Research Report

Application of Best Linear Prediction and Penalized Best Linear Prediction to ETS Tests

ETS RR–20-08

Shelby J. Haberman

December 2020

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Application of Best Linear Prediction and Penalized Best Linear Prediction to ETS Tests

Shelby J. Haberman

Consultant, Jerusalem, Israel

Best linear prediction (BLP) and penalized best linear prediction (PBLP) are techniques for combining sources of information to produce task scores, section scores, and composite test scores. The report examines issues to consider in operational implementation of BLP and PBLP in testing programs administered by ETS.

Keywords True scores; scoring accuracy; assessment accuracy; composite scores

doi:10.1002/ets2.12290

Best linear prediction (BLP) has been widely explored for combining information from different components of a test to improve inferences concerning the test components. This approach has been used to determine both whether subscores of tests are worth reporting at all and, if so, how they should be reported (Haberman, 2008; Haberman et al., 2009; Puhan et al., 2008, 2010; Sinharay et al., 2007, 2010, 2011, 2018; Wainer et al., 2001). One indication of professional recognition of BLP is the 2009 National Council on Measurement in Education Award for Technical or Scientific Contributions to the Field of Educational Measurement awarded to Shelby Haberman, Sandip Sinharay, and Gautam Puhan. Applications have also been made to combining human and machine-generated scores for evaluation of either individual constructed responses or test sections composed of constructed responses (Haberman, 2011; Haberman et al., 2015; Haberman & Qian, 2007).

A potential complication in using BLP for subscores has been the differential impact on subgroups (Haberman & Sinharay, 2011, 2013; Sinharay & Haberman, 2014). As shown in these references, the impact of the issue varies appreciably from assessment to assessment, although the impact can often be minor. Penalized best linear prediction (PBLP) is a generalization of BLP that can be used in scoring assessments at different levels when the impact of subgroup membership is relatively large with BLP. Use of PBLP rather than BLP generally involves an attempt to reduce subgroup biases that present issues of fairness. A change in a scoring procedure that substantially advantages or disadvantages a group of test takers of interest needs a strong validity justification and may require legal scrutiny.

This report describes the use of BLP and PBLP for task scoring, section scoring, and composite scoring. In the examples under study, at least two tasks used for scoring are constructed responses with scores that are not deterministic. For example, reasonable people can disagree about the quality of an essay or about the quality of a spoken response. In each such task under consideration, the task score for the assessments under review normally involves both a human rater and computer-generated numerical variables that describe the response. Three ETS assessments are examined: the *TOEFL iBT*® test, the *GRE*® General test, and the *PRAXIS*® assessment's Core Writing section. In all these cases, task scores and section scores are of interest; however, only in TOEFL iBT is a composite score relevant given current reporting practices. The relevant levels depend on the testing program. Human scores studied are always numerical holistic scores for a constructed response. It should be emphasized that constructed responses do exist in which scoring is unambiguous and can be done by computer. For example, in the *SAT*® test Mathematics section, test takers are asked to provide a numerical answer rather than to select a response from a list of alternatives. If the sum of 105 and 224 is requested, then the correct answer is unambiguously defined. In the examples under study, this type of constructed response does not happen to arise.

In the case of TOEFL iBT (*TOEFL*®) test, the Writing and Speaking sections only include prompts that require constructed responses, and these responses also affect the total score derived by computing the sum of the scale scores for the

Corresponding author: S. J. Haberman, E-mail: haberman.statistics@gmail.com

Listening, Reading, Speaking, and Writing sections. The Writing section consists of two essay prompts that are normally scored by both a trained human rater and the *e-rater*® automated scoring engine (Attali & Burstein, 2006). A human score for a response is normally an integer from 1 to 5, with 5 the best score and 1 the worst score, although in certain cases, a score of 0 is recorded. The Speaking section as of August 2019 consists of four speaking tasks that are normally scored by a trained human rater and by the *SpeechRater*® automated scoring service (Higgins et al., 2011). Human scores are normally integers from 1 to 4, with 1 the worst score and 4 the best score, although a score of 0 can be obtained in certain cases. The Listening section contains 28 selected responses that are used to compute the scale score, while the Reading section contains 30 selected responses used to compute the scale score. Test takers may receive additional items for Listening and Reading that do not contribute to the reported score but are used to pretest items and to link items and scores for different test administrations. Neither Listening nor Reading includes constructed responses. Score users receive linked scale scores for each test section together with the total score. Scores received for each section are integers between 0 and 30. The total score is the sum of these four section scores.

In the case of GRE General, the section on Analytical Writing (GRE Writing) is composed of two prompts that require essay responses. These responses are normally scored by one human rater and by e-rater. They yield a single unlinked reported score provided to test takers and institutions. This report considers the scores for individual essays and the score for the section. For each prompt, a human score is normally an integer from 1 to 6, with 6 the best score and 1 the lowest score. In special cases, a score of 0 is possible.

The PRAXIS Core Writing (PRAXIS Writing) test consists of two essay prompts—an argumentative prompt and an informative/explanatory prompt—and 34 selected responses. A total raw score for the essay responses, a total raw score for the selected responses, a weighted total raw score for the test, and a linked scale score for the test are produced. Each essay is normally scored by one human rater and by e-rater. For human raters, the best score is 6, and the worst is 1, although a 0 can occur in special cases. This report considers the scores of individual essays, the total score for the two essays, and the raw score for the entire test.

In the examples under study, traditional item response theory (IRT) is not applicable owing to the limited number of constructed responses involved and the use of computer-generated features that are continuous variables. Therefore analysis will involve linear methods and classical test theory (Lord & Novick, 1968).

In this report, the section Best Linear Prediction and Penalized Best Linear Prediction provides a general description of BLP and PBLP. The discussion relies heavily on Yao et al. (2019a, 2019b); however, further background is provided by Haberman and Qian (2004, 2007), Haberman and Yao (2015), and Haberman et al. (2015). As in these publications, the predictions are of true scores, a classical psychometric concept (Lord & Novick, 1968), although the use of true scores in the cited papers and reports is not entirely conventional. There are two separate cases of true scores to predict, both of which are important to consider. These definitions strongly affect applications.

The case of scoring accuracy involves the true rating score. This version of a true score is only really relevant when the basis of scoring of constructed responses is human scoring. For a specific constructed response for a particular task, the true rating score is the expected score assigned by a rater chosen from the rater pool. The rating prediction error is the difference between the observed score and the true rating score. If only human raters are used to score a constructed response, the true rating score and the observed score are not typically the same, for different human raters may score the same response differently. Thus there is a rating prediction error when the true rating score for a constructed response is predicted by one or more observed scores provided by human raters. On the other hand, computer scoring of a selected response involves a true rating score and an observed score that are the same. If one ignores the possibility of equipment malfunction, the same response always receives the same score so that no rating prediction error exists.

Computer-generated numerical features of a constructed response may be used to predict a true rating score in cases, such as essays and speech samples, in which different human raters may not necessarily provide the same score. Alternatively, human scores and computer-generated numerical features may be used together to predict the true rating score. In either of these cases, observed scores and true rating scores are typically not the same so that rating prediction error exists. Rating prediction error exists even though the computer-generated features should not vary in the same way as human scores do. The computer features assigned to a constructed response should always be the same for the same computer program; otherwise, the program is defective. Rating prediction error arises because the true rating score of the constructed response is not observed and the computer-generated features never provide perfect descriptions of the constructed response. Use of linear combinations permits a true rating score to apply to section scores and composite scores as well as to individual tasks. Once again, rating prediction error here involves scoring accuracy. The section Scoring Accuracy considers the procedures required to apply scoring accuracy to assessments.

The second case involves the true assessment score. The true assessment score is the expected observed score for a test taker on a parallel task, section, or test. The true assessment score, which applies to all assessments, corresponds to the customary definition of reliability used in psychometrics. In principle, the true assessment score for the TOEFL Listening section might be thought of as the expected score the test taker would receive by taking the same test again; however, this definition is not quite right, because the act of taking a test once can influence performance at a later time, even if the time elapsed is quite small. A more accurate definition for TOEFL Listening is the expected score of the test taker on a randomly selected Listening section. Indeed, in some areas of the world, security concerns dictate that the test taker must in fact randomly receive one of several test forms administered in that location simultaneously. Assessment prediction errors involve assessment accuracy. Virtually no realistic case exists in which the true assessment score and the observed score are always the same. The Listening score reported does vary to some extent for different forms the test taker might take at a specific test administration. For a given construct, a loss of scoring accuracy will reduce assessment accuracy, which, in turn, will reduce the validity of the observed score as an indicator of the construct to be measured. It should be emphasized that no assessment can be satisfactory if scoring accuracy is inadequate, whereas very accurate scoring does not imply that a test is satisfactory for its intended purpose. In some sense, scoring accuracy is a necessary but not a sufficient condition for a good assessment. The Assessment Accuracy section examines application of assessment accuracy to assessments.

In the Conclusions section, basic recommendations are provided concerning use of PBLP, and some consideration is given to issues that require significant research and/or software enhancement.

Best Linear Prediction and Penalized Best Linear Prediction

PBLP builds on BLP. In the case of BLP, a random vector $\mathbf{X}$ of finite dimension $K \geq 1$ includes the observed real observations X_k for $1 \leq k \leq K$. These observations can have widely varying interpretations—they can be task scores, various computer-generated features of task responses, item scores for selected responses, or section scores. It is assumed that these observations have finite variances and a positive-definite covariance matrix Cov($\mathbf{X}$) with row j and column k denoted by $\mathrm{Cov}(X_j, X_k)$ for positive integers j and k no greater than K. This assumption implies that the variance $\sigma^2(X_k) = \mathrm{Cov}(X_k, X_k)$ is positive. More generally, the variance of $\mathbf{c}'\mathbf{X} = \sum_{k=1}^{K} c_k X_k$ is positive for every K-dimensional vector $\mathbf{c}$ with elements c_k, $1 \leq k \leq K$, that are not all 0. For different applications involving the same assessment, more than one definition of $\mathbf{X}$ may be used.

Example 1. Consider a case with $K = 21$. For $1 \leq k \leq 20$, let X_k be the item score for item k, and let each such item score have a finite variance. Let X_{21} be the sum of the X_k, $1 \leq k \leq 20$. Then $\mathbf{X}$ does not have a positive-definite covariance matrix, for $\mathbf{c}'\ \mathbf{X}$ is identically 0 if $c_k = 1$ for $1 \leq k \leq 20$ and $c_{21} = -1$. This example does not imply that sums of item scores may not be considered. For example, each element X_k of $\mathbf{X}$ may itself be a function of the sum of the item scores for a separate section of an assessment. This situation is discussed further in Example 7 for the scaled scores for the TOEFL test. Here $K = 4$, X_1 is the Listening score, X_2 is the Reading score, X_3 is the Speaking score, and X_4 is the Writing score. Each scaled score for a section is a form-dependent nonlinear transformation of the sum of the item scores for that section. Each reported scaled score is an integer between 0 and 30.

True Scores and Measurement Errors

The observed vector $\mathbf{X}$ is the sum of an unobserved true vector $\boldsymbol{\tau}$ with elements τ_k with finite variances, $1 \leq k \leq K$, and an unobserved error vector ϵ with elements ϵ_k with finite variances, $1 \leq k \leq K$. Here τ_k is the true score of X_k, and $\epsilon_k = X_k - \tau_k$ is the error of measurement of X_k. It is assumed that $\boldsymbol{\tau}$ and ϵ are uncorrelated in the sense that the covariance $\mathrm{Cov}(\tau_j, \epsilon_k)$ of τ_j and ϵ_k is 0 for all positive integers j and k no greater than K. If the covariance matrix Cov $(\boldsymbol{\tau}, \epsilon)$ of the vector $\boldsymbol{\tau}$ of true scores and the vector ϵ of true scores is defined to be the $K \times K$ matrix with row j, $1 \leq j \leq K$, and column k, $1 \leq k \leq K$, equal to $\mathrm{Cov}(\tau_j, \epsilon_k)$, then it follows that Cov $(\boldsymbol{\tau}, \epsilon)$ is the $K \times K$ matrix $\mathbf{0}_{KK}$ with all elements 0. Because the covariance matrix Cov $(\epsilon, \boldsymbol{\tau})$ of the vector ϵ of measurement errors and the vector $\boldsymbol{\tau}$ of true scores is the transpose of Cov $(\boldsymbol{\tau}, \epsilon)$, Cov $(\epsilon, \boldsymbol{\tau})$ is also $\mathbf{0}_{KK}$. Unlike common treatments of classical test theory (Lord & Novick, 1968), it is not necessarily assumed that, for

$K > 1$, the error variables ϵ_j and ϵ_k, $1 \le j < k \le K$, are uncorrelated. Nonetheless, it is assumed that the expected value of ϵ_k is 0 for $1 \le k \le K$ so that the expectation $E(\tau_k)$ of τ_k is equal to the expectation $E(X_k)$ of X_k. In terms of random vectors, the expectation $E(\mathbf{X}) = E(\boldsymbol{\tau})$ and $E(\boldsymbol{\epsilon}) = \mathbf{0}_K$. Here $E(\mathbf{X})$ is the K-dimensional vector with element k equal to $E(X_k)$ for $1 \le k \le K$ and $\mathbf{0}_K$ is the K-dimensional vector with all elements 0. The assumptions on $\boldsymbol{\tau}$ and ϵ also imply that

$$\mathrm{Cov}(\mathbf{X}, \boldsymbol{\tau}) = \mathrm{Cov}(\boldsymbol{\tau}, \boldsymbol{\tau}) + \mathrm{Cov}(\epsilon, \boldsymbol{\tau}) = \mathrm{Cov}(\boldsymbol{\tau}), \tag{1}$$

$$\mathrm{Cov}(\mathbf{X}, \epsilon) = \mathrm{Cov}(\boldsymbol{\tau}, \epsilon) + \mathrm{Cov}(\epsilon, \epsilon) = \mathrm{Cov}(\epsilon), \tag{2}$$

$$\mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{X}, \boldsymbol{\tau} + \epsilon) = \mathrm{Cov}(\boldsymbol{\tau}) + \mathrm{Cov}(\epsilon). \tag{3}$$

True scores are used to distinguish between the observed measurement and the quantity measured. The issue arises in practice in all cases in which a measurement must be made. Consider a purchase of a five-pound sack of potatoes at a supermarket. Because whole potatoes cannot be sorted so that a collection of them weighs exactly five pounds, the actual weight of the sack varies. There is a true weight of the item, and there is a measured weight from a supermarket scale. There is also a measured weight from a scale in a physics laboratory. The latter is likely to be much more accurate. In addition, the supermarket usually has multiple scales that are likely to exhibit some variation in measurements for the same bag. A further problem is that, for two supermarket scales, the difference between the measured weights for one bag need not be the same as the difference between the measured weights for a different bag. Thus even in this simple scenario, measurement error is not an entirely straightforward matter.

This supermarket example illustrates two different concepts of true scores and measurement errors. Consider estimation of the average weight of the population of five-pound sacks at the supermarket by use of a random sample of $K = 10$ sacks. Each sack selected is weighed on a randomly selected scale at the supermarket and yields an observed measurement X_k for $1 \le k \le K$. In scoring accuracy, which is considered in the Scoring Accuracy section, the true score τ_k for sack k is the average measured weight of the specified sack weighed for different supermarket scales. The measurement error ϵ_k is the difference between the observed weight and the true score. In assessment accuracy, the subject of the section Assessment Accuracy, the true score is the average weight of a sack from the population of sacks as measured by a randomly selected scale. Here the true score τ_k is the same for all measured sacks, and the measurement error ϵ_k contains both a component for variation of average measured weights of five-pound sacks for different scales and a component for variation of measured weights given the same sack due to use of different scales. A simple variation would involve division of the potatoes into different types so that X_1 might involve a randomly chosen sack of russet potatoes, while X_2 might involve a randomly selected bag of red potatoes. In this case, τ_k would likely depend on the type of potato selected. In all cases, the true score is not the weight of the potato sack obtained by an extremely precise scale.

In educational assessment, the problem of observed and true scores is much more severe, for the quantity to be measured is usually more difficult to define than is the weight of a sack of potatoes. Language proficiency or knowledge of algebra is not nearly as clearly defined. As in the case of the sacks of potatoes, the concept of a true score is somewhat restricted. Scoring accuracy generally only applies to results for constructed responses, such as essays rated by human raters. It involves variations in test scores or item scores related to human judgment. Assessment accuracy generally involves variation in test results due to random selection of a test form from a pool of similar forms. This concept involves variation in performance on parallel tests. These concepts of accuracy, true scores, and measurement errors are based on means, variances, and covariance and not on specific models.

For a fixed vector $\mathbf{c}$ with elements c_k, $1 \le k \le K$, a linear combination $\nu = \mathbf{c}'\boldsymbol{\tau} = \sum_{k=1}^{K} c_k \tau_k$ is to be predicted. It is assumed that $\mathbf{c}$ is nonzero in the sense that, for some positive integer $k \le K$, $c_k \ne 0$. Here ν is the true score of the observed linear combination $O = \mathbf{c}'\,\mathbf{X}$, and $\eta = O - \nu$ is the error of measurement of O. The expectation

$$E(O) = \mathbf{c}' E(\mathbf{X}) = \sum_{j=1}^{K} c_k E\left(X_k\right) \tag{4}$$

of O is the same as the expectation

$$E(\nu) = \mathbf{c}' E(\boldsymbol{\tau}) = \sum_{k=1}^{K} c_k E\left(\tau_k\right) \tag{5}$$

of ν, so that η has expectation 0. The variance of η is

$$\sigma^2(\eta) = \mathbf{c}'\mathrm{Cov}(\epsilon)\,\mathbf{c} = \sum_{k=1}^{K}\sum_{k'=1}^{K} c_k c_{k'}\mathrm{Cov}\left(\epsilon_k, \epsilon_{k'}\right). \tag{6}$$

Similar formulas apply to the true score ν and the observation O. Thus

$$\sigma^2(\nu) = \mathbf{c}'\mathrm{Cov}(\tau)\,\mathbf{c} \tag{7}$$

and

$$\sigma^2(O) = \mathbf{c}'\mathrm{Cov}(\mathbf{X})\,\mathbf{c}. \tag{8}$$

By Equation 1, the covariance $\mathrm{Cov}(O,\nu)$ of the observation O and its true score ν is equal to the variance $\sigma^2(\nu)$ of ν. The reliability coefficient of O is

$$\rho^2(O) = \frac{\sigma^2(\nu)}{\sigma^2(O)} = 1 - \frac{\sigma^2(\eta)}{\sigma^2(O)}, \tag{9}$$

so that $\rho^2(O)$ is nonnegative and does not exceed 1. A simple special case sometimes of interest has $\mathbf{c} = \boldsymbol{\delta}_k$ for some positive integer $k \le K$, where $\boldsymbol{\delta}_k$ has element k equal to 1 and all other elements equal to 0. In this case, the true score τ_k of the observation X_k is predicted.

Example 2. Kelley (1923) considered the case of $K=1$ and $c_1=1$. Here the predicted variable $\nu = \mathbf{c}'\boldsymbol{\tau}$ is the true score τ_1 of the observed score $O=X_1$. It is assumed that X_1 has a positive variance $\sigma^2(O)=\sigma^2(X_1)$. The error of measurement $\eta=\epsilon_1$ has expectation 0 and variance $\sigma^2(\eta)=\sigma^2(\epsilon_1)$. The variance $\sigma^2(\nu)=\sigma^2(\tau_1)$, and $\sigma^2(O)=\sigma^2(\nu)+\sigma^2(\eta)$. The covariance $\mathrm{Cov}(O,\nu)=\mathrm{Cov}(X_1,\tau_1)=\sigma^2(\nu)=\sigma^2(\tau_1)$.

Example 3. In TOEFL Writing, one case that has been considered with BLP uses one human score per prompt and nine computer-generated feature scores per prompt (Yao et al., 2019a, 2019b). Let $K=20$, let X_1 be the human score for the first prompt (the integrated task), let X_2 be the human score for the second prompt (the independent task), let X_3 to X_{11} be the feature scores for the first prompt, and let X_{12} to X_{20} be the feature scores for the second prompt. Let $\mathbf{c}$ have elements $c_1=c_2=1$, and let all other elements be 0. Then ν is the true score $\tau_1+\tau_2$ of the sum score $O=X_1+X_2$, and $\eta=\epsilon_1+\epsilon_2$. In this case, a raw score is considered for the assessment based on the human scores for the two prompts. The human observed score X_1 and true score τ_1 for the integrated task correspond to $\mathbf{c}=\boldsymbol{\delta}_1$, while the human score X_2 and true score τ_2 for the independent task correspond to $\mathbf{c}=\boldsymbol{\delta}_2$. A variation on this case uses $K=4$. The definitions of X_1 and X_2 remain unchanged, O is still X_1+X_2, ν is still $\tau_1+\tau_2$, and η is still $\epsilon_1+\epsilon_2$, but X_3 is a score on the integrated task from e-rater, and X_4 is a score on the independent task from e-rater. The score X_3 is normally a linear combination of computer-generated features from the first response, and X_4 is normally a linear combination of computer-generated features from the second response. In current practice, the true score ν is approximated by $\frac{1}{3}\left(2X_1+X_3\right)+\frac{1}{2}\left(X_2+X_4\right)$ in the case of $K=4$. This practice involves an approximation based on an early use of BLP (Haberman, 2011). In the section Best Linear Prediction, methods are provided to evaluate the impact of using this approximation rather than BLP. Other variations exist based on changes of e-rater features over time. For instance, current practice uses 11 features for the integrated task and 10 for the independent task.

Example 4. In GRE Writing, a case that has been examined with BLP uses one human score per prompt and nine computer-generated feature scores per prompt (Yao et al., 2019a, 2019b). As in Example 3, let $K=20$, let X_1 be the human score for the first prompt (the issue prompt), let X_2 be the human score for the second prompt (the argument prompt), let X_3 to X_{11} be the feature scores for the first prompt, and let X_{12} to X_{20} be the feature scores for the second prompt. Let $\mathbf{c}$ have elements $c_1=c_2=0.5$, and let all other elements be 0. Then ν is the average true score $(\tau_1+\tau_2)/2$ of the average sum score $O=(X_1+X_2)/2$, and $\eta=(\epsilon_1+\epsilon_2)/2$. In this case, a raw score is considered for the assessment based on the human scores for the two prompts. The human observed score X_1 and true score τ_1 for the issue prompt correspond to $\mathbf{c}=\boldsymbol{\delta}_1$, while the human score X_2 and true score τ_2 for the argument prompt correspond to $\mathbf{c}=\boldsymbol{\delta}_2$. As in Example 3, an alternative approach uses $K=4$ with X_1, X_2, O, ν, and η unchanged but with X_3 the e-rater score for the issue prompt and X_4 the e-rater score for the argument prompt. In GRE Writing, the current practice for prediction of ν involves a somewhat rounded approximation based on the formulation with $K=4$ uses $(X_1+X_2+X_3+X_4)/4$. In addition, current practice uses 10 feature scores for each prompt to find the e-rater score.

Example 5. In TOEFL Speaking, several rather distinct approaches may be considered for the four Speaking prompts currently in use. The approach most similar to that in Examples 3 and 10 uses 1 human score and 28 computer-generated feature scores for each prompt. Here $K = 116$. For prompt j, $1 \leq j \leq 4$, X_j is the human score and $X_{4+k+28(j-1)}$ is the kth feature score for $1 \leq k \leq 28$. The variable X_7 is the third feature score for the first prompt, while X_{39} is the seventh feature score for the second prompt. With this approach, $\mathbf{c}$ has $c_j = 1$ for $1 \leq j \leq 4$, and all other elements of $\mathbf{c}$ are 0, so that the observed score $O = X_1 + X_2 + X_3 + X_4$, $\nu = \tau_1 + \tau_2 + \tau_3 + \tau_4$, and $\eta = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$. For prompt j, $1 \leq j \leq 4$, the observed score X_j and true score τ_j correspond to $\mathbf{c} = \boldsymbol{\delta}_j$. Variations on this approach will be considered in this report because of the relatively high dimension of the observation vector and because symmetry considerations appear to apply relatively well in this case. One simple case uses sums over the four items to reduce the dimension K to 29. In this case, X_1 is the sum of the four human scores for the four prompts, and X_{k+1} is the sum of the kth feature scores for the four prompts. Thus $\mathbf{c} = \boldsymbol{\delta}_1$ so that $O = X_1$, $\nu = \tau_1$, and $\eta = \epsilon_1$. Note that the value of O has not changed due to the change in formulation. A further alternative formulation has $K = 2$, X_1 the sum of the four human scores, and X_2 the sum of the four SpeechRater scores. In current practice, an approximation for the case of $K = 2$ is used based on the PBLP approach for the case of $K = 2$ described in the section Penalized Best Linear Prediction. For a study of BLP for a previous version of TOEFL Speaking with six prompts, see Zhang et al. (2019).

Example 6. In PRAXIS Writing, consider 1 human score and 10 computer-generated features for the argumentative prompt, 1 human score and 11 features for the informative/explanatory prompt, and the sum of the selected-response item scores. In this case, $K = 24$, X_1 is the human score for the argumentative prompt, X_2 is the human score for the informative/explanatory prompt, X_3 is the sum of the item scores for the selected responses, X_4 to X_{13} are feature scores for the first prompt, and X_{14} to X_{24} are feature scores for the second prompt. The vector $\mathbf{c}$ with $c_1 = c_2 = 1.5$, $c_3 = 1$, and $c_k = 0$ for $4 \leq k \leq 24$ is used to generate a composite raw score. Thus $O = 1.5(X_1 + X_2) + X_3$, $\nu = 1.5(\tau_1 + \tau_2) + \tau_3$, and $\eta = 1.5(\epsilon_1 + \epsilon_2) + \epsilon_3$. The case of $\mathbf{c} = \boldsymbol{\delta}_j$ corresponds to X_j and τ_j for $1 \leq j \leq 3$. Current reporting practice does provide the sum X_3 of the scores for the selected responses, but no scores are provided for individual essay responses. A raw score is provided based on $\mathbf{c}$ with $c_1 = c_2 = 1$ and $c_k = 0$ for $3 \leq k \leq K$. Thus $O = X_1 + X_2$ is the sum of the human essay scores, and $\nu = \tau_1 + \tau_2$ is the corresponding true score. An alternative uses e-rater scores rather than individual feature scores. Here $K = 5$, X_1, X_2, and X_3 are defined as before, but X_4 is the e-rater score for the first prompt, and X_5 is the e-rater score for the second prompt. Once again, variations can occur due to changes in e-rater features. Because the data for Yao et al. (2019a, 2019b) were obtained before PRAXIS Writing used e-rater for the informative/explanatory prompt, only a rather restricted analysis was made for the argumentative prompt. In addition, that analysis only used nine essay features.

Example 7. The total TOEFL score illustrates a composite score based on linked scaled scores. Let $K = 4$. As discussed in Example 1, let X_1 be the scaled Listening score, let X_2 be the scaled Reading score, let X_3 be the scaled Speaking score, and let X_4 be the scaled Writing score. Each of these scores is a real number between 0 and 30 when unrounded, obtained by a nonlinear transformation of a section raw score, where the transformation is specific to the test form and administration. Let $\mathbf{c} = \mathbf{1}_4$ be the four-dimensional vector with all elements equal to 1. The total unrounded score is $O = \mathbf{c}'\mathbf{X} = X_1 + X_2 + X_3 + X_4$, $\eta = \epsilon_1 + \epsilon_2 + \epsilon_3 + \epsilon_4$, and $\nu = \tau_1 + \tau_2 + \tau_3 + \tau_4$. Note that the X_4 in this example is a function of the O in Example 3, while X_3 is a function of the O of Example 5. In this example, the scale score for Listening corresponds to $\mathbf{c} = \boldsymbol{\delta}_1$, the scale score for Reading corresponds to $\mathbf{c} = \boldsymbol{\delta}_2$, the scale score for Speaking corresponds to $\mathbf{c} = \boldsymbol{\delta}_3$, and the scale score for Writing corresponds to $\mathbf{c} = \boldsymbol{\delta}_4$.

Best Linear Prediction

In BLP, the linear predictor $\hat{\nu}$ of ν based on the observed vector $\mathbf{X}$ is selected to minimize mean squared error. Thus the real constant α and K-dimensional vector $\boldsymbol{\beta}$ with elements β_k, $1 \leq k \leq K$, are found such that for the linear predictor

$$\hat{\nu} = \alpha + \boldsymbol{\beta}'\mathbf{X} = \alpha + \sum_{k=1}^{K} \beta_k X_k, \tag{10}$$

and error of prediction $\hat{\eta} = \nu - \hat{\nu}$, the mean squared error

$$\text{MSE}\left(\nu, \hat{\nu}\right) = E\left(\hat{\eta}^2\right) \leq \text{MSE}\left(\nu, Y\right) = E\left(\left[\nu - Y\right]^2\right) \tag{11}$$

for any linear predictor

$$Y = a + \mathbf{b}'\mathbf{X} = a + \sum_{k=1}^{K} b_k X_k \tag{12}$$

of ν based on $\mathbf{X}$ determined by the real number a and the K-dimensional vector $\mathbf{b}$ with elements b_k for $1 \le k \le K$. The random variable $\hat{\nu}$ is the best linear predictor of ν, and $\hat{\eta} = \nu - \hat{\nu}$ is the corresponding prediction error.

Regression Equations

Use of standard regression theory shows that

$$\mathrm{Cov}\,(\mathbf{X})\,\boldsymbol{\beta} = \mathrm{Cov}\,(\mathbf{X}, \nu) = \mathrm{Cov}\,(\boldsymbol{\tau})\,\mathbf{c}, \tag{13}$$

and

$$\alpha = E\,(\nu) - \boldsymbol{\beta}' E\,(\mathrm{X})\,. \tag{14}$$

Here

$$\mathrm{Cov}\,(\mathbf{X}, \nu) = \mathrm{Cov}\,(\mathbf{X}, \boldsymbol{\tau})\,\mathbf{c} = \mathrm{Cov}\,(\boldsymbol{\tau})\,\mathbf{c}, \tag{15}$$

is the K-dimensional vector with element k, $1 \le k \le K$, equal to the covariance of X_k and ν.

The mean squared error MSE $\left(\nu, \hat{\nu}\right)$ for prediction of the true score ν by the BLP $\hat{\nu}$ is only equal to the mean squared error MSE(ν, Y) for prediction of the true score ν by the linear predictor Y if $a = \alpha$, $\mathbf{b} = \boldsymbol{\beta}$, and $Y = \hat{\nu}$.

Properties of the Best Linear Predictor

The BLP $\hat{\nu}$ has the basic properties

$$\hat{\nu} = E\,(\nu) + \boldsymbol{\beta}'\,[\mathrm{X} - E\,(\mathrm{X})]\,, \tag{16}$$

and has expectation $E(\nu) = E(O)$ and variance $\sigma^2\left(\hat{\nu}\right) = \boldsymbol{\beta}'\mathrm{Cov}\,(\mathrm{X})\,\boldsymbol{\beta}$.

By Equation 13, the covariance of $\mathbf{X}$ and $\hat{\nu}$ is

$$\mathrm{Cov}\left(\mathrm{X}, \hat{\nu}\right) = \mathrm{Cov}\,(\mathrm{X})\,\boldsymbol{\beta} = \mathrm{Cov}\,(\boldsymbol{\tau})\,\mathbf{c}, \tag{17}$$

so that

$$\mathrm{Cov}\left(\mathbf{X}, \hat{\eta}\right) = \mathrm{Cov}\,(\mathbf{X}, \nu) - \mathrm{Cov}\left(\mathbf{X}, \hat{\nu}\right) = \mathbf{0}_K. \tag{18}$$

Thus the prediction error $\hat{\eta} = \nu - \hat{\nu}$ and the predicting vector $\mathbf{X}$ are uncorrelated.

If

$$\hat{\boldsymbol{\tau}} = E\,(\mathrm{X}) + \mathrm{Cov}\,(\boldsymbol{\tau})\,[\mathrm{Cov}\,(\mathrm{X})]^{-1}\,[\mathbf{X} - E\,(\mathbf{X})] \tag{19}$$

has elements $\hat{\tau}_k$, $1 \le k \le K$, and $\hat{\epsilon} = \boldsymbol{\tau} - \hat{\boldsymbol{\tau}}$ has elements $\hat{\epsilon}_k$ for $1 \le k \le K$, then $\hat{\nu} = \mathbf{c}'\hat{\boldsymbol{\tau}}$. The best linear predictor of $\tau_k = \boldsymbol{\delta}_k'\boldsymbol{\tau}$ is $\hat{\tau}_k$. The prediction error $\hat{\epsilon}_k = \tau_k - \hat{\tau}_k = \boldsymbol{\delta}_k'\hat{\epsilon}$.

Proportional Reduction in Mean Squared Error

The proportional reduction in mean squared error provides a measure of the relative value of the linear predictor Y of the true score ν in cases in which the true score ν has a positive variance $\sigma^2(\nu)$. The proportional reduction in mean squared error of Y is then

$$\mathrm{PRMSE}\,(\nu, Y) = 1 - \frac{\mathrm{MSE}\,(\nu, Y)}{\sigma^2\,(\nu)}. \tag{20}$$

Here the variance $\sigma^2(\nu)$ of the true score ν is the smallest mean squared error achievable by use of a constant linear predictor $z = z + \mathbf{0}_K'\,\mathbf{X}$ of ν always equal to the real number z, for $\mathrm{MSE}(\nu, z) = E([\nu - z]^2) = \sigma^2(\nu) + [E(\nu) - z]^2$ exceeds $\sigma^2(\nu)$ unless $z = E(\nu) = E(O)$. Thus PRMSE(ν, Y) measures the relative improvement in mean squared error from use of Y to predict ν rather than use of a constant predictor. The measure PRMSE(ν, Y) is never greater than 1, and it equals 1 if, and only if, $Y = \nu$ with probability 1. The inequality $\mathrm{MSE}\,(\nu, Y) \ge \mathrm{MSE}\left(\nu, \hat{\nu}\right)$, with equality only if $Y = \hat{\nu}$, implies that

$$\mathrm{PRMSE}\left(\nu, \hat{\nu}\right) \ge \mathrm{PRMSE}\,(\nu, Y)\,, \tag{21}$$

with equality only if $Y = \hat{\nu}$. The special case of $Y = E(O)$ then implies that PRMSE $\left(\nu, \hat{\nu}\right)$ is nonnegative.

A decomposition of mean squared error leads to a chain rule for proportional reduction in mean squared error. Because the true score ν and its approximation $\hat{\nu}$ have the same expectation, the mean squared error MSE $\left(\nu, \hat{\nu}\right)$ is the variance $\sigma^2\left(\hat{\eta}\right)$ of the prediction error $\hat{\eta} = \nu - \hat{\nu}$. Because $\hat{\nu}$ is a linear function of $\mathbf{X}$ and Cov $\left(\mathbf{X}, \hat{\eta}\right) = \mathbf{0}_K$, $\hat{\eta}$ and $\hat{\nu}$ are uncorrelated so that the variance $\sigma^2(\nu)$ is the sum of the variance $\sigma^2\left(\hat{\eta}\right)$ of the prediction error and the variance $\sigma^2\left(\hat{\nu}\right)$ of the best linear predictor $\hat{\nu}$. It follows that

$$\text{MSE}\left(\nu, \hat{\nu}\right) = \sigma^2\left(\hat{\eta}\right) = \sigma^2(\nu) - \sigma^2\left(\hat{\nu}\right) = \sigma^2(\nu) - \boldsymbol{\beta}'\text{Cov}(\mathbf{X})\,\boldsymbol{\beta}. \tag{22}$$

If the true score ν has a positive variance $\sigma^2(\nu)$, then the proportional reduction in mean squared error relative to the constant predictor $E(\nu) = E(O)$ of ν is

$$\text{PRMSE}\left(\nu, \hat{\nu}\right) = 1 - \frac{\text{MSE}\left(\nu, \hat{\nu}\right)}{\sigma^2(\nu)} = 1 - \frac{\sigma^2\left(\hat{\eta}\right)}{\sigma^2(\nu)}. \tag{23}$$

Because $\sigma^2\left(\hat{\eta}\right) = \sigma^2(\nu) - \sigma^2\left(\hat{\nu}\right)$,

$$\frac{\sigma^2\left(\hat{\eta}\right)}{\sigma^2(\nu)} = 1 - \frac{\sigma^2\left(\hat{\nu}\right)}{\sigma^2(\nu)} \tag{24}$$

$$\text{PRMSE}\left(\nu, \hat{\nu}\right) = \frac{\sigma^2\left(\hat{\nu}\right)}{\sigma^2(\nu)}. \tag{25}$$

In Equation 11, because $\hat{\eta} = \nu - \hat{\nu}$ is uncorrelated with $\mathbf{X}$, $\hat{\eta}$ is uncorrelated with $\hat{\nu} - Y$. Thus the following decomposition of mean squared error applies:

$$\text{MSE}(\nu, Y) = \text{MSE}\left(\nu, \hat{\nu}\right) + \text{MSE}\left(\hat{\nu}, Y\right) = \sigma^2\left(\hat{\eta}\right) + \text{MSE}\left(\hat{\nu}, Y\right). \tag{26}$$

Thus the increase in mean squared error from prediction of ν by Y rather than the BLP $\hat{\nu}$ is the mean squared error from prediction of the BLP by Y.

Equation 26 also leads to inequalities and decompositions for the proportional reduction in error. If the variance $\sigma^2\left(\hat{\nu}\right)$ of the BLP $\hat{\nu}$ is positive, then division by $\sigma^2(\nu)$ shows that

$$\begin{aligned} 1 - \text{PRMSE}(\nu, Y) &= 1 - \text{PRMSE}\left(\nu, \hat{\nu}\right) + \left[1 - \text{PRMSE}\left(\hat{\nu}, Y\right)\right]\text{PRMSE}\left(\nu, \hat{\nu}\right) \\ &= 1 - \text{PRMSE}\left(\nu, \hat{\nu}\right)\text{PRMSE}\left(\hat{\nu}, Y\right), \end{aligned} \tag{27}$$

so that

$$\text{PRMSE}(\nu, Y) = \text{PRMSE}\left(\nu, \hat{\nu}\right)\text{PRMSE}\left(\hat{\nu}, Y\right). \tag{28}$$

Prediction by Observed Score

An important special case involves use of the observed score O as an unbiased estimate of its true score ν. Here

$$\text{MSE}(\nu, O) = \sigma^2(\eta) = \sigma^2(O)\left[1 - \rho^2(O)\right]. \tag{29}$$

If the reliability $\rho^2(O)$ is positive, then

$$\text{PRMSE}(\nu, O) = 1 - \frac{\sigma^2(\eta)}{\sigma^2(\nu)} = 2 - \frac{1}{\rho^2(O)}, \tag{30}$$

so that PRMSE(ν, O) is negative if $2\rho^2(O) < 1$ and PRMSE $(\nu, O) <$ PRMSE $\left(\nu, \hat{\nu}\right)$ if $\mathbf{c}$ is unequal to $\boldsymbol{\beta}$. The only possible way for $\mathbf{c}$ to equal $\boldsymbol{\beta}$ is for Cov$(\boldsymbol{\epsilon})\mathbf{c}$ to be $\mathbf{0}_K$ so that η is 0 with probability 1 and $\rho^2(O) = 1$.

Example 8. In Example 2, $\sigma^2(X_1)\beta_1 = \sigma^2(\tau_1)$ so that

$$\beta_1 = \frac{\sigma^2\left(\tau_1\right)}{\sigma^2\left(X_1\right)} = \rho^2\left(X_1\right), \tag{31}$$

$$\alpha = \left[1 - \rho^2\left(X_1\right)\right] E\left(X_1\right), \tag{32}$$

$$\hat{\tau}_1 = E\left(X_1\right) + \rho^2\left(X_1\right)\left[X_1 - E\left(X_1\right)\right], \tag{33}$$

$$\mathrm{MSE}\left(\tau_1, \hat{\tau}_1\right) = \sigma^2\left(\tau_1\right)\left[1 - \rho^2\left(X_1\right)\right], \tag{34}$$

and $\mathrm{PRMSE}\left(\tau_1, \hat{\tau}_1\right) = \rho^2\left(X_1\right)$ if $\sigma^2(\tau_1) > 0$. For comparison, $\mathrm{PRMSE}(\tau_1, X_1) = 2 - 1/\rho^2(X_1)$. The only way for $\mathrm{PRMSE}\left(\tau_1, \hat{\tau}_1\right)$ to equal $\mathrm{PRMSE}(\tau_1, X_1)$ is for $\sigma^2(\epsilon_1)$ to be 0, so that $\rho^2(X_1) = 1$. These results correspond to Kelley's (1923) formula.

In other words, for either the criterion of scoring accuracy based on selection of raters at random or the criterion of assessment accuracy based on selection of one test form out of a collection of similar forms, if the variance of the error of measurement is not 0, then the proportional reduction in mean squared error associated with using the observed score X_1 as an estimate of its expected score τ_1 is smaller than the proportional reduction in mean squared error associated with the BLP for that expected score. In fact, if the reliability of the observed score drops below .5, the proportional reduction in mean squared error associated with that observed score will be negative because its mean squared error is larger than that associated with using the average score $E(X_1)$.

A fundamental aspect of BLP is that the intuitive approximation of a true score ν by an observed score O may not be very efficient. This issue is most important if the observed score is not very reliable and the true scores τ_k, $1 \le k \le K$, are highly correlated. This problem is quite prevalent when subscores based on only a few items are used in an assessment. This point is emphasized in the references at the beginning of the introduction.

A second aspect of BLP that also must be emphasized is that in cases in which a linear approximation $Y = a + \mathbf{b}'\mathbf{X}$ of the true score ν is used rather than the intuitive approximation O, the performance of the approximation Y should be evaluated in terms of the proportional reduction $\mathrm{PRMSE}(\nu, Y)$ of mean squared error and not in terms of the reliability $\rho^2(Y) = \mathrm{PRMSE}(\xi, Y)$, where $\xi = a + \mathbf{b}'\,\mathbf{X}$. Unless Y is the BLP $\hat{\nu}$ of ν, a bias results because ν and ξ are different random variables. This issue arises, for example, in the evaluation of TOEFL Writing in Example 3 by use of $Y = \frac{1}{3}\left(2X_1 + X_3\right) + \frac{1}{2}\left(X_2 + X_4\right)$ in the case of $K = 4$. Here $\nu = (\tau_1 + \tau_2)/2$ and $\xi = \frac{1}{3}\left(2\tau_1 + \tau_3\right) + \frac{1}{2}\left(\tau_2 + \tau_4\right)$. This difference between proportional reduction of mean squared error and reliability reflects a much more general issue. If f is a real number and g is a positive real number, then reliability is scale invariant in the sense that $\rho^2(Y) = \rho^2(f + gY)$, and proportional reduction of mean squared error is scale invariant in the sense that $\mathrm{PRMSE}\ (f + g\nu, f + gY) = \mathrm{PRMSE}(\nu, Y)$. Nonetheless, it is not generally true that $\mathrm{PRMSE}(\nu, f + gY) = \mathrm{PRMSE}(\nu, Y)$ when either f is not 0 or g is not 1.

Penalized Best Linear Prediction

PBLP is a generalization of BLP normally employed to reduce subgroup bias. Such bias can unfairly advantage or disadvantage a distinct group of test takers. Here an additional polytomous random variable G is observed with positive integer values $h \le H$ for an integer $H > 1$. For any positive integer $h \le H$, the probability $P(G = h)$ is positive that $G = h$. The choice of variable depends on the specific assessment involved and the population that takes the assessment. In TOEFL iBT, native language, native country, and current country of residence are important, for performance on an assessment of English fluency can be affected by a native country's quality of instruction in the English language, the extent to which students are trained to take standardized tests, the similarity of English to the native language, and the extent of exposure to spoken English. The issue is not simply a matter of general fluency. For example, countries differ in their relative emphases on speaking English and their relative emphases on reading English. In many analyses for TOEFL related to fairness or linking, a division of examinees into 19 groups has been employed in which region of test administration is the primary variable, but in cases in which the test is administered in an English-speaking country, native language is also used. In GRE General and in PRAXIS Core, the main concern has been disparate impact by race/ethnicity, and that variable has been employed to examine subgroup biases.

Analysis in PBLP considers the conditional distributions of the vector $\mathbf{X}$ of observed scores, the vector $\boldsymbol{\tau}$ of true scores, and the vector ϵ of measurement errors given the random variable G. Because $G = h$ with positive probability for positive integers $h \le H$, these conditional distributions are uniquely defined. For a random variable Z with a finite expectation, let $E(Z|G = h)$ denote the conditional expectation of Z given h. For a K-dimensional random vector $\mathbf{V}$ with elements V_k, $1 \le k \le K$, with finite expectations, let the conditional expectation $E(\mathbf{V}|G = h)$ of $\mathbf{V}$ given $G = h$ be the

K-dimensional vector with elements $E(V_k|G=h)$ for $1 \le k \le K$. The random variable $E(Z|G)$ then has value $E(Z|G=h)$ if $G=h$ for a positive integer $h \le H$, and the random vector $E(\mathbf{V}|G)$ has value $E(\mathbf{V}|G=h)$ if $G=h$ for a positive integer $h \le H$. The expectation of $E(Z|G)$ is $E(Z)$, and the expectation of $E(\mathbf{V}|G)$ is $E(\mathbf{V})$. PBLP involves the conditional expectations $E\left(\mathbf{X}|G\right)$, $E\left(\boldsymbol{\tau}|G\right)$, and $E\left(\epsilon|G\right) = E\left(\mathbf{X}|G\right) - E\left(\boldsymbol{\tau}|G\right)$. It is assumed that the conditional expected measurement error $E(\boldsymbol{\epsilon}|G=h) = \mathbf{0}_K$ for $1 \le h \le H$ so that $E\left(\mathbf{X}|G\right) = E\left(\boldsymbol{\tau}|G\right)$.

Conditional expectations lead to a standard decomposition of variances and covariance matrices. Let Z have finite variance. The covariance of $Z - E(Z|G)$ and $E(Z|G)$ is 0 so that

$$\sigma^2\left(Z\right) = \sigma^2\left(Z - E\left(Z|G\right)\right) + \sigma^2\left(E\left(Z|G\right)\right). \tag{35}$$

A similar result applies to random vectors. Let $\mathbf{V}$ have a finite covariance matrix. Then the covariance matrix $\mathrm{Cov}(\mathbf{V} - E(\mathbf{V}|G), E(\mathbf{V}|G))$ is the $K \times K$ matrix with all elements 0 so that

$$\mathrm{Cov}\left(\mathbf{V}\right) = \mathrm{Cov}\left(\mathbf{V} - E\left(\mathbf{V}|G\right)\right) + \mathrm{Cov}\left(E\left(\mathbf{V}|G\right)\right). \tag{36}$$

For use with PBLP, for any real $d > 1$, let

$$\begin{aligned}\sigma_d^2\left(Z\right) &= \sigma^2\left(Z - E\left(Z|G\right) + d^{1/2}E\left(Z|G\right)\right)\\ &= \sigma^2\left(Z - E\left(Z|G\right)\right) + d\sigma^2\left(E\left(Z|G\right)\right)\\ &= \sigma^2\left(Z\right) + \left(d-1\right)\sigma^2\left(E\left(Z|G\right)\right),\end{aligned} \tag{37}$$

and

$$\begin{aligned}\mathrm{Cov}_d\left(\mathbf{V}\right) &= \mathrm{Cov}\left(\mathbf{V} - E\left(\mathbf{V}|G\right) + d^{1/2}E\left(\mathbf{V}|G\right)\right)\\ &= \mathrm{Cov}\left(\mathbf{V} - E\left(\mathbf{V}|G\right)\right) + d\mathrm{Cov}\left(E\left(\mathbf{V}|G\right)\right)\\ &= \mathrm{Cov}\left(\mathbf{V}\right) + \left(d-1\right)\mathrm{Cov}(E\left(\mathbf{V}|G\right).\end{aligned} \tag{38}$$

PBLP seeks to address a possible challenge in BLP involving subgroup effects. The conditional expectation variable $E(O|G) = E(\nu|G)$ and the conditional expectation variable $E\left(\hat{\nu}|G\right)$ can be somewhat different even though $E(O) = E(\nu)$ and $E\left(\hat{\nu}\right)$ are the same. In PBLP, a penalty function is imposed to reduce this disparity. For a finite penalty multiplier $d \ge 1$, the penalty for a prediction of ν by the linear predictor Y of Equation 12 is $(d-1)\mathrm{MSE}(E(\nu|G), E(Y|G))$. By assumption, this penalty is equal to $(d-1)\mathrm{MSE}(E(O|G), E(Y|G))$. The penalized mean squared error for prediction of ν by Y is then

$$\begin{aligned}\mathrm{MSE}_d\left(\nu, Y\right) &= \mathrm{MSE}\left(\nu, Y\right) + \left(d-1\right)\mathrm{MSE}\left(E\left(O|G\right), E\left(Y|G\right)\right)\\ &= \mathrm{MSE}\left(\nu - E\left(\nu|G\right) + d^{1/2}E\left(\nu|G\right), Y - E\left(Y|G\right) + d^{1/2}E\left(Y|G\right)\right).\end{aligned} \tag{39}$$

The penalized best linear predictor $\hat{\nu}_d$ based on $\mathbf{X}$ is selected to minimize the penalized mean squared error. The real constant α_d and K-dimensional vector $\boldsymbol{\beta}_d$ with elements β_{kd}, $1 \le k \le K$, are selected so that $\hat{\nu}_d = \alpha_d + \boldsymbol{\beta}_d'\mathbf{X}$ satisfies

$$\mathrm{MSE}_d\left(\nu, \hat{\nu}_d\right) \le \mathrm{MSE}_d\left(\nu, Y\right) \tag{40}$$

for any linear predictor Y defined as in Equation 12. The case of $d = 1$ corresponds to no penalty so that $\hat{\nu}_1 = \hat{\nu}$ and $\mathrm{MSE}_1\left(\nu, \hat{\nu}_1\right) = \mathrm{MSE}\left(\nu, \hat{\nu}\right)$. In addition, if $E(\mathbf{X}|G=h) = E(\mathbf{X})$ for $1 \le h \le H$, then $\hat{\nu}_d = \hat{\nu}$ and $\mathrm{MSE}_d\left(\nu, \hat{\nu}_d\right) = \mathrm{MSE}\left(\nu, \hat{\nu}\right)$.

Regression Equations

Computation of α_d and $\boldsymbol{\beta}_d$ is very similar to computation of α and $\boldsymbol{\beta}$. The equations

$$\mathrm{Cov}_d\left(\mathbf{X}\right)\boldsymbol{\beta}_d = \mathrm{Cov}_d\left(\boldsymbol{\tau}\right)\mathbf{c} \tag{41}$$

and

$$\alpha_d = E\left(O\right) - \boldsymbol{\beta}_d'E\left(\mathbf{X}\right) \tag{42}$$

apply so that

$$\hat{\nu}_d = E\left(O\right) + \boldsymbol{\beta}_d'\left[\mathbf{X} - E\left(\mathbf{X}\right)\right], \tag{43}$$

has expectation $E(O)$ and variance $\sigma^2\left(\widehat{\nu}_d\right) = \boldsymbol{\beta}_d'\mathrm{Cov}\,(\mathbf{X})\,\boldsymbol{\beta}_d$. The prediction error is $\widehat{\eta}_d = \nu - \widehat{\nu}_d$.

If

$$\widehat{\boldsymbol{\tau}}_d = E\,(\mathbf{X}) + \mathrm{Cov}_d\,(\boldsymbol{\tau})\left[\mathrm{Cov}_d\,(\mathbf{X})\right]^{-1}[\mathbf{X} - E\,(\mathbf{X})] \tag{44}$$

has elements $\widehat{\tau}_{kd}$, $1 \le k \le K$, and $\widehat{\epsilon}_d$ has elements $\widehat{\epsilon}_{kd}$ for $1 \le k \le K$, then $\widehat{\nu} = \mathbf{c}'\widehat{\boldsymbol{\tau}}_d$ and $\widehat{\eta}_d = \mathbf{c}'\widehat{\epsilon}_d$.

Optimization of predictors by different criteria generally yields different results. Equation 26 leads to

$$\mathrm{MSE}\left(\nu, \widehat{\nu}_d\right) = \mathrm{MSE}\left(\nu, \widehat{\nu}\right) + \mathrm{MSE}\left(\widehat{\nu}, \widehat{\nu}_d\right), \tag{45}$$

so that, in terms of mean squared error, the PBLP $\widehat{\nu}_d$ is never a better predictor of the true score ν than is the BLP $\widehat{\nu}$, and $\mathrm{MSE}\left(\nu, \widehat{\nu}_d\right) > \mathrm{MSE}\left(\nu, \widehat{\nu}\right)$ unless $\boldsymbol{\beta}_d = \boldsymbol{\beta}$ and $\alpha_d = \alpha$. The same argument based on $\nu - E(\nu|G) + d^{1/2}E(\nu|G)$ and $\widehat{\nu}_d - E\left(\widehat{\nu}_d|G\right) + d^{1/2}E\left(\widehat{\nu}_d|G\right)$ leads to

$$\mathrm{MSE}_d\left(\nu, \widehat{\nu}\right) = \mathrm{MSE}_d\left(\nu, \widehat{\nu}_d\right) + \mathrm{MSE}_d\left(\widehat{\nu}_d, \widehat{\nu}\right), \tag{46}$$

so that the penalized mean squared error for prediction of the true score ν by the BLP $\widehat{\nu}$ is greater than the penalized mean squared error for prediction of ν by the PBLP $\widehat{\nu}_d$ unless $\boldsymbol{\beta}_d = \boldsymbol{\beta}$ and $\alpha_d = \alpha$.

Measuring Subgroup Bias

To further consider subgroup bias, let

$$\Delta_d\,(h) = E\,(O|G = h) - E\left(\widehat{\nu}_d|G = h\right) = \left[E\,(O|G = h) - E\,(O)\right] - \boldsymbol{\beta}_d'\left[E\,(\mathbf{X}|G = h) - E\,(\mathbf{X})\right] \tag{47}$$

for positive integers $h \le H$, and let $\Delta_d(G)$ be the random variable with value $\Delta_d(h)$ if $G = h$. Then $E(\Delta_d(G)) = 0$ so that $E([\Delta_d(G)]^2) = \sigma^2(\Delta_d(G))$. In the case of $d = 1$, $\Delta_1\,(G) = E\,(O|G) - E\left(\widehat{\nu}|G\right)$. In terms of penalized mean squared error,

$$\mathrm{MSE}_d\left(\nu, \widehat{\nu}_d\right) = \mathrm{MSE}\left(\nu, \widehat{\nu}_d\right) + (d - 1)\,\sigma^2\left(\Delta_d\,(G)\right), \tag{48}$$

while

$$\mathrm{MSE}_d\left(\nu, \widehat{\nu}\right) = \mathrm{MSE}\left(\nu, \widehat{\nu}\right) + (d - 1)\,\sigma^2\left(\Delta_1\,(G)\right). \tag{49}$$

The inequality restrictions on $\mathrm{MSE}_d\left(\nu, \widehat{\nu}\right)$, $\mathrm{MSE}_d\left(\nu, \widehat{\nu}_d\right)$, $\mathrm{MSE}\left(\nu, \widehat{\nu}\right)$, and $\mathrm{MSE}\left(\nu, \widehat{\nu}_d\right)$ imply that $\sigma^2(\Delta_1(G))$ is at least as large as $\sigma^2(\Delta_d(G))$, with strict inequality if $\sigma^2(\Delta_1(G))$ is not zero. This strict inequality occurs unless the BLP $\widehat{\nu}$ had no issue at all so that $E(O|G = h)$ was equal to $E\left(\widehat{\nu}|G = h\right)$ for $1 \le h \le H$.

In the case of prediction of ν by the constant $E(O) = E(\nu)$, the penalized mean squared error $\mathrm{MSE}_d(\nu, E(O))$ is $\sigma_d^2(\nu)$, and

$$\mathrm{MSE}_d\left(\nu, \widehat{\nu}_d\right) = \sigma_d^2\left(\eta_d\right) = \sigma_d^2\,(\nu) - \sigma_d^2\left(\widehat{\nu}_d\right). \tag{50}$$

If $\sigma_d^2\,(\nu)$ is positive, as is certainly true if the variance $\sigma^2(\nu)$ of the true score ν is positive, then the corresponding proportional reduction in mean squared error is

$$\mathrm{PRMSE}_d\left(\nu, \widehat{\nu}_d\right) = 1 - \frac{\sigma_d^2\left(\eta_d\right)}{\sigma_d^2\,(\nu)} = \frac{\sigma_d^2\left(\widehat{\nu}_d\right)}{\sigma_d^2\,(\nu)}. \tag{51}$$

The chain equation

$$\mathrm{PRMSE}_d\,(\nu, Y) = \mathrm{PRMSE}_d\left(\nu, \widehat{\nu}\right)\mathrm{PRMSE}_d\left(\widehat{\nu}, Y\right) \tag{52}$$

follows from Equation 28 if Y is defined as in Equation 12.

If the covariance matrix $\mathrm{Cov}(E(\mathbf{X}|G))$ of the conditional expectation $E(\mathbf{X}|G)$ is positive-definite, then, as d approaches ∞, $d\left(\mathbf{c} - \boldsymbol{\beta}_d\right)$, $d\alpha_d$, and $d\left(O - \widehat{\nu}_d\right)$ have finite limits so that $d^2\sigma^2(\Delta_d(G))$ has a finite limit, and $\mathrm{MSE}_d\left(\nu, \widehat{\nu}_d\right)$ and $\mathrm{MSE}\left(\nu, \widehat{\nu}_d\right)$ both approach $\sigma^2(O)[1 - \rho^2(O)]$. In addition, $\mathrm{PRMSE}_d\left(\nu, \widehat{\nu}_d\right)$ and $\mathrm{PRMSE}\left(\nu, \widehat{\nu}_d\right)$ both approach $2 - 1/\rho^2(O)$, which, as noted in Equation 30, is the PRMSE associated with use of the observed score O as an estimate of the true score ν. This result concerning the covariance matrix of $E(\mathbf{X}|G)$ can only hold if H, the number of possible values of G, is no greater than the dimension K of $\mathbf{X}$.

In practice, it is important to verify that PRMSE $\left(\nu, \widehat{\nu}_d\right)$ is relatively close to PRMSE $\left(\nu, \widehat{\nu}\right)$, although some reduction does occur unless $\boldsymbol{\beta}_d = \boldsymbol{\beta}$ and $\alpha_d = \alpha$, for PRMSE $\left(\nu, \widehat{\nu}_d\right) =$ PRMSE $\left(\nu, \widehat{\nu}\right)$ PRMSE $\left(\widehat{\nu}, \widehat{\nu}_d\right)$. It is also important to consider the size of $\sigma^2(\Delta_d(G))$ and to examine the individual differences $\Delta_d(h)$ for $1 \le h \le H$.

For some understanding of relative sizes of group effects in cases in which the variance $\sigma^2\left(\widehat{\nu}_d\right)$ of $\widehat{\nu}_d$ is positive, it can be helpful to apply a linear linking to $\widehat{\nu}_d$. Let $\sigma(O)$ be the standard deviation of O, and let $\sigma\left(\widehat{\nu}_d\right)$ be the standard deviation of $\widehat{\nu}_d$. Let

$$\widehat{\nu}_{ds} = E(O) + \frac{\sigma(O)}{\sigma\left(\widehat{\nu}_d\right)}\left[\widehat{\nu}_d - E(O)\right], \tag{53}$$

so that $\widehat{\nu}_{ds}$ has the same expectation and variance as O. Then, for cases in which $\boldsymbol{\beta}_d$ is not $\mathbf{0}_K$, the scaled measure

$$\Delta_{ds}(h) = E(O|G=h) - E\left(\widehat{\nu}_{ds}|G=h\right) = [E(O|G=h) - E(O) - \frac{\sigma(O)}{\sigma\left(\widehat{\nu}_d\right)}\left[E\left(\widehat{\nu}_d|G=h\right) - E(O)\right] \tag{54}$$

is also informative. This measure can help distinguish between general shrinkage toward the expectation $E(O)$ and other important effects of subgroups. If $\mathbf{c} = \boldsymbol{\delta}_k$ for a positive integer $k \le K$, then $\widehat{\nu}_{ds}$ can be denoted by $\widehat{\tau}_{kds}$.

It is reasonable to try a variety of choices of d to compare results. The ultimate choice of the penalty multiplier $d - 1$ depends on relative priorities between fairness and accuracy that depend on the assessment, its legal considerations, and the magnitude of the issue. It is possible that a satisfactory balance between these competing criteria is unachievable. For example, it may be the case that O is not reliable enough for a satisfactory assessment but modest values of d that yield more satisfactory proportional reductions in mean squared error lead to excessive subgroup differences.

Example 9. Examples 2 and 8 provide a simple illustration of issues for the very simple case in which the number K of elements of $\mathbf{X}$ is 1 and the true score τ_1 is to be predicted. In this case, the slope

$$\beta_{1d} = \frac{\sigma^2\left(X_1\right)\rho^2\left(X_1\right) + (d-1)\sigma^2\left(E\left(X_1|G\right)\right)}{\sigma^2\left(X_1\right) + (d-1)\sigma^2\left(E\left(X_1|G\right)\right)} \tag{55}$$

is a weighted linear combination of the reliability $\rho^2(X_1)$ of X_1 and the constant 1 and

$$\alpha_d = \left(1 - \beta_{1d}\right)E\left(X_1\right), \tag{56}$$

so that

$$\widehat{\tau}_{1d} = E\left(X_1\right) + \beta_{1d}\left[X_1 - E\left(X_1\right)\right] \tag{57}$$

has variance $\beta_{1d}^2\sigma^2\left(X_1\right)$. The value of β_{1d} is at least as large as the reliability $\beta_1 = \rho^2(X_1)$ of X_1, and the variance of the PBLP $\widehat{\tau}_{1d}$ is at least as large as the variance of the BLP $\widehat{\tau}_1$. If $E(X_1|G=h) = E(X_1)$ for $1 \le h \le H$, then $\beta_{1d} = \beta_1 = \rho^2(X_1)$, $\alpha_d = \alpha$, and $\widehat{\tau}_{1d} = \widehat{\tau}_1$. Otherwise, β_{1d} is nonnegative, greater than β_1, and no greater than 1, with $\beta_{1d} = 1$ if, and only if, $\rho^2(X_1) = 1$. As a function of d, β_{1d} is strictly increasing. As d approaches ∞, $d(1 - \beta_{1d})$, $d\alpha_d$, and $d\left[X_1 - \widehat{\nu}_d\right)]$ have finite limits. The variance $\sigma^2(\Delta_d(G))$ is $(1 - \beta_{1d})^2\sigma^2(E(X_1|G))$ so that the mean squared error

$$\text{MSE}\left(\tau_1, \widehat{\tau}_{1d}\right) = \sigma^2\left(\tau_1\right)\left[1 - \rho^2\left(X_1\right)\right] + \left[\beta_{1d} - \rho^2\left(X_1\right)\right]^2\sigma^2\left(X_1\right). \tag{58}$$

The value of $\text{MSE}_d\left(\tau_1, \widehat{\tau}_{1d}\right)$ is the sum of MSE $\left(\tau_1, \widehat{\tau}_{1d}\right)$ and $(d-1)(1-\beta_{1d})^2\sigma^2(E(X_1|G))$. As d approaches ∞, $(d-1)\sigma^2(\Delta_d(G))$ converges to 0, and MSE $\left(\tau_1, \widehat{\tau}_{1d}\right)$ converges to $\sigma^2(X_1)[1 - \rho^2(X_1)]$.

In this example, for $1 \le h \le H$, $\Delta_d(h) = (1 - \beta_{1d})[E(X_1|G=h) - E(X_1)]$ but $\Delta_{ds}(h) = 0$. The difference in results arises because the subgroup bias reflects only the normal reduction in variability of $\widehat{\tau}_{1d}$ relative to X_1.

PBLP is not simply a matter of scaling the observed O, as in Example 9, when the true score ν of the observed score O is approximated by $\widehat{\nu}_d$ and no constant multiplier f exists such that $\boldsymbol{\beta}_d = f\mathbf{c}$ so that the weight β_{kd} the PBLP assigns to variable X_k is not proportional for $1 \le k \le K$ to the weight c_k assigned to X_k in the case of the observed score O. In such cases, the use of information other than O to approximate the true score ν has potential consequences in terms of fairness that PBLP seeks to address by balancing accuracy and subgroup bias. In addition to the formal aspects of BLP and PBLP, it is always important to consider the relevance of each variable X_k to the construct of interest. Features used in e-rater and SpeechRater must be relevant on substantive grounds to the evaluation of writing or speaking and not merely predictive of human scores.

Estimation for Best Linear Prediction and Penalized Best Linear Prediction

Practical estimation for BLP and PBLP is challenging because the vectors $\boldsymbol{\tau}$ and ϵ are not observed so that a simple regression analysis cannot be employed. In the simplest cases, the pairs $\left(\boldsymbol{\tau}_i, \epsilon_i\right)$, $1 \le i \le n$, $n > K$, are mutually independent and have the same distribution as the pair $(\boldsymbol{\tau}, \epsilon)$. The vectors $\mathbf{X}_i = \boldsymbol{\tau}_i + \epsilon_i$ are observed for $1 \le i \le n$. For positive integers $k \le K$, element k of $\mathbf{X}_i$ is X_{ik}, element k of $\boldsymbol{\tau}_i$ is τ_{ik}, and element k of $\boldsymbol{\epsilon}_i$ is ϵ_{ik}. No particular difficulty exists as far as estimation of summary parameters dependent on $\mathbf{X}$. If bars are used for estimates of parameters, then the sample mean

$$\bar{E}(\mathbf{X}) = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i \tag{59}$$

provides an unbiased estimate of $E(\mathbf{X})$. For $1 \le k \le K$, element k of $\bar{E}(\mathbf{X})$ is $\bar{E}\left(X_k\right)$. If $O_i = \mathbf{c}' \mathbf{X}_i$ for $1 \le i \le n$, then the sample mean

$$\bar{E}(O) = n^{-1} \sum_{i=1}^{n} O_i \tag{60}$$

is an unbiased estimate of $E(O)$. For consistency with some of the analysis of assessment accuracy, it is convenient to use the biased covariance estimate

$$\overline{\mathrm{Cov}}(\mathbf{X}) = n^{-1} \sum_{i=1}^{n} \left[\mathbf{X}_i - \bar{E}(\mathbf{X})\right] \left[\mathbf{X}_i - \bar{E}(\mathbf{X})\right]' \tag{61}$$

for Cov($\mathbf{X}$) and the biased variance estimate

$$\bar{\sigma}^2(O) = n^{-1} \sum_{i=1}^{n} \left[O_i - \bar{E}(O)\right]^2 \tag{62}$$

for $\sigma^2(O)$. For $1 \le j \le K$ and $1 \le k \le K$, row j and column k of $\overline{\mathrm{Cov}}(\mathbf{X})$ is $\overline{\mathrm{Cov}}\left(X_j, X_k\right)$. For $1 \le h \le H$, $\bar{E}\left(\delta_h(G)\right)$ is an unbiased estimate of $P(G = h)$. Let $\bar{E}(\mathbf{X}|G = h)$ be $\bar{E}(\mathbf{X})$ if $\bar{E}\left(\delta_h(G)\right) = 0$, and let

$$\bar{E}(\mathbf{X}|G = h) = \frac{\bar{E}\left(\delta_h(G)\mathbf{X}\right)}{\bar{E}\left(\delta_h(G)\right)} \tag{63}$$

if $\bar{E}\left(\delta_h(G)\right) > 0$. Then $\bar{E}(\mathbf{X}|G = h)$ estimates $E(\mathbf{X}|G = h)$. Element k of $\bar{E}(\mathbf{X}|G = h)$ is $\bar{E}\left(X_k|G = h\right)$ for $1 \le k \le K$. Let $\bar{E}(O|G = h)$ be $\bar{E}(O)$ if $\bar{E}\left(\delta_h(G)\right) = 0$, and let

$$\bar{E}(O|G = h) = \frac{\bar{E}\left(\delta_h(G)O\right)}{\bar{E}\left(\delta_h(G)\right)} \tag{64}$$

if $\bar{E}\left(\delta_h(G)\right) > 0$. Then $\bar{E}(O|G = h)$ estimates $E(O|G = h)$.

Let

$$\overline{\mathrm{Cov}}(E(\mathbf{X}|G)) = \sum_{h=1}^{H} \bar{E}\left(\delta_h(G)\right) \left[\bar{E}(\mathbf{X}|G = h) - \bar{E}(\mathbf{X})\right] \left[\bar{E}(\mathbf{X}|G = h) - \bar{E}(\mathbf{X})\right]' \tag{65}$$

estimate Cov($E(\mathbf{X}|G)$), and let

$$\bar{\sigma}^2(E(O|G)) = \sum_{h=1}^{H} \bar{E}\left(\delta_h(G)\right) \left[\bar{E}(O|G = h) - \bar{E}(O)\right]^2 \tag{66}$$

estimate $\sigma^2(E(O|G))$. Let

$$\overline{\mathrm{Cov}}_d(\mathbf{X}|G) = \overline{\mathrm{Cov}}(\mathbf{X}) + (d - 1)\overline{\mathrm{Cov}}(E(\mathbf{X}|G)) \tag{67}$$

estimate Cov$_d$($E(\mathbf{X})$.

Further estimation requires far more difficult analysis because the observations $\mathbf{X}_i$, $1 \le i \le n$, do not generally provide an estimate of Cov$(\boldsymbol{\tau})$. In the section Scoring Accuracy, estimation procedures are explored for scoring accuracy. In the section Assessment Accuracy, estimation is studied for assessment accuracy. For the moment, it is worth noting that, given an estimate $\overline{\mathrm{Cov}}\,(\boldsymbol{\tau})$ of $\mathrm{Cov}\,(\boldsymbol{\tau})$, the remaining estimates are readily found. Let $\bar{\sigma}^2\,(\nu) = \mathbf{c}'\overline{\mathrm{Cov}}\,(\boldsymbol{\tau})\,\mathbf{c}$ estimate $\sigma^2(\nu)$. Let

$$\overline{\mathrm{Cov}}\,(\mathbf{X})\,\bar{\boldsymbol{\beta}} = \overline{\mathrm{Cov}}\,(\boldsymbol{\tau}) \tag{68}$$

determine the estimate $\bar{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$. If $\overline{\mathrm{Cov}}\,(\mathbf{X})$ is nonsingular, then the definition of $\bar{\boldsymbol{\beta}}$ is clear. In general, $\bar{\boldsymbol{\beta}}$ can be defined by selecting the solution of Equation 68 that minimizes $\left\|\bar{\boldsymbol{\beta}}\right\|^2 = \overline{\boldsymbol{\beta}'}\bar{\boldsymbol{\beta}}$ (Rao & Mitra, 1972). Let

$$\bar{\alpha} = \bar{E}\,(O) - \overline{\boldsymbol{\beta}'}\bar{E}\,(\mathbf{X}) \tag{69}$$

estimate α, and let

$$\bar{\nu} = \bar{\alpha} + \overline{\boldsymbol{\beta}'}\mathrm{X} = \bar{E}\,(O) + \overline{\boldsymbol{\beta}'}\left[\mathrm{X} - \bar{E}\,(\mathrm{X})\right] \tag{70}$$

estimate $\hat{\nu}$. Let

$$\bar{\sigma}^2\,\left(\hat{\nu}\right) = \overline{\boldsymbol{\beta}'}\overline{\mathrm{Cov}}\,(\mathrm{X})\,\bar{\boldsymbol{\beta}} \tag{71}$$

estimate $\sigma^2\,\left(\hat{\nu}\right)$. Let $\bar{\tau}_k = \bar{\nu}$ if $\mathbf{c} = \boldsymbol{\delta}_k$ for a positive integer $k \le K$, and let $\bar{\boldsymbol{\tau}}$ be the vector of dimension K with elements $\bar{\tau}_k$ for $1 \le k \le K$. In the typical case of $\overline{\mathrm{Cov}}\,(\mathbf{X})$ nonsingular,

$$\bar{\boldsymbol{\tau}} = \bar{E}\,(\mathbf{X}) + \overline{\mathrm{Cov}}\,(\boldsymbol{\tau})\left[\overline{\mathrm{Cov}}\,(\mathbf{X})\right]^{-1}\left[\mathbf{X} - \bar{E}\,(\mathbf{X})\right] \tag{72}$$

has elements $\bar{\tau}_k$ for $1 \le k \le K$. Thus $\bar{\boldsymbol{\tau}}$ estimates $\hat{\boldsymbol{\tau}}$, and $\bar{\tau}_k$ estimates $\hat{\tau}_k$.

Let

$$\overline{\mathrm{MSE}}\,\left(\nu, \hat{\nu}\right) = \bar{\sigma}^2\,(\nu) - \bar{\sigma}^2\,\left(\hat{\nu}\right) \tag{73}$$

estimate $\mathrm{MSE}\,\left(\nu, \hat{\nu}\right)$. If $\bar{\sigma}^2\,(\nu)$ is positive, let

$$\overline{\mathrm{PRMSE}}\,\left(\nu, \hat{\nu}\right) = 1 - \frac{\overline{\mathrm{MSE}}\,\left(\nu, \hat{\nu}\right)}{\bar{\sigma}^2\,(\nu)} = \frac{\bar{\sigma}^2\,\left(\hat{\nu}\right)}{\bar{\sigma}^2\,(\nu)} \tag{74}$$

estimate $\mathrm{PRMSE}\,\left(\nu, \hat{\nu}\right)$. More generally, if Y is defined as in Equation 12, then

$$\overline{\mathrm{MSE}}\,(\nu, Y) = \overline{\mathrm{MSE}}\,\left(\nu, \hat{\nu}\right) + \overline{\mathrm{MSE}}\,\left(\hat{\nu}, Y\right) \tag{75}$$

estimates $\mathrm{MSE}(\nu, Y)$, where

$$\overline{\mathrm{MSE}}\,\left(\hat{\nu}, Y\right) = \left[a - \bar{\alpha} - \left(\mathbf{b} - \bar{\boldsymbol{\beta}}\right)'\bar{E}\,(\mathrm{X})\right]^2 + \left(\mathbf{b} - \bar{\boldsymbol{\beta}}\right)'\overline{\mathrm{Cov}}\,(\mathbf{X})\left(\mathbf{b} - \bar{\boldsymbol{\beta}}\right). \tag{76}$$

If $\bar{\sigma}^2\,(\nu)$ is positive, let

$$\overline{\mathrm{PRMSE}}\,(\nu, Y) = 1 - \frac{\overline{\mathrm{MSE}}\,(\nu, Y)}{\bar{\sigma}^2\,(\nu)} \tag{77}$$

estimate $\mathrm{PRMSE}(\nu, Y)$.

Similar arguments apply to PBLP. Let

$$\overline{\mathrm{Cov}}_d\,(\boldsymbol{\tau}) = \overline{\mathrm{Cov}}\,(\boldsymbol{\tau}) + (d-1)\,\overline{\mathrm{Cov}}\,(E\,(\mathbf{X}|G)) \tag{78}$$

estimate $\mathrm{Cov}_d\,(\boldsymbol{\tau})$. Then

$$\overline{\mathrm{Cov}}_d\,(\mathbf{X})\,\bar{\boldsymbol{\beta}}_d = \overline{\mathrm{Cov}}_d\,(\boldsymbol{\tau})\,\mathbf{c} \tag{79}$$

estimates $\boldsymbol{\beta}_d$. If $\overline{\mathrm{Cov}}_d\,(\mathbf{X})$ is singular, then $\bar{\boldsymbol{\beta}}_d$ minimizes $\left\|\bar{\boldsymbol{\beta}}_d\right\|^2$ among solutions of Equation 79. Then

$$\bar{\alpha}_d = \bar{E}\,(O) - \overline{\boldsymbol{\beta}'_d}\bar{E}\,(\mathbf{X}) \tag{80}$$

estimates α_d and $\bar{\nu}_d = \bar{\alpha}_d + \overline{\boldsymbol{\beta}'_{\mathrm{X}}}$ estimates $\hat{\nu}_d$ so that

$$\bar{\nu}_d = \bar{E}\,(O) + \overline{\boldsymbol{\beta}'_d}\left[\mathrm{X} - \bar{E}\,(\mathrm{X})\right]. \tag{81}$$

Let $\overline{\sigma}^2\left(\hat{\nu}_d\right) = \overline{\boldsymbol{\beta}'_d}\,\overline{\mathrm{Cov}}\,(\mathbf{X})\,\overline{\boldsymbol{\beta}}_d$ estimate $\sigma^2\left(\hat{\nu}_d\right)$. Let $\overline{\tau}_{kd} = \overline{\nu}$ if $\mathbf{c} = \boldsymbol{\delta}_k$ for a positive integer $k \le K$, and let $\overline{\boldsymbol{\tau}}_d$ be the vector of dimension K with elements $\overline{\tau}_{kd}$ for $1 \le k \le K$. In the typical case of $\overline{\mathrm{Cov}}_d\,(\mathbf{X})$ nonsingular,

$$\overline{\boldsymbol{\tau}}_d = \overline{E}\,(\mathbf{X}) + \overline{\mathrm{Cov}}_d\,(\boldsymbol{\tau})\left[\overline{\mathrm{Cov}}_d\,(\mathbf{X})\right]^{-1}\left[\mathbf{X} - \overline{E}\,(\mathbf{X})\right] \tag{82}$$

has elements $\overline{\tau}_{kd}$ for $1 \le k \le K$. Thus $\overline{\boldsymbol{\tau}}_d$ estimates $\hat{\boldsymbol{\tau}}_d$, and $\overline{\tau}_{kd}$ estimates $\hat{\tau}_{kd}$.

Let

$$\overline{E}\left(\hat{\nu}_d|G = h\right) = \overline{E}\,(O) + \overline{\boldsymbol{\beta}'_d}\left[\overline{E}\,(\mathbf{X}|G = h) - \overline{E}\,(O)\right] \tag{83}$$

estimate $E\left(\hat{\nu}_d|G = h\right)$ for positive integers $h \le H$. Let

$$\overline{\sigma}^2\left(E\left(\hat{\nu}_d|G\right)\right) = \sum_{h=1}^{H} \overline{E}\left(\delta_h\,(G)\right)\left[\overline{E}\left(\hat{\nu}_d|G = h\right)\right]^2 \tag{84}$$

estimate $\sigma^2\left(E\left(\hat{\nu}_d|G\right)\right)$. Let

$$\overline{\Delta}_d\,(h) = \overline{E}\,(O|G = h) - \overline{E}\left(\hat{\nu}_d|G = h\right) \tag{85}$$

estimate $\Delta_d(h)$ for positive integers $h \le H$. Let

$$\overline{\sigma}^2\left(\Delta_d\,(G)\right) = \sum_{h=1}^{H} \overline{E}\left(\delta_h\,(G)\right)\left[\overline{\Delta}_d\,(h)\right]^2 \tag{86}$$

estimate $\sigma^2(\Delta_d(G))$.

Let

$$\overline{\sigma}^2_d\,(\nu) = \overline{\sigma}^2\,(\nu) + (d - 1)\,\overline{\sigma}^2\,(E\,(O|G)) \tag{87}$$

estimate $\sigma^2_d\,(\nu)$, and let

$$\overline{\mathrm{MSE}}_d\left(\nu, \hat{\nu}_d\right) = \left[\overline{\sigma}^2\,(\nu) - \overline{\sigma}^2\left(\hat{\nu}_d\right)\right] + (d - 1)\left[\overline{\sigma}^2\,(E\,(O|G)) - \overline{\sigma}^2\left(E\left(\hat{\nu}_d|G\right)\right)\right] \tag{88}$$

estimate $\mathrm{MSE}_d\left(\nu, \hat{\nu}_d\right)$. If Y is defined as in Equation 12, then

$$\overline{\mathrm{MSE}}_d\,(\nu, Y) = \overline{\mathrm{MSE}}_d\left(\nu, \hat{\nu}_d\right) + \overline{\mathrm{MSE}}_d\left(\hat{\nu}_d, Y\right) \tag{89}$$

estimates $\mathrm{MSE}_d(\nu, Y)$, where

$$\overline{\mathrm{MSE}}_d\left(\hat{\nu}_d, Y\right) = d\left[a - \overline{\alpha}_d - \left(\mathbf{b} - \overline{\boldsymbol{\beta}}_d\right)'\,\overline{E}\,(\mathbf{X})\right]^2 + \left(\mathbf{b} - \overline{\boldsymbol{\beta}}\right)'\,\overline{\mathrm{Cov}}_d\,(\mathbf{X})\left(\mathbf{b} - \overline{\boldsymbol{\beta}}\right). \tag{90}$$

If $\overline{\sigma}^2_d\,(\nu)$ is positive, let

$$\overline{\mathrm{PRMSE}}_d\left(\nu, \hat{\nu}_d\right) = \frac{\overline{\sigma}^2\left(\hat{\nu}_d\right) + (d - 1)\,\overline{\sigma}^2\left(E\left(\hat{\nu}_d|G\right)\right)}{\overline{\sigma}^2_d\,(\nu)} \tag{91}$$

estimate $\mathrm{PRMSE}_d\left(\nu, \hat{\nu}_d\right)$, and let

$$\overline{\mathrm{PRMSE}}_d\,(\nu, Y) = 1 - \frac{\overline{\mathrm{MSE}}_d\left(\hat{\nu}_d, Y\right)}{\overline{\sigma}^2_d\,(\nu)} \tag{92}$$

estimate $\mathrm{PRMSE}_d(\nu, Y)$.

Let $\overline{\sigma}\,(O)$, the nonnegative square root of $\overline{\sigma}^2\,(O)$, estimate $\sigma(O)$, and let $\overline{\sigma}\left(\hat{\nu}_d\right)$, the nonnegative square root of $\overline{\sigma}^2\left(\hat{\nu}_d\right)$, estimate $\sigma\left(\hat{\nu}_d\right)$. If $\overline{\sigma}\left(\hat{\nu}_d\right) > 0$, then estimate $\Delta_{ds}(h)$, $1 \le h \le H$, by

$$\overline{\Delta}_{ds}\,(h) = [\overline{E}\,(O|G = h) - \overline{E}\,(O) - \frac{\overline{\sigma}\,(O)}{\overline{\sigma}\left(\hat{\nu}_d\right)}\left[\overline{E}\left(\hat{\nu}_d|G = h\right) - \overline{E}\,(O)\right]. \tag{93}$$

Estimates proposed are intuitive estimates that are consistent for the quantities estimated. They can be expected to be increasingly accurate as sample sizes increase. The proposed estimates suffice for all analyses of BLP and PBLP as long as the assumption of simple random sampling is reasonable. In cases of simple random sampling, grouped jackknifing may be employed to estimate the accuracy of estimates. In practice, sampling may be much more complex so that sampling weights may be needed and accuracy of estimates may be reduced. Evaluation of accuracy of estimates is also affected. Thus users of the methodology for estimation should consider how sampling has been conducted.

Scoring Accuracy

The criterion of scoring accuracy involves the error in prediction of the expected score given the specific constructed responses in the assessment that require human scoring. Let J be a positive integer no greater than K, and let X_k, $1 \le k \le J$, be scores based on human scoring. In typical cases, for positive integers j and k no greater than K, $\mathrm{Cov}(\epsilon_j, \epsilon_k)$ is 0 unless $j = k \le K$, and $\sigma^2(\epsilon_k) > 0$ for $k \le K$. Thus the error of measurement ϵ_k has positive variance for X_k if $k \le J$, measurement errors ϵ_j and ϵ_k are uncorrelated if j and k are distinct positive integers no greater than J, and errors of measurement ϵ_k are 0 if k is a positive integer no greater than K but greater than J.

In the TOEFL, GRE, and PRAXIS assessments used as examples, the X_k, $k \le J$, are human holistic scores for either essay responses or spoken responses. The assumption made is that scoring errors on constructed responses do not correlate with scoring errors on other constructed responses. It is also assumed that scoring errors on selected responses do not exist due to automated scoring of such items. In the case of scoring accuracy, all that is required to estimate $\mathrm{Cov}(\epsilon)$ is to estimate $\sigma^2(\epsilon_k)$ for $k \le J$. Because $\mathrm{Cov}(\boldsymbol{\tau}) = \mathrm{Cov}(\mathbf{X}) - \mathrm{Cov}(\epsilon)$, the covariance matrix of $\boldsymbol{\tau}$ can then be estimated. Let $\overline{\sigma}^2\left(\epsilon_k\right)$ be the nonnegative estimate of $\sigma^2(\epsilon_k)$ for $1 \le k \le J$. Then the estimate $\overline{\mathrm{Cov}}(\epsilon)$ of $\mathrm{Cov}(\boldsymbol{\epsilon})$ is the diagonal matrix with row k and column k equal to 0 for any integer k such that $J < k \le K$ and equal to $\overline{\sigma}^2\left(\epsilon_k\right)$ if $1 \le k \le J$. It then follows that $\overline{\mathrm{Cov}}(\boldsymbol{\tau}) = \overline{\mathrm{Cov}}(\mathbf{X}) - \overline{\mathrm{Cov}}(\epsilon)$ estimates $\mathrm{Cov}(\boldsymbol{\tau})$.

To avoid an estimated covariance matrix that is not nonnegative definite can require a reduction in normal estimates of $\overline{\sigma}^2\left(\epsilon_k\right)$ for $1 \le k \le J$. Such a case has never been encountered in practice, but the issue is noted here for completeness. Such a case would in practice require changes in the sampling procedures used for estimation of variances of measurement errors for scoring or a major investigation into scoring quality for constructed responses.

Agreement Samples

Agreement samples arise when a constructed task response is normally scored by only one human rater. This situation arises in all examples considered in this report. In such a case, to estimate rater variability, a random sample of responses must be drawn, and responses in the sample must be scored by a second human rater. It is very important to emphasize the use of a random sample rather than a sample that may have been obtained by checking constructed responses with unusual discrepancies between the scores assigned by a human rater and predicted by computer-generated features.

The basic procedure is quite straightforward for a positive integer $k \le J$. A simple random subsample I_k of $n_k > 0$ positive integers no greater than n is drawn without replacement from the set $\overline{n}$ of positive integers no greater than n. In typical cases, n_k/n is relatively small, say, .05. For each i in I_k, a second human score X'_{ik} is observed. It is assumed that $\epsilon'_{ik} = X'_{ik} - \tau_{ik}$ has expectation 0 and that variance $\sigma^2(\epsilon_k)$ and ϵ'_{ik} and ϵ_{ik} are uncorrelated. The estimate of $\sigma^2(\epsilon_k)$ is then

$$\overline{\sigma}^2\left(\epsilon_k\right) = \left(2n_k\right)^{-1} \sum \left(X_{ik} - X'_{ik}\right)^2 . \tag{94}$$

For examples of application of agreement samples to scoring accuracy for the testing programs under study, see Yao et al. (2019a, 2019b) and Zhang et al. (2019).

Double Human Scoring

Although not currently encountered in the assessments examined, a traditional approach has each score X_k, $1 \le k \le J$, equal to the average of two human scores X'_k and X''_k such that $\epsilon'_k = X'_k - \tau_k$ and $\epsilon''_k = X''_k - \tau_k$ have expectation 0, are uncorrelated, and have common variance $2\sigma^2(\epsilon'_k)$, so that the measurement error $\epsilon_k = (\epsilon'_k + \epsilon''_k)/2$ has variance $\sigma^2(\epsilon'_k)/2$. For $1 \le i \le n$, X_{ik} is the average of X'_{ik} and X''_{ik}, where the pairs (X'_{ik}, X''_{ik}), $1 \le i \le n$, are mutually independent and have the same distribution as (X'_k, X''_k). In this case the estimate of $\sigma^2(\epsilon_k)$ is

$$\overline{\sigma}^2\left(\epsilon_k\right) = (4n)^{-1} \sum_{i=1}^{n} \left(X''_{ik} - X'_{ik}\right)^2 . \tag{95}$$

Agreement samples can be employed to estimate the scoring accuracy that would be achieved were double human scoring used rather than single human scoring. This approach is illustrated in Yao et al. (2019a, 2019b) and Zhang et al. (2019).

Assessment Accuracy

Assessment accuracy is closely related to traditional notions of test reliability. Here the vector $\boldsymbol{\tau}$ of true assessment scores is the expectation of the observed score $\mathbf{X}$ given the test taker on a randomly selected test from a collection of parallel assessments. Here the covariance matrix Cov($\boldsymbol{\epsilon}$) is not easily estimated under normal circumstances. Two options are examined here. One is closely related to Cronbach's alpha (Cronbach, 1951). The other involves use of repeater data. For the type of data considered, available software related to IRT does not yet appear to be adequate for operational use.

Cronbach's Alpha

Cronbach's alpha is a traditional approach to assessment accuracy that does not require actual observation of a parallel assessment; however, this approach only applies under quite restrictive conditions, and the application here is somewhat unconventional. As noted in Examples 10, 11, and 12, the approach applies relatively well to GRE Analytical Writing and TOEFL Speaking but not to TOEFL Writing.

To begin, consider a generalization of Cronbach's alpha for use with vectors. Let $\mathbf{X}$ be a sum of observed vectors $\mathbf{Z}_a$, $1 \le a \le A$, such that each $\mathbf{Z}_a$ has a finite covariance matrix Cov($\mathbf{Z}_a$). Let $\mathbf{Z}_a$, $1 \le a \le A$, in turn have a decomposition $\mathbf{Z}_a = \boldsymbol{\zeta}_a + \boldsymbol{\gamma}_a$, where each $\boldsymbol{\zeta}_a$ has a finite covariance matrix $\text{Cov}\left(\boldsymbol{\zeta}_a\right)$ and each $\boldsymbol{\gamma}_a$ has a finite covariance matrix $\text{Cov}\left(\boldsymbol{\gamma}_a\right)$. The true assessment score $\boldsymbol{\tau}$ is the sum of the vectors $\boldsymbol{\zeta}_a$ for $1 \le a \le A$, and the measurement error ϵ is the sum of the vectors $\boldsymbol{\gamma}_a$ for $1 \le a \le A$. It is assumed that the errors $\boldsymbol{\gamma}_a$, $1 \le a \le A$, are mutually uncorrelated, and it is assumed that $\boldsymbol{\zeta}_a$ and $\boldsymbol{\gamma}_b$ are uncorrelated for positive integers a and b no greater than K.

To study sampling, let $\mathbf{Z}_*$ be the $K \times A$ matrix with columns $\mathbf{Z}_a$ for $1 \le a \le A$. In terms of sampling, for $1 \le i \le n$, let $\mathbf{Z}_{i^*}$, $1 \le i \le n$, be mutually independent $K \times A$ matrices with the same distribution as $\mathbf{Z}_*$, and let $\mathbf{Z}_{ai}$ be column a of $\mathbf{Z}_{i^*}$ for $1 \le i \le n$ and $1 \le a \le A$.

With the approach of Cronbach's alpha (Yao et al., 2019a), Cov $(\boldsymbol{\tau})$ is estimated by

$$\overline{\text{Cov}}\,(\boldsymbol{\tau}) = \frac{A}{A-1}\left[\overline{\text{Cov}}\,(\mathbf{X}) - \sum_{a=1}^{A} \overline{\text{Cov}}\left(\mathbf{Z}_a\right)\right]$$

$$= \frac{A}{A-1}\sum_{a=2}^{A}\sum_{b=1}^{a-1}\left[\overline{\text{Cov}}\left(\mathbf{Z}_a, \mathbf{Z}_b\right) + \overline{\text{Cov}}\left(\mathbf{Z}_b, \mathbf{Z}_a\right)\right]. \tag{96}$$

Here

$$\overline{\text{Cov}}\left(\mathbf{Z}_a, \mathbf{Z}_b\right) = \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{Z}_{ai} - \overline{E}\left(\mathbf{Z}_a\right)\right]\left[\mathbf{Z}_{ai} - \overline{E}\left(\mathbf{Z}_a\right)\right]' \tag{97}$$

estimates Cov($\mathbf{Z}_a$, $\mathbf{Z}_b$) so that $\overline{\text{Cov}}\,(\boldsymbol{\tau})$ estimates

$$\text{Cov}_*\,(\boldsymbol{\tau}) = \frac{A}{A-1}\sum_{a=2}^{A}\sum_{b=1}^{a-1}\left[\text{Cov}\left(\mathbf{Z}_a, \mathbf{Z}_b\right) + \text{Cov}\left(\mathbf{Z}_b, \mathbf{Z}_a\right)\right]. \tag{98}$$

For positive and distinct integers a and b no greater than A,

$$\text{Cov}\left(\mathbf{Z}_a, \mathbf{Z}_b\right) = \text{Cov}\left(\boldsymbol{\zeta}_a, \boldsymbol{\zeta}_b\right). \tag{99}$$

Therefore

$$\text{Cov}_*\,(\boldsymbol{\tau}) = \frac{A}{A-1}\sum_{a=2}^{A}\sum_{b=1}^{a-1}\left[\text{Cov}\left(\boldsymbol{\zeta}_a, \boldsymbol{\zeta}_b\right) + \text{Cov}\left(\boldsymbol{\zeta}_b, \boldsymbol{\zeta}_a\right)\right]. \tag{100}$$

The challenge with using $\overline{\text{Cov}}\,(\boldsymbol{\tau})$ is that

$$\text{Cov}\,(\boldsymbol{\tau}) = \sum_{a=1}^{A}\sum_{b=1}^{A}\text{Cov}\left(\boldsymbol{\zeta}_a, \boldsymbol{\zeta}_b\right) \tag{101}$$

rather than $\text{Cov}_*(\boldsymbol{\tau})$, so the relationship of $\text{Cov}_*(\boldsymbol{\tau})$ and $\text{Cov}(\boldsymbol{\tau})$ must be considered.

To compare the two expressions, observe that

$$\text{Cov}\left(\boldsymbol{\zeta}_a, \boldsymbol{\zeta}_b\right) + \text{Cov}\left(\boldsymbol{\zeta}_b, \boldsymbol{\zeta}_a\right) = \text{Cov}\left(\boldsymbol{\zeta}_a\right) + \text{Cov}\left(\boldsymbol{\zeta}_b\right) - \text{Cov}\left(\boldsymbol{\zeta}_a - \boldsymbol{\zeta}_b, \boldsymbol{\zeta}_a - \boldsymbol{\zeta}_b\right) \tag{102}$$

for all positive integers a and b no greater than A. Therefore

$$\text{Cov}_*(\boldsymbol{\tau}) = A\sum_{a=1}^{A}\text{Cov}\left(\boldsymbol{\zeta}_a\right) - \frac{A}{2(A-1)}\sum_{a=1}^{A}\sum_{b=1}^{A}\text{Cov}\left(\boldsymbol{\zeta}_a - \boldsymbol{\zeta}_b\right) \tag{103}$$

$$\text{Cov}(\boldsymbol{\tau}) = A\sum_{a=1}^{A}\text{Cov}\left(\boldsymbol{\zeta}_a\right) - \frac{1}{2}\sum_{a=1}^{A}\sum_{b=1}^{A}\text{Cov}\left(\boldsymbol{\zeta}_a - \boldsymbol{\zeta}_b\right). \tag{104}$$

It follows that the difference

$$\text{Cov}(\boldsymbol{\tau}) - \text{Cov}_*(\boldsymbol{\tau}) = \frac{1}{2(A-1)}\sum_{a=1}^{A}\sum_{b=1}^{A}\text{Cov}\left(\boldsymbol{\zeta}_a - \boldsymbol{\zeta}_b\right). \tag{105}$$

Then $\text{Cov}(\boldsymbol{\tau}) - \text{Cov}_*(\boldsymbol{\tau})$ is nonnegative definite so that $\overline{\text{Cov}}(\boldsymbol{\tau})$ may provide an underestimate of $\text{Cov}(\boldsymbol{\tau})$ in the sense that, for a K-dimensional vector $\mathbf{z}$, $\overline{\sigma}^2\left(\mathbf{z}'\boldsymbol{\tau}\right) = \mathbf{z}'\overline{\text{Cov}}(\boldsymbol{\tau})\,\mathbf{z}$ estimates $\sigma_*^2\left(\mathbf{z}'\boldsymbol{\tau}\right) \le \sigma^2\left(\mathbf{z}'\boldsymbol{\tau}\right)$. The bias issue disappears if $\boldsymbol{\zeta}_a - E\left(\boldsymbol{\zeta}_a\right)$ and $\boldsymbol{\zeta}_1 - E\left(\boldsymbol{\zeta}_1\right)$] are equal with probability 1 for each positive integer a such that $2 \le a \le A$.

In general, the discrepancy between $\text{Cov}(\boldsymbol{\tau})$ and $\text{Cov}_*(\boldsymbol{\tau})$ is much less of a problem if the number A of summands is large because of the divisor $2(A-1)$ in Equation 105. In addition, if the dimension K is sufficiently large, it may be possible to stratify the summands $\mathbf{Z}_a$ into strata such that, for a within a stratum, the differences $\boldsymbol{\zeta}_a - E\left(\boldsymbol{\zeta}_a\right)$ are very similar. This approach is used in stratified alpha (Cronbach et al., 1965). The challenge for constructed-response scoring is that it is often not possible to have values of A that exceed 2 owing to time constraints and owing to cost constraints on tests that restrict the number of constructed responses used.

In the case of transformations, Cronbach's alpha remains relevant. If $\mathbf{G}$ is an $L \times K$ matrix and $\mathbf{g}$ is an L-dimensional vector for some positive integer L, then the covariance matrix of the true assessment score $\boldsymbol{\upsilon} = \mathbf{g} + \mathbf{G}\boldsymbol{\tau}$ of $\mathbf{Y} = \mathbf{g} + \mathbf{G}\mathbf{X}$ can be estimated by

$$\overline{\text{Cov}}(\boldsymbol{\upsilon}) = \mathbf{G}\overline{\text{Cov}}(\boldsymbol{\tau})\,\mathbf{G}'. \tag{106}$$

A more general result can be useful in the case of nonlinear transformations that are continuously differentiable. Let $\mathbf{Y} = \mathbf{f}(\mathbf{X})$ for a continuously differentiable function $\mathbf{f}$ of dimension L on the K-dimensional vectors, and let $\boldsymbol{\upsilon}$ be the true score for $\mathbf{Y}$. Let $\nabla\mathbf{f}$ be the $L \times K$ matrix of partial derivatives of $\mathbf{f}$. Then

$$\overline{\text{Cov}}(\boldsymbol{\upsilon}) = [\nabla\mathbf{f}(\overline{E}(\mathbf{X})]\overline{\text{Cov}}(\boldsymbol{\tau})\left[\nabla\mathbf{f}(\overline{E}(\mathbf{X})\right]' \tag{107}$$

estimates $\text{Cov}(\boldsymbol{\upsilon})$. This result is most appropriate if $\mathbf{X}$ has a small covariance matrix or if $\nabla\mathbf{f}$ is continuously differentiable and does not vary much over the range of $\mathbf{X}$. This result can be useful in treating some scaled scores obtained by nonlinear transformations of raw scores.

IRT can be employed in place of Cronbach's alpha to estimate the true score $\boldsymbol{\tau}$ in a variety of cases that can include scaled scores (Haberman, 2013; Kolen et al., 1996).[1]

Unfortunately, this approach is quite difficult to apply to cases with test sections with only two constructed responses, and the approach does not apply directly to continuous measurements like those encountered in automated scoring.

One technical issue can arise. It is possible that $\overline{\text{Cov}}(\epsilon) = \overline{\text{Cov}}(\mathbf{X}) - \overline{\text{Cov}}(\boldsymbol{\tau})$ is not nonnegative definite. Such a case raises major questions concerning sampling and concerning appropriateness of methodology based on Cronbach's alpha for the application. As a consequence, formal treatment of this situation is not pursued here.

Example 10. In GRE Analytical Writing, two prompts are used. The first is an issue prompt that requires generation of an argument concerning a specified general issue, and the second is an argument prompt that requires analysis of a given argument. Although the prompts are not identical, their empirical behavior is quite similar so that Cronbach's alpha does provide a basis for analysis. In current practice, 1 human score and 10 computer-generated features are available for each prompt response. Here $K = 11$ and $A = 2$. For prompt a, $a = 1$ or 2, Element 1 of $\mathbf{Z}_a$ is the human score for that prompt response, and the remaining 10 elements of $\mathbf{Z}_a$ are the corresponding computer-generated numerical features for the response. The vector $\mathbf{X} = \mathbf{Z}_1 + \mathbf{Z}_2$. The value $\mathbf{c} = \frac{1}{2}\boldsymbol{\delta}_1$ in this application.

As in Example 4, it is also possible to use a human score and an e-rater score for each prompt. Here $K = 2$, $\mathbf{c}$ is $1/2\boldsymbol{\delta}_1$, $\mathbf{Z}_a$ has as first element the human score for prompt a and as second element the e-rater score for that prompt, and $\mathbf{X} = \mathbf{Z}_1 + \mathbf{Z}_2$. In the actual scoring procedure of GRE, Y is the arithmetic average of the two elements of $\mathbf{X}$ rather than $\widehat{\nu} = \widehat{\tau}_1$. The loss in precision from use of Y can be measured with Equations 75 and 77.

Example 11. In TOEFL iBT Speaking, four prompts are used. Although the prompts are not identical in nature, behavior is again sufficiently similar to use the approach of Cronbach's alpha. Here 1 human score and 28 computer-generated numerical feature scores are used for each prompt response. Here $K = 29$ and $A = 4$. For prompt a, $1 \le a \le 4$, the first element of $\mathbf{Z}_a$ is the human score for the prompt, and the remaining elements are the corresponding computer-generated numerical features for the response. The sum of the $\mathbf{Z}_a$ is $\mathbf{X}$, and $\mathbf{c} = \boldsymbol{\delta}_1$. An alternative has $K = 2$, the first element of $\mathbf{Z}_a$ equal to the human score, and the other element of $\mathbf{Z}_a$ equal to the SpeechRater score. For some $d > 1$, a PBLP predictor $\widehat{\nu}_d$ is used based on the selected $\mathbf{Z}_a$, $1 \le a \le 4$. To evaluate the effect of PBLP based on just SpeechRater and human scores versus PBLP based on feature scores and human scores, define $\widehat{\nu}_d$ based on $K = 29$, and let $Y = a + \mathbf{b}'\,\mathbf{X}$ be defined based on the corresponding $\widehat{\nu}_d$ based on $K = 2$. This procedure is feasible because the SpeechRater score is a linear function of the feature scores. For an analysis with six rather than four prompts, see Zhang et al. (2019).

Repeater Data

Repeater data provide a direct approach to the error of prediction of a score in a parallel assessment; however, this approach must treat selection bias in almost all realistic situations. This approach requires data from multiple test administrations at multiple times, for a test taker cannot be observed more than once at the same time. In most cases that have been examined, to each observation i, $1 \le i \le n$, corresponds an integer code R_i that uniquely identifies the test taker. This code is typically a customer number used to keep track of test takers who take a test more than once. The key feature of the code R_i is that, for $1 \le i < j \le n$, $R_i = R_j$ if, and only if, the test taker for response vector $\mathbf{X}_i$ is also the test taker for response vector $\mathbf{X}_j$. Without loss of generality, assume that sorting of data has been used such that when $R_i = R_j$ and $i < j$, response $\mathbf{X}_i$ corresponds to an earlier administration than response $\mathbf{X}_j$. Of interest for repeater analysis are test takers in the sample who took the test at least twice. Let $\mathcal{R}$, the repeater set, be the set of integers r such that $R_i = r$ for more than one positive integer $i \le n$. For r in $\mathcal{R}$, let $i(r, 1)$ be the smallest positive integer i such that $R_i = r$ so that $i(r, 1)$ is the earliest observation of the test taker with code r. Let $i(r, 2)$ be the second observation on the test taker with code r so that $i(r, 2)$ exceeds $i(r, 1)$ but no code $R_i = r$ for $i(r, 1) < i < i(r, 2)$. Analysis would be straightforward if the observations $\mathbf{X}_{i(r,\,1)}$ from the first occasion and $\mathbf{X}_{i(r,\,2)}$ from the second occasion were to have the same distribution as $\mathbf{X}$; however, in practice, membership in $\mathcal{R}$ is likely to be related to the observed response vectors, for test takers usually do not take an examination more than once unless they have reason to believe that they can improve their performance sufficiently to warrant the time and monetary expense involved in repeating an assessment. In addition, the act of taking an assessment can improve knowledge of how to take the assessment so that improvement in performance is often encountered on the second assessment. One approach to this problem based on Haberman (1984) involves minimum discriminant information adjustment (MDIA). This approach requires that $\mathcal{R}$ have at least $(3K + 2)(K + 1)/2$ members.

With MDIA, a positive weight w_r is assigned to each r in $\mathcal{R}$. The w_r, r in $\mathcal{R}$, satisfy the condition that the logarithm $\log w_r$ is a linear combination of the following quantities: (a) the constant 1; (b) the elements $X_{i(r,\,1)k}$, $1 \le k \le K$, of the vector $\mathbf{X}_{i(r,\,1)}$ from the first testing; (c) the elements $X_{i(r,\,2)k}$, $1 \le k \le K$, of the vector $\mathbf{X}_{i(r,\,2)}$ from the second testing; (d) the products $X_{i(r,\,1)k}X_{i(r,\,1)m}$, $1 \le k \le m \le K$, from the first testing; (e) the products $X_{i(r,\,2)k}X_{i(r,\,2)m}$, $1 \le k \le m \le K$, from the second testing; and (f) the differences of products $X_{i(r,\,1)k}X_{i(r,\,2)m} - X_{i(r,\,2)k}X_{i(r,\,1)m}$, $1 \le k < m \le K$. In addition, the w_r satisfy the following linear constraints:

$$\sum_{r \in \mathcal{R}} w_r = 1, \tag{108}$$

$$\sum_{r \in \mathcal{R}} w_r X_{i(r,1)} = \sum_{r \in \mathcal{R}} w_r X_{i(r,2)} = \overline{E}\left(X_k\right), 1 \le k \le K, \tag{109}$$

$$\sum_{r\in\mathcal{R}} w_r X_{i(r,1)k} X_{i(r,1)m} = \sum_{r\in\mathcal{R}} w_r X_{i(r,2)k} X_{i(r,2)m} = \overline{E}\left(X_k X_m\right), 1 \le k \le m \le K, \tag{110}$$

$$\sum_{r\in\mathcal{R}} w_r X_{i(r,1)k} X_{i(r,2)m} = \sum_{r\in\mathcal{R}} w_r X_{i(r,2)k} X_{i(r,1)m}, 1 \le k < m \le K. \tag{111}$$

The requirements on the weights w_r uniquely determine them. If $K=1$, then Equation 111 disappears and the linear combination requirement on $\log w_r$ does not include the product differences $X_{i(r,1)k}X_{i(r,2)m} - X_{i(r,2)k}X_{i(r,1)m}$ for $1 \le k < m \le K$. Let $N(\mathcal{R})$ be the number of codes of test takers in the repeater set $\mathcal{R}$, and let $u_r = 1/N(\mathcal{R})$, r in $\mathcal{R}$, define uniform weights on $\mathcal{R}$. Then the weights w_r, r in $\mathcal{R}$, minimize the discriminant information $\sum_{r\in\mathcal{R}} w_r \log\left(w_r/u_r\right)$ subject to Equations 108, 109, 110, and 111. In some sense, the weights w_r are chosen to be as close as possible to uniform weights subject to the constraints of Equations 108, 109, 110, and 111. A program in Fortran 95 to perform the computations is available within ETS (Haberman, 2014) and is available outside of ETS for noncommercial use. The program uses the Newton–Raphson algorithm associated with log-linear models to find the weights w_r.

The weights w_r, r in $\mathcal{R}$, are defined as long as some positive real numbers v_r, r in $\mathcal{R}$, exist such that

$$\sum_{r\in\mathcal{R}} r_j = 1, \tag{112}$$

$$\sum_{r\in\mathcal{R}} v_r X_{i(r,1)} = \sum_{r\in\mathcal{R}} v_r X_{i(r,2)} = \overline{E}\left(X_k\right), 1 \le k \le K, \tag{113}$$

$$\sum_{r\in\mathcal{R}} v_r X_{i(r,1)k} X_{i(r,1)m} = \sum_{r\in\mathcal{R}} v_r X_{i(r,2)k} X_{i(r,2)m} = \overline{E}\left(X_k X_m\right), 1 \le k \le m \le K, \tag{114}$$

$$\sum_{r\in\mathcal{R}} v_r X_{i(r,1)k} X_{i(r,2)m} = \sum_{r\in\mathcal{R}} v_r X_{i(r,2)k} X_{i(r,1)m}, 1 \le k < m \le K. \tag{115}$$

Given the weights w_r, r in $\mathcal{R}$, the estimated covariance matrix of the vector $\boldsymbol{\tau}$ of true assessment scores is

$$\overline{\mathrm{Cov}}\left(\boldsymbol{\tau}\right) = \sum_{r\in\mathcal{R}} w_r \left[\mathbf{X}_{i(r,1)j} - \overline{E}\left(\mathbf{X}\right)\right]\left[\mathbf{X}_{i(r,2)j} - \overline{E}\left(\mathbf{X}\right)\right]'. \tag{116}$$

It is always the case that $\overline{\mathrm{Cov}}\left(\epsilon\right) = \overline{\mathrm{Cov}}\left(\mathbf{X}\right) - \overline{\mathrm{Cov}}\left(\boldsymbol{\tau}\right)$ is nonnegative definite.

The use of MDIA can be unsatisfactory even if the w_r, r in $\mathcal{R}$, are defined if the coefficient of variation of the w_r is quite large and the number of members of $\mathcal{R}$ is small or of moderate size. This result arises from general results concerning asymptotic variances of estimates of expectations by use of MDIA (Haberman, 1984). The issue arises if membership in $\mathcal{R}$ is quite strongly related to the corresponding two response vectors. Applications of MDIA can sometimes be more complex and involve the variables G_i, $1 \le i \le n$, as well as $\mathbf{X}_i$. In the examples considered in Yao et al. (2019a, 2019b), the values of $N(\mathcal{R})$ were 168,595 for TOEFL Writing, 4,739 for GRE Writing, and 23,673 for PRAXIS Writing. The respective coefficients of variation were 3.77, 1.05, and 1.54 so that MDIA was satisfactory. In these cases, although repeater samples indeed differ in characteristics from complete samples, the differences were not sufficiently dramatic, given the large number of repeaters observed, to cause a major problem. Serious problems are most likely to arise in cases in which test takers just need to perform above a low cut point, and the vast preponderance meet the cut point the first time they take the test. In such a case, the repeaters are a very highly selected group, for a test taker is highly unlikely to take a test twice if the initial performance was satisfactory for the purpose of the test.

Example 12. Consider the TOEFL scaled section scores of Example 7. Here repeater data can be employed. It is important to note that in prediction of specific true assessment section scores, all section scores are applied, and in prediction of the total score, observed section scores need not receive equal weight. Repeater data can also be applied to the raw scores and item scores in TOEFL Writing even though use of the methodology based on Cronbach's alpha is not satisfactory because of quite different behavior of the responses to the two prompts.

Example 13. Consider the case of TOEFL Writing introduced in Example 3. For the data used in Yao et al. (2019a, 2019b), the proportional reduction in mean squared error only decreases from .81 for BLP ($d=1$) to .80 for PBLP with $d=5$. At the same time $\overline{\sigma}^2(E\left(\Delta_d\left(G\right)\right)$ decreases from .0038 to .0014. One cautionary note is that the largest absolute value of $\overline{\Delta}_{ds}\left(h\right)$

decreases only from .137 to .096. Use of $d = 100$ involves a more substantial change in results. Here proportional reduction in mean squared error is down to .76, and $\overline{\sigma}^2(E\left(\Delta_d(G)\right)$ is only .0001. The largest absolute value of $\overline{\Delta}_{ds}(h)$ is .067. It is worth noting that use of two human scorers for each prompt without any use of computer-generated features leads to a maximum value of $|\overline{\Delta}_{1s}(h)|$ of .007 and to a proportional reduction in mean squared error with BLP of .83, which is more satisfactory than any result with both one human rater and nine e-rater features.

Use of repeater data has the great virtue of flexibility, although imperfect attempts to correct for selection bias are a consistent danger. For evaluation of assessment reliability for cases with few items, it is often the case that few good alternatives exist. Analysis with repeater data may reveal weaknesses in an assessment not evident from examination of scoring accuracy. For example, in TOEFL Writing using BLP with one human score and nine essay features, the estimated proportional reduction in mean squared error for BLP for assessment accuracy of .81 was much lower than the corresponding value of .89 for scoring accuracy.

Conclusions

This report outlines the procedures required to apply BLP and PBLP in assessments. These methods permit more efficient use of information from an assessment and more careful examination of fairness problems. The theoretical basis of BLP and PBLP is described in the sections Best Linear Prediction and Penalized Best Linear Prediction. The problem of different definitions of true scores is explored in the sections Scoring Accuracy and Assessment Accuracy. Estimation procedures are provided in the section Estimation for Best Linear Prediction and Penalized Best Linear Prediction.

It must be emphasized that many details can require attention in practice. Sampling discussed in this report has been simple random sampling; however, complex sampling does arise in assessments. Monitoring of results over time is always important. Testing populations can change substantially over time so that an analysis one year may no longer be appropriate 5 years later.

This report does not consider several issues important in practice. Linking is important, and a change from use of the observed score O to use of $\overline{\nu}_d$ requires changes in linking approaches that depend on the specific assessment. Rounding and truncation can add quite substantial effects on results. This issue, although not familiar in psychometrics, has been discussed for more than 100 years (Sheppard, 1898). More recent discussions include Dempster and Rubin (1983), Kolassa (1989), and Kolassa and McCullagh (1990). As a general rule, rounding and truncation of intermediate results should be avoided as much as possible until final scores must be reported.

This report deals with cases in which the relevant expectations and covariance matrices can be estimated accurately, that is, sample sizes are large and dimensions of vectors and matrices are modest in size and not related to sample size. To apply BLP and PBLP to testing programs with small cohorts of test takers and methods to score constructed responses that involve very high dimensions and irregular statistical properties is a far more difficult enterprise than the problems considered in this report.

Acknowledgments

The author thanks the reviewers and the editor for their comments, Mo Zhang and Lili Yao for manuscript review and for additional information on data used in publications, and Neil Dorans for manuscript review and for assistance in obtaining information concerning current practices.

Note

[1] Further cases can be found in revisions of Haberman (2013) at https://github.com/EducationalTestingService/MIRT/blob/master/Documents/irtprogram.pdf

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment*, *4*(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified-parallel tests. *Educational and Psychological Measurement*, *25*(2), 291–312. https://doi.org/10.1177/001316446502500201

Dempster, A. P., & Rubin, D. (1983). Rounding error in regression: The appropriateness of Sheppard's corrections. *Journal of the Royal Statistical Society, Series B*, *45*(1), 51–59. https://doi.org/10.1111/j.2517-6161.1983.tb01230.x

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*(3), 971–988. https://doi.org/10.1214/aos/1176346715

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*(2), 204–229. https://doi.org/10.3102/1076998607302636

Haberman, S. J. (2011). *Use of e-rater in scoring of the TOEFL iBT writing test* (Research Report No. RR-11-25). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02261.x

Haberman, S. J. (2013). *A general program for item-response analysis that employs the stabilized Newton–Raphson algorithm* (Research Report No. RR-13-32). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2013.tb02339.x

Haberman, S. J. (2014). *A program for adjustment by minimum discriminant information* (Research Memorandum No. RM-14-01). Educational Testing Service.

Haberman, S. J., & Qian, J. (2004). *The best linear predictor for true score from a direct estimate and several derived estimates* (Research Report No. RR-04-35). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2004.tb01962.x

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*(1), 6–23. https://doi.org/10.3102/1076998606298036

Haberman, S. J., & Sinharay, S. (2011). *How does the knowledge of subgroup membership of examinees affect the prediction of true subscores?* (Research Report No. RR-11-43). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2011.tb02279.x

Haberman, S. J., & Sinharay, S. (2013). Does subgroup membership information lead to better estimation of true subscores? *British Journal of Mathematical and Statistical Psychology*, *66*(3), 452–469. https://doi.org/10.1111/j.2044-8317.2012.02061.x

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 79–95. https://doi.org/10.1348/000711007x248875

Haberman, S. J., & Yao, L. (2015). Repeater analysis for combining information from different assessments. *Journal of Educational Measurement*, *52*(2), 223–251. https://doi.org/10.1111/jedm.12075

Haberman, S. J., Yao, L., & Sinharay, S. (2015). Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 363–385. https://doi.org/10.1111/bmsp.12052

Higgins, D., Xi, X., Zechner, K., & Williamson, D. (2011). A three-stage approach to the automated scoring of spontaneous spoken responses. *Computer Speech and Language*, *25*(2), 282–306. https://doi.org/10.1016/j.csl.2010.06.001

Kelley, T. L. (1923). *Statistical method*. Macmillan.

Kolassa, J. E. (1989). *Topics in series approximations to distribution functions* [Unpublished doctoral dissertation]. University of Chicago.

Kolassa, J. E., & McCullagh, P. (1990). Edgeworth series for lattice distributions. *Annals of Statistics*, *18*, 981–985. https://doi.org/10.1007/978-1-4757-4275-6_3

Kolen, M. J., Zeng, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores using IRT. *Journal of Educational Measurement*, *33*(2), 129–140. https://doi.org/10.1111/j.1745-3984.1996.tb00485.x

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

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2008). *Comparison of subscores based on classical test theory methods* (Research Report No. RR-08-54). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2008.tb02140.x

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: An evaluation of methods using empirical data. *Applied Measurement in Education*, *23*(3), 266–285. https://doi.org/10.1080/08957347.2010.486287

Rao, C. R., & Mitra, S. K. (1972). Generalized inverse of a matrix and its applications. *Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability: Vol. 1. Theory of statistics* (pp. 601–620). University of California Press.

Sheppard, W. F. (1898). On the calculation of the most probable values of frequency-constants, for data arranged according to equidistant division of a scale. *Proceedings of the London Mathematical Society*, *s1-29*(1), 353–380. https://doi.org/10.1112/plms/s1-29.1.353

Sinharay, S., & Haberman, S. J. (2014). An empirical investigation of population invariance in the value of subscores. *International Journal of Testing*, *14*(1), 22–48. https://doi.org/10.1080/15305058.2013.822712

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: To report or not to report. *Educational Measurement: Issues and Practice*, *26*(4), 21–28. https://doi.org/10.1111/j.1745-3992.2007.00105.x

Sinharay, S., Puhan, G., & Haberman, S. J. (2010). Reporting diagnostic subscores in educational testing: Temptations, pitfalls, and some solutions. *Multivariate Behavioral Research*, *45*(3), 553–573. https://doi.org/10.1080/00273171.2010.483382

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*(3), 29–40. https://doi.org/10.1111/j.1745-3992.2011.00208.x

Sinharay, S., Puhan, G., Haberman, S. J., & Hambleton, R. K. (2018). Subscores: When to communicate them, what are their alternatives, and some recommendations. In D. Zapata-Rivera (Ed.), *Score reporting: Research and applications* (pp. 35–49). Routledge. https://doi.org/10.4324/9781351136501-4

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). Erlbaum.

Yao, L., Haberman, S. J., & Zhang, M. (2019a). Penalized best linear prediction of true test scores. *Psychometrika*, *84*(1), 186–211. https://doi.org/10.1007/s11336-018-9636-7

Yao, L., Haberman, S. J., & Zhang, M. (2019b). *Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors* (Research Report No. RR-19-13). Educational Testing Service. https://doi.org/10.1002/ets2.12248

Zhang, M., Yao, L., Haberman, S. J., & Dorans, N. J. (2019). Assessing scoring accuracy and assessment accuracy for spoken responses. In K. Zechner & K. Evanini (Eds.), *Automated speaking assessment* (pp. 32–58). Routledge. https://doi.org/10.4324/9781315165103-3

Suggested citation:

Haberman, S. J. (2020). *Application of best linear prediction and penalized best linear prediction to ETS tests* (Research Report No. RR-20-08). Educational Testing Service. https://doi.org/10.1002/ets2.12290

Action Editor: James Carlson

Reviewers: Wei Wang, Tsung-Han Ho, Jiyun Zu, and Sandip Sinharay