. This remark applies even if the true essay score is regarded as the average holistic score an essay would receive if rated by an arbitrarily large number of human raters (Lord & Novick, 1968, p. 2).

In this study, a continuation of work presented earlier (Qian & Haberman, 2003), the criterion of mean squared error is used to determine the best linear predictor of a true score based on a direct estimate and on covariates. In Section 1, this predictor is considered under the assumption that all relevant population parameters are known. In this ideal case, the best linear predictor is shown to be a weighted average of two components. The first component is the direct estimate. The second component is the regression of the direct estimate onto the covariates. The weights assigned to the components depend on the reliability of the direct estimate and on the multiple correlation between the direct estimate and the covariates. The mean squared error of the optimal linear predictor is shown to depend on the variance of the direct estimate, on the reliability of the direct estimate, and on the multiple correlation of the true score and the covariates. Results of this section can be regarded as a generalization of Kelley's formula to the case of covariates (Kelley, 1947, p. 409). Required arguments are familiar from treatments of linear prediction in classical test theory (Holland & Hoskens, 2003; Lord & Novick, 1968).

In Section 2, estimation of the best linear predictor and of the mean squared error are considered. Estimation is described for a simple random sample of essays from a large population. Because reliability must be estimated, it is assumed that $m > 1$ independently obtained holistic scores are available in the sample for each essay and that all covariates are observed for each essay. Given these data, estimation of parameters is relatively straightforward, at least for large samples. Standard treatments of classical test theory provide basic background (Lord & Novick, 1968, chap. 8), as do classical treatments of statistical inference (Rao, 1973, chap. 4).

In Section 3, the methods developed in Sections 1 and 2 are applied to essays from GMAT and from the Test of English as a Foreign Language™ (TOEFL®). A notable feature of the analysis is the relatively low weight assigned to the human holistic score. This result reflects some limitations in the reliability of holistic scores and a relatively high multiple correlation of human holistic scores and computer-derived essay features.

As discussed in Section 4, results in this report suggest that scoring procedures such as those used in GMAT should be given considerably higher weight to computer-generated essay features than is currently the case. Policy issues may arise that involve public perceptions concerning the reduced weight given to the human rater, and there is some question concerning the effect on examinee performance if they are aware that a very large fraction of the grade on their essay is determined by a computer program.

1 The Best Linear Predictor of the True Score

To obtain the best linear predictor of the true score from a direct estimate and from the available covariates, some elementary notation and a basic probability model are required. Let θ, the true score, be a random variable with expectation $E(\theta)$ and positive variance $V(\theta)$, let h, the direct estimate, be a random variable such that the error $e = h - \theta$ in estimation of θ has expectation 0 and positive variance $V(e)$ (Lord & Novick, 1968, p. 31). Thus the observed score h has mean $E(h) = E(\theta)$ and variance

$$V(h) = V(\theta) + V(e). \tag{1}$$

The reliability coefficient is

$$\tau^2 = V(\theta)/V(h) = V(\theta)/[V(\theta) + V(e)] \tag{2}$$

(Lord & Novick, 1968, p. 208). Under the assumptions made concerning the variances of the true score θ and the error e, $0 < \tau^2 < 1$.

Let $\mathbf{d}$ be a q-dimensional vector of covariates d_j, $1 \leq j \leq q$, with mean $\mathbf{E}(\mathbf{d})$ and positive definite covariance matrix $\mathbf{C}(\mathbf{d})$. Assume that the estimation error e is uncorrelated with the covariates d_j, $1 \leq j \leq q$. Let $\mathbf{C}(\mathbf{d}, e)$ denote the vector of covariances of the error e and the covariates d_j, $1 \leq j \leq q$. This information suffices to specify the best linear predictor of the true score θ based on the observed score h and the vector $\mathbf{d}$ of covariates.

To describe the best linear predictor of the true score, first consider the standard formula for the best linear predictor of the direct estimate h based on the covariate vector $\mathbf{d}$. For q-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ with respective coordinates x_i and y_i, let

$$\mathbf{x}'\mathbf{y} = \sum_{i=1}^{q} x_i y_i.$$

Then the best linear predictor of h from $\mathbf{d}$ is

$$f = E(h) + \boldsymbol{\gamma}'[\mathbf{d} - \mathbf{E}(\mathbf{d})], \tag{3}$$

where

$$\boldsymbol{\gamma} = [\mathbf{C}(\mathbf{d})]^{-1}\mathbf{C}(\mathbf{d}, h). \tag{4}$$

Note that $\mathbf{C}(\mathbf{d}, h)$ is the vector of covariances of d_j and h for $1 \leq j \leq q$ (Lord & Novick, 1968, p. 267).

The best linear predictor of the direct estimate h from the covariate vector $\mathbf{d}$ is the same as the best linear predictor of the true score θ from the covariate vector $\mathbf{d}$. This claim is easily verified. Because the error e is assumed to have expectation 0 and to be uncorrelated with the covariate vector $\mathbf{d}$, the covariance vector $C(\mathbf{d}, \theta)$ for the covariates d_j and the true score θ is the same as the covariance vector $C(\mathbf{d}, h)$ for the covariates d_j and the direct estimate h (Holland & Hoskens, 2003). As already noted, the direct estimate h and the true score θ satisfy $E(h) = E(\theta)$. Thus

$$f = E(\theta) + \boldsymbol{\gamma}'[\mathbf{d} - \mathbf{E}(\mathbf{d})]$$

and

$$\boldsymbol{\gamma} = [\mathbf{C}(\mathbf{d})]^{-1}\mathbf{C}(\mathbf{d}, \theta).$$

It follows that f is also the best linear predictor of the true score θ from the covariate vector $\mathbf{d}$.

The residual for prediction of the direct estimate h by the covariate vector $\mathbf{d}$ is

$$r = h - f.$$

The corresponding residual for prediction of the true score θ by the covariate vector $\mathbf{d}$ is

$$u = \theta - f,$$

so that $r = u + e$.

The mean squared error for linear prediction of the direct estimate h by the covariate vector $\mathbf{d}$ is then

$$V(r) = V(h) - V(f), \tag{5}$$

where

$$V(f) = \boldsymbol{\gamma}'\mathbf{C}(\mathbf{d})\boldsymbol{\gamma} \tag{6}$$

(Rao, 1973, p. 266). If $\rho(h, \mathbf{d})$ is the multiple correlation coefficient of the direct estimate h and $\rho^2(h, \mathbf{d})$ is the square of $\rho(h, \mathbf{d})$, then

$$\rho^2(h, \mathbf{d}) = V(f)/V(h), \tag{7}$$

so that

$$V(r) = V(h)[1 - \rho^2(h, \mathbf{d})]. \tag{8}$$

In like manner, the mean squared error for linear prediction of the true score θ by the covariate vector $\mathbf{d}$ is

$$V(u) = V(\theta) - V(f). \tag{9}$$

It is assumed in this paper that the residual variance $V(u)$ is positive, so that the true score is not determined by an affine function of the covariate vector $\mathbf{d}$. By (1),

$$V(r) = V(u) + V(e). \tag{10}$$

Thus the multiple correlation $\rho(\theta, \mathbf{d})$ of the true score θ and the covariate vector $\mathbf{d}$ satisfies

$$\rho^2(\theta, \mathbf{d}) = V(f)/V(\theta), \tag{11}$$

$$V(u) = V(\theta)[1 - \rho^2(\theta, \mathbf{d})], \tag{12}$$

and

$$\rho^2(\theta, \mathbf{d}) = \rho^2(h, \mathbf{d})/\tau^2. \tag{13}$$

By (12), (13), and the assumption that the residual variance $V(u)$ is positive, it follows that the multiple correlation coefficient $\rho(\theta, \mathbf{d})$ is less than 1, so that

$$\rho^2(h, \mathbf{d}) < \tau^2. \tag{14}$$

Given these basic results, it is then relatively easily shown that the best linear predictor of the true score θ based on the direct estimate h and on the covariate vector $\mathbf{d}$ is

$$t = \alpha h + (1 - \alpha) f, \tag{15}$$

where

$$\alpha = V(u)/V(r). \tag{16}$$

By (10), the weight α assigned to the direct estimate is always between 0 and 1. A similar comment applies to the weight $1 - \alpha$ assigned to the best linear predictor f of the direct estimate based on the covariate vector $\mathbf{d}$. The weight α assigned to the direct estimate can be expressed in terms of the reliability τ^2 and the multiple correlation coefficient $r(\theta, \mathbf{d})$ of the true score θ and the covariate vector $\mathbf{d}$, for (2), (8), and (13) imply that

$$\alpha = \frac{\tau^2[1 - \rho^2(\theta, \mathbf{d})]}{1 - \tau^2\rho^2(\theta, \mathbf{d})}.$$

The weight α increases with an increase in the reliability τ^2 and decreases with an increase in the multiple correlation $\rho(\theta, \mathbf{d})$ of the true score θ and the covariate vector $\mathbf{d}$. If $\rho(\theta, \mathbf{d})$ is 0, then the weight is the same as in Kelley's formula.

To verify that the best linear predictor t of the true score satisfies (15), consider the mean squared error

$$L(a, c, \mathbf{b}) = E([\theta - a - ch - \mathbf{b}'\mathbf{d}]^2) \tag{17}$$

from prediction of the true score θ by a function $a + ch + \mathbf{b}'\mathbf{d}$, where a and c are real constants and $\mathbf{b}$ is a constant q-dimensional vector. The mean squared error $L(a, c, \mathbf{b})$ is minimized if

$$a = E(\theta) - cE(h) - \mathbf{b}'\mathbf{E}(\mathbf{d}) = (1 - c)E(\theta) - \mathbf{b}'\mathbf{E}(\mathbf{d}), \tag{18}$$

$$cV(h) + \mathbf{b}'\mathbf{C}(\mathbf{d}, \theta) = \mathbf{C}(h, \theta), \tag{19}$$

and

$$c\mathbf{C}(\mathbf{d}, h) + \mathbf{C}(\mathbf{d})\mathbf{b} = \mathbf{C}(\mathbf{d}, \theta) \tag{20}$$

(Rao, 1973, p. 266). Recall that the covariance vector $\mathbf{C}(\mathbf{d}, h)$ is the same as the covariance vector $\mathbf{C}(\mathbf{d}, \theta)$, so that (20) implies that

$$\mathbf{b} = (1 - c)\boldsymbol{\gamma}. \tag{21}$$

By (18),

$$a + ch + \mathbf{b}'\mathbf{d} = ch + (1 - c)f. \tag{22}$$

In addition, the covariance $C(h, \theta)$ of the direct estimate h and the true score θ is the variance $V(\theta)$ of θ (Lord & Novick, 1968, p. 57). By (5), (6), (16), and (20), the optimal c is α, so that the optimal predictor is t.

The residual from prediction of θ by t is

$$v = \theta - t = (1 - \alpha)u - \alpha e. \tag{23}$$

Because u and e have 0 expectations, v also has 0 expectation. Because u, a linear function of θ and $\mathbf{d}$, is uncorrelated with the error e, it follows from (10) that the mean squared error of prediction of the true score θ by the direct estimate h and the covariate vector $\mathbf{d}$ is the variance $V(v)$ of v, and

$$V(v) = (1 - \alpha)^2 V(u) + \alpha^2 V(e) = V(e)V(u)/V(r) = \left(\frac{1}{V(e)} + \frac{1}{V(u)}\right)^{-1}. \tag{24}$$

Note that $V(v)$ is less than either the variance $V(e)$ of the error of the direct estimate or the variance $V(u)$ of the error from use of the predictor f as an estimate of the true score θ. If the multiple correlation $\rho(\theta, \mathbf{d})$ is 0, then the variance $V(v)$ is the variance of Kelley's estimate.

2 Estimation of the Best Linear Predictor

To estimate the best linear predictor t, consider a random sample of size $n > q + 1$ from the population used to define t. Assume that the underlying population is either infinite or so large that finite sampling corrections can be ignored. For each observation i, $1 \leq i \leq n$, let $m_i \geq 1$ direct estimates h_{ij}, $1 \leq j \leq m_i$, $1 \leq i \leq n$, be available, and assume that at least one m_i exceeds 1 and that the m_i are selected without regard to any characteristics of the essays under study. The requirement of some multiple direct estimates is essential in order to determine the variance $V(e)$. In use of e-rater, essays used to construct the regression analysis are assessed by more than one rater, so that the requirement imposed here is consistent with current practice with e-rater. In the analysis of essays in Section 4, each m_i will be 2; however, little is lost by consideration of the more general case.

Let the true score for observation i be θ_i, so that the error for replication j and observation i is $e_{ij} = h_{ij} - \theta_i$. Let the vector of covariates for observation i be $\mathbf{d}_i$. For each observation i and replication j, it is assumed that the joint distribution of h_{ij}, θ_i, and $\mathbf{d}_i$ is the same as the joint distribution of h, θ, and $\mathbf{d}$. The added assumptions are imposed that the errors e_{ij} for the direct estimates are all uncorrelated. To assist in some formulas, a variable $\bar{e}$ will be introduced that is uncorrelated with $\mathbf{d}$ and θ, has mean 0, and has variance $V(e)/m$, where

$$m = \frac{1}{n^{-1} \sum_{i=1}^{n} m_i^{-1}}$$

is the harmonic mean of the m_i. If m is an integer and m_i is at least m, then $\bar{e}$ has the same mean and variance as does the average $\bar{e}_i$ of the e_{ij}, $1 \leq j \leq m$.

Given these conditions, estimation of the best linear predictor t is straightforward. For each observation i, let $\bar{h}_i$ be the average of the h_{ij}, $1 \leq j \leq m$, so that the average error $\bar{e}_i = \bar{h}_i - \theta$ for observation i has mean 0 and variance $V(e)/m_i$ and is uncorrelated with $\mathbf{d}_i$. One may then estimate the expectation $E(h) = E(\theta)$ by the grand mean

$$\bar{h} = n^{-1} \sum_{i=1}^{n} \bar{h}_i. \tag{25}$$

The expectation $\mathbf{E}(\mathbf{d})$ is then estimated by the sample mean

$$\bar{\mathbf{d}} = n^{-1} \sum_{i=1}^{n} \mathbf{d}_i. \tag{26}$$

The covariance matrix $\mathbf{C}(\mathbf{d})$ is estimated by the sample covariance

$$\bar{\mathbf{C}}(\mathbf{d}) = (n-1)^{-1} \sum_{i=1}^{n} (\mathbf{d}_i - \bar{\mathbf{d}})(\mathbf{d}_i - \bar{\mathbf{d}})', \tag{27}$$

where $\mathbf{xy}'$ is the q by q matrix with elements $x_j y_k$ for $1 \leq j \leq q$ and $1 \leq k \leq q$ if $\mathbf{x}$ and $\mathbf{y}$ are vectors of dimension q with respective coordinates x_j and y_j for $1 \leq j \leq q$. The covariance vector $\mathbf{C}(\mathbf{d}, \theta) = \mathbf{C}(\mathbf{d}, h)$ is then estimated by

$$\bar{\mathbf{C}}(\mathbf{d}, h) = (n-1)^{-1} \sum_{i=1}^{n} (\bar{h}_i - \bar{h})(\mathbf{d}_i - \bar{\mathbf{d}}). \tag{28}$$

Thus the vector $\boldsymbol{\gamma}$ of regression coefficients may be estimated by

$$\mathbf{g} = [\bar{\mathbf{C}}(\mathbf{d})]^{-1} \bar{\mathbf{C}}(\mathbf{d}, h). \tag{29}$$

The approximation to f is then

$$\hat{f} = \bar{h} + \mathbf{g}'(\mathbf{d} - \bar{\mathbf{d}}). \tag{30}$$

For observation i, $\hat{h}$ is

$$\hat{h}_i = \bar{h} + \mathbf{g}'(\mathbf{d}_i - \bar{\mathbf{d}}). \tag{31}$$

To complete estimation, it is necessary to approximate α. To do so, $V(e)$ and $V(u)$ must be estimated. Estimation of $V(e)$ is a straightforward manner given customary results for one-way analysis of variance. An unbiased estimate of $V(e)$ is

$$\bar{V}(e) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m_i} (h_{ij} - \bar{h}_i)^2}{\sum_{i=1}^{n} (m_i - 1)} \tag{32}$$

(Lord & Novick, 1968, p. 158).

The case of $V(u)$ is a bit more complex. Let

$$\bar{r}_i = \bar{h}_i - \hat{f}_i \tag{33}$$

be the residual from regression of the $\bar{h}_i$ on the $\mathbf{d}_i$ for $1 \leq i \leq n$. Then the residual mean square error

$$\bar{V}(\bar{r}) = (n - q - 1)^{-1} \sum_{i=1}^{n} \bar{r}_i^2 \tag{34}$$

is a consistent estimate of the variance

$$V(\bar{e} + u) = V(u) + V(e)/m.$$

If $\mathbf{d}$ has a continuous distribution, if each m_i is m, and if the residual u is independent of $\mathbf{d}$, then $\bar{V}(\bar{r})$ is unbiased (Rao, 1973, p. 227). It follows that $V(u)$ has the estimate

$$\bar{V}(u) = \bar{V}(\bar{r}) - m^{-1}\bar{V}(e). \tag{35}$$

At this point, the natural estimate of α is

$$\hat{\alpha} = \bar{V}(u)/[\bar{V}(e) + \bar{V}(u)]. \tag{36}$$

The only complication is that $\bar{V}(e)$ and $\bar{V}(u)$ need not be positive. One may adopt the convention that $\hat{\alpha}$ is 0 if $\bar{V}(u) \leq 0$ (Bock & Petersen, 1975).

Given $\hat{\alpha}$, h, and $\hat{f}$, t may be estimated by

$$\hat{t} = \hat{\alpha}h + (1 - \hat{\alpha})\hat{f}. \tag{37}$$

The mean square error $V(v)$ may then be approximated by

$$\hat{V}(v) = \bar{V}(e)\bar{V}(u)/[\bar{V}(e) + \bar{V}(u)]. \tag{38}$$

3 Data Sources and Empirical Results

The results of Sections 1 and 2 are readily applied to essay assessment. In this section, data and variables used in the analysis are described, and results of the analysis are presented.

Data Sources and Prompts Used in Essay Assessment

The data used in the study are essays generated by four essay prompts, with the first two prompts from GMAT and the other two from TOEFL. For each prompt, about 5,000 essays are available. Essays are only used if assigned scores from 1 to 6 by both initial raters and if they contain at least 25 words (Haberman, 2004). These restrictions remove responses that do not satisfy minimal criteria for essays responsive to the prompt. For each essay, the initial $m = 2$ holistic scores obtained from readers are used in the analysis.

Covariates in the Analysis

Several choices of covariates vectors were considered in the analysis. These vectors are based on the following essay features (Attali & Burstein, 2004; Burstein, Chodorow, & Leacock, in press; Haberman, 2004).

Number of Words

The number W of words in the essay.

Number of Characters

The number C of alphanumeric characters in the essay.

Average Word Length

The ratio $A = C/W$ is the average number of characters per word.

Error Rates

For a given essay, let N_G be the number of grammatical errors detected by e-rater Version 2.0, let N_U be the number of usage errors detected, let N_M be the number of detected errors in mechanics, and let N_S be the number of detected errors in style. The corresponding rates per word are $R_G = N_G/W$, $R_U = N_U/W$, $R_M = N_M/W$, and $R_S = N_S/W$. A summary total is $R_T = R_G + R_U + R_M + R_S$. A special case of mechanical errors, spelling errors, is also of interest. Here N_P is the number of detected spelling errors, and $R_P = N_P/W$ is the rate per word.

Number of Arguments

Let D be the number of discourse elements in the essay, and let D_8 be the minimum of D and 8. (In a standard five-paragraph essay, there are 8 discourse elements.)

Average Argument Length

The ratio $L = W/D$ is the average number of words in a discourse element.

Standard Frequency Index

The Breland Standard Frequency Index (SFI) (Breland, 1996; Breland, Jones, & Jenkins, 1994) is a measure of word frequency. The measure is on a logarithmic scale, and lower numbers indicate less frequent words. In Version 2.0 of e-rater, the fifth lowest SFI value (B_5) is used for essay words in the list of 179,195 words with an SFI. The median B of the SFI for essay words in the list is also considered in the regression analysis in this report.

Measures of Word Diversity

Simpson's index S (Gini, 1912; Simpson, 1949) measures the probability that two distinct randomly selected words from an essay are the same. The ratio T is the ratio D/M in an essay of the number D of distinct content words to the total number M of content words. Here content words are words that are normally used in search engines and indexes. Thus words such as "the" and "and" are excluded.

Selection of Specific Words

Let Z_j be the jth most frequently used content word among all essays available for a particular prompt, and let F_j be the number of times Z_j appears in an essay. The variable U_j is $(F_j/M)^{1/2}$. Two other variables, τ and e_6, used in the regression analysis are obtained from the content vector analysis of e-rater (Attali & Burstein, 2004; Burstein et al., in press; Haberman, 2004). The variable τ is the score group with the highest similarity measure to the observed essay in terms of the observed ratios F_j/M, and e_6 is a cosine measure of similarity of the F_j/M to the observed F_j/M in the highest score group of essays. The variables τ and e_6 are not entirely satisfactory for use in the analysis considered in this paper, for their calculation is affected by essays other than the essay under study. They are considered in this report to provide some indication of the behavior of the regression used in e-rater; however, any results involving τ and e_6 should be approached with great caution. The definition of U_j is also affected by the specific essays found in the sample, but the effect is rather small in large samples (Haberman, 2004).

Sources of Variables

Variables W, C, A, N_G, N_U, N_M, N_S, R_G, R_U, R_M, R_S, R_T, D, D_8, L, B_5, B, T, τ, and e_6 are computed by e-rater software. The variables S and U_j were obtained by one of the authors (Haberman, 2004).

Covariate Vectors Used

In all, seven covariate vectors were considered. In Vector 1, the elements were W, W^2, L, D_8, R_G, R_U, R_M, R_S, A, T, τ, e_6, and B_5. This vector is used in e-rater version 2.0.

In Vector 2, the e-rater variables from content vector analysis were removed from Vector 1, so that the elements were W, W^2, L, D_8, R_G, R_U, R_M, R_S, A, T, and B_5. This omission is considered to eliminate variables defined by reference to essays other than the one to be rated.

In Vector 3, the only variables are $\log(C)$ and $\log(R_T)$. This vector is a rather minimal selection that only considers a length measure and an error rate measure.

In Vector 4, Vector 3 is supplemented by B, so that $\log(C)$, $\log(R_T)$, and B are the coordinates. Addition of B provides a measure of vocabulary level.

In Vector 5, $C^{1/2}$, A, $(R_G + R_U)^{1/2}$, $R_P^{1/2}$, $(R_M - R_P)^{1/2}$, and B are the covariates. This choice is based on empirical work by one of the authors (Haberman, 2004). There is a length measure, a word length measure, error rate measures that reflect types of errors that appear to correlate with human holistic scores, and a vocabulary measure.

In Vector 6, $C^{1/2}$, $(R_G + R_U)^{1/2}$, $R_P^{1/2}$, $(R_M - R_P)^{1/2}$, B, and $S^{1/2}$ are the covariates. The measure of word length has been replaced by a measure of word diversity.

In Vector 7, $C^{1/2}$, $(R_G + R_U)^{1/2}$, $R_P^{1/2}$, $(R_M - R_P)^{1/2}$, B, $S^{1/2}$, and U_j, $1 \le j \le 50$, are the covariates. Thus Vector 6 is supplemented by measures of specific word choice.

Results

Results are summarized in Tables 1 and 2. In Table 1, the sample size and $\bar{V}(e)$ are provided for each prompt. In Table 2, $\bar{V}(u)$, $\hat{\alpha}$, and $\hat{V}(v)$ are provided. Of note is the consistent finding that the estimated optimal weight on the human score is less than 0.5, with the optimal weight at times less than 0.2. For each prompt, it is possible to find a vector of covariates such that the estimated variance of v is less than 0.1. The covariates used in e-rater perform quite well relative to other selections, although interpretation of results is complicated if e_6 and τ are included. It is worth noting that an appreciable improvement in results, especially for GMAT prompts, is achieved by use of more U_j terms than are found in Vector 7. For instance, in the first GMAT prompt, use of the first 172 of the U_j rather than just the first 50 yields $\hat{V}(v)$ of 0.059, while in the second GMAT prompt, use of the first 174 of the U_j yields $\hat{V}(v)$ of 0.033 (Haberman, 2004).

For some perspective on these results, note that the estimated mean squared error from

Table 1.
Variability of Holistic Scores

Program	Prompt	Count	$\bar{V}(e)$
GMAT	1	5183	0.356
GMAT	2	5158	0.346
TOEFL	1	4895	0.275
TOEFL	2	4884	0.259

use of the average of m holistic scores is $\bar{V}(e)/m$. For the first GMAT prompt, it follows that 10 raters yield a mean squared error comparable to that provided by one human rater and a careful selection of features. Achievable results for TOEFL are comparable to those for three or four readers.

4 Findings

This study determines the best linear predictor of a true score based on a direct estimate and a vector of covariates and determines the resulting mean squared error. A simple estimation procedure is also presented for this linear predictor. Application of results to essay scoring suggests that the true score for holistic essay scores assigned by raters can be estimated with relatively good accuracy by use of one human rater and by use of covariates generated by computer analysis of essays.

The proposed estimation procedure differs considerably from the procedure currently found in GMAT in that a continuous approximation of the true essay score is produced that gives the human holistic score for the essay a relatively small weight. Use of the continuous approximation requires the perception that there is a population of raters who might grade an essay and that there is a distribution of human holistic scores that has a mean and a variance. In this framework, there is no pretense of a true rating of the essay that is an integer from 1 to 6 provided by an infinitely skilled reader.

Because the essay ratings suggested in this study are essentially continuous, it is possible to consider equating of essay scores. Given that the mean squared error of the proposed essay rating is somewhat smaller than the mean squared error of the current system of score assignment, it is also plausible that the proposed weighting might improve reliability and

Table 2.
Mean Squared Errors and Weights for Selected Covariate Vectors

Program	Prompt	Vector	$\bar{V}(u)$	$\hat{\alpha}$	$\hat{V}(v)$
GMAT	1	1	0.083	0.190	0.067
GMAT	1	2	0.210	0.371	0.132
GMAT	1	3	0.251	0.414	0.147
GMAT	1	4	0.240	0.402	0.143
GMAT	1	5	0.215	0.376	0.134
GMAT	1	6	0.211	0.373	0.133
GMAT	1	7	0.105	0.228	0.081
GMAT	2	1	0.036	0.095	0.033
GMAT	2	2	0.071	0.171	0.059
GMAT	2	3	0.126	0.267	0.092
GMAT	2	4	0.107	0.236	0.082
GMAT	2	5	0.080	0.188	0.065
GMAT	2	6	0.076	0.179	0.062
GMAT	2	7	0.051	0.128	0.044
TOEFL	1	1	0.083	0.232	0.064
TOEFL	1	2	0.093	0.253	0.070
TOEFL	1	3	0.144	0.344	0.095
TOEFL	1	4	0.127	0.315	0.087
TOEFL	1	5	0.111	0.288	0.079
TOEFL	1	6	0.101	0.268	0.074
TOEFL	1	7	0.096	0.258	0.071
TOEFL	2	1	0.097	0.272	0.070
TOEFL	2	2	0.115	0.308	0.080
TOEFL	2	3	0.169	0.395	0.102
TOEFL	2	4	0.152	0.370	0.096
TOEFL	2	5	0.125	0.325	0.084
TOEFL	2	6	0.123	0.322	0.083
TOEFL	2	7	0.113	0.305	0.079

validity of essay scores; however, this possibility can only be verified with further research.

The proposed method of essay scoring has potential problems. It is not clear whether the public can be persuaded that a reduced weight to human holistic scores is desirable, no matter what statistical arguments may be made. Perhaps this potential concern can be reduced by emphasizing that the essay features used by the computer analysis do provide measures of writing quality that are strongly related to human holistic scores and that the collection of human holistic scores of essay responses has been employed to determine the final predictor of the essay score.

A further potential difficulty is that behavior of essay writers might change if they are aware of the scoring procedure used to evaluate the essay. Exploiting this knowledge might be difficult in practice, and, in any event, research concerning the relationship of essay features to human holistic scores is publicly available, at least to a substantial extent (Haberman, 2004).

In conclusion, it appears that the proposed regression-based method of essay assessment should be seriously considered in those cases in which essays are available in computer-readable form and in which human holistic scoring is employed.

References

Attali, Y., & Burstein, J. (2004). *Automated essay scoring with e-rater v.2.0.* Paper presented at the Annual Conference of the International Association for Educational Assessment (IAEA), Philadelphia, PA.

Bock, R. D., & Petersen, A. C. (1975). A multivariate correction for attenuation. *Biometrika*, *62*, 673–678.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological science*, *7*, 96–99.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (College Board Rep. no. 94-4). Princeton, NJ: ETS.

Burstein, J., Chodorow, M., & Leacock, C. (in press). Automated essay evaluation: The Criterion Online Service. *AI Magazine*, *25*.

Gini, C. (1912). *Variabilitá e mutabilitá: Contributo allo studio delle distribuzioni e delle relazioni statische*. Bologna, Italy: Cuppini.

Haberman, S. J. (2004). *Statistical and measurement properties of features used in essay assessment.* Manuscript in preparation.

Holland, P. W., & Hoskens, M. (2003). Classical test theory as a first-order item response theory: Application to true-score prediction from a possibly non-parallel test. *Psychometrika*, *68*.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University Press.

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

Qian, J., & Haberman, S. J. (2003). *The best linear predictor for true score from a direct estimate and a derived estimate.* Paper presented at the annual Joint Statistical Meetings of the American Statistical Association, San Francisco, CA.

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

Simpson, E. H. (1949). The measurement of diversity. *Nature*, *163*, 688.