Reliability-Based Feature Weighting for Automated Essay Scoring

Applied Psychological Measurement

2015, Vol. 39(4) 303–313

© The Author(s) 2014

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621614561630

apm.sagepub.com

Yigal Attali¹

Abstract

From their earliest origins, automated essay scoring systems strived to emulate human essay scores and viewed them as their ultimate validity criterion. Consequently, the importance (or weight) and even identity of computed essay features in the composite machine score were determined by statistical techniques that sought to optimally predict human scores from essay features. However, it is evident that machine evaluation of essays is fundamentally different from human evaluation and therefore is not likely to measure the same set of writing skills. As a consequence, feature weights of human-prediction machine scores (reflecting their importance in the composite scores) are bound to reflect statistical artifacts. This article suggests alternative feature weighting schemes based on the premise of maximizing reliability and internal consistency of the composite score. The article shows, in the context of a large-scale writing assessment, that these alternative weighting schemes are significantly different from human-prediction weights and give rise to comparable or even superior reliability and validity coefficients.

Keywords

writing, essay assessment, automated scoring, validity, reliability

In recent years, several high-stakes graduate admission programs have implemented automated essay scoring (AES) in the operational scoring of their writing assessments, including the Graduate Management Admission Test (GMAT), the Graduate Record Examination (GRE) General test, and the Test of English as a Foreign Language (TOEFL).

In contrast to human essay scoring, which is based on forming a holistic impression of the overall quality of the essay or of several distinct writing traits, AES is analytic in nature. That is, AES, is based on computing multiple features of the text that are then aggregated into a composite essay score. Each of these features is meant to measure some aspect of the quality of the essay, for example, by identifying grammar errors or vocabulary levels. As Bennett and Bejar (1998) note, an advantage of the analytic nature of automated scoring is that it makes it possible to control (to some degree) what Whitely (1983) calls the construct representation (the meaning of scores based on internal evidence) and nomothetic span (the meaning of scores based on relationships with external variables) of automated scores. That is, the relative weight of essay

¹Educational Testing Service, Princeton, NJ, USA

Corresponding Author:

Yigal Attali, Educational Testing Service, Rosedale Rd., MS-16-R, Princeton, NJ 08541, USA.

Email: yattali@ets.org

features in the composite score (e.g., word use vs. grammatical conventions vs. fluency measures) can have a substantial effect on the meaning of the automated score.

Nevertheless, weighting schemes for AES have been based almost exclusively on the concept of optimizing the relation of automated scores with human scores of the same essays, for example, by using multiple regression (Burstein, Tetreault, & Madnani, 2013; Foltz, Streeter, Lochbaum, & Landauer, 2013; Schultz, 2013). That is, multiple regression of essay features on the human scores for the same essays is performed, and as a result, optimal feature weights for the prediction of human scores from the set of essay features are derived. Human-prediction weighting is based on the rationale that human essay scores should serve as the ultimate validity criterion for AES. However, machine evaluation of essays is fundamentally different than human evaluation. The machine does not read and understand the text and its semantic analysis of the text is limited, compared with human evaluation. Therefore, assessing such aspects as the quality of argumentation or the development of characters in a narrative will be difficult for the machine. However, automated evaluation of the text could provide superior assessment of other aspects of writing quality, such as adherence to conventions of writing or writing fluency. As a result, the machine is not likely to measure the same set of writing skills as human raters do.

Because the machine does not measure the construct in the same way as human raters do, feature weights that are based on the human-prediction approach (HP) do not necessarily reflect the relative importance of human-measured traits. The most evident example of this possibility is the inordinately high weight that this approach assigns for essay length (EL; Attali, 2007a), exceeding in many cases 50% of the total score (see below for additional information). Although human raters may certainly be affected by EL (Powers, 2005), the high weight assigned to EL may reflect the difficulty of automated scores to measure higherorder writing skills (like organization and development) that are merely statistically correlated with EL. Moreover, this example underscores the possibility that human-prediction weights would not produce a composite machine score with the strongest measurement properties because the criterion for these weights was not based on optimizing the measurement properties of the features.

What could then be the basis for feature weighting for AES? Measurement theory and practice suggest several possibilities (Wang & Stanley, 1970). With judgment-based weighting, experts could set weights according to judged importance of AES features or measured dimensions of writing. Ben-Simon and Bennett (2007) carried out this approach with two independent committees of writing experts. They found that judgment-based scores generalized well to new prompts, compared with the “brute-force” human-prediction-based scores. In the absence of expert judgments, an alternative approach that is similarly not based on empirical data is to assign equal weights to features or dimensions. The rationale of this approach is that the features are all measures of a single underlying construct, similar to a fixed set of items or components of a test. Attali (2007a) compared equal weights with human-prediction weights and found similar test–retest reliabilities and correlations with human scores.

Another key alternative to human-prediction weighting is to choose weights that maximize the reliability (internal consistency) of the composite score. If components are congeneric (i.e., they measure the same trait or perfectly correlated traits), then the optimum weighting from a psychometric perspective is the one that maximizes the reliability of the composite, because such a composite will have the highest possible correlation with any other variable (Li, Rosenthal, & Rubin, 1996). In this case, the optimal weight for a component (assuming equal variances for components) would be proportional to $\sqrt{r}/(1 - r)$, where r is the reliability of the component (Li et al., 1996). Even if components are not congeneric, weights that maximize reliability are of interest because they provide a limit against which to evaluate the efficiency

(in terms of optimizing correlations with other variables) of any given weighting (Haertel, 2006). For the general case of components that do not necessarily measure the same trait, maximal reliability and the weights associated with it can be found from the solution to the eigenvalue problem $\mathbf{u}'(\mathbf{R}^* - \lambda \mathbf{R}) = \mathbf{0}'$, where $\mathbf{R}$ is the component correlation matrix, $\mathbf{R}^*$ is the same matrix with component reliabilities on the diagonal, and $\mathbf{u}$ is the eigenvector corresponding to the largest eigenvalue λ (Haertel, 2006). In this study, the reliability of the individual features (the components that constitute the composite score) was estimated from the correlations between feature scores on two writing tasks that were completed at the same time (alternate-form reliability estimates).

Another related key alternative is to choose weights on the basis of a principal component analysis (PCA). It has been shown (Bentler, 1968) that when standardized item scores are weighted by their respective loading under the first principal component (PC), the internal consistency of the composite score is maximized. In the context of AES, the item scores to be weighted will be feature scores.

The purpose of this article was to evaluate the reliability- and PCA-based weighting schemes and to compare them with the human-prediction scheme. This evaluation was conducted in the context of a large-scale admissions assessment, which includes two essay writing tasks. In Task 1, a brief passage is presented in which the author makes a case for some course of action or interpretation of events by presenting claims backed by reasons and evidence. The student's task is to discuss the logical soundness of the author's case by critically examining the line of reasoning and the use of evidence. In Task 2, the student is asked to discuss and express his or her perspective on a topic of general interest. For both tasks, the scoring rubric emphasized ideas, development and organization, word choice, sentence fluency, and conventions. The assessment treats the two tasks as two "items" and reports only one writing score, computed as the average of the two essay scores. The analyses in this article are based on 380,463 examinees who completed the assessment within a 6-month period in 2011.

The evaluation of human-prediction and reliability-based weighting schemes was based on the e-rater (Attali & Burstein, 2006) feature set. E-rater is a state-of-the-art AES system that is currently being used in the operational scoring of the GRE and TOEFL writing assessments. Table 1 presents a brief explanation of the features used in this study. Attali and Burstein (2006) and Burstein et al. (2013) provide a thorough description of the methods used to compute each of the features. The HP was implemented by conducting, for each of the two tasks, a multiple-regression analysis of the 10 features on the operational human rating for the essay. The reliability approach was based on estimating the reliability of each feature as the correlation between the feature scores on both tasks. Two types of reliability-based scores were computed. The optimized reliability (OR) score was based on finding the optimized weights for each task through the feature correlation matrix. The simple reliability (SR) score was based on the assumption that the features are congeneric and approximating optimal weights through $\sqrt{r}/(1 - r)$, where r is the reliability of the feature score. The PC score was based on a PCA for each task.

Comparisons of the four types of scores were performed in terms of several reliability and validity coefficients. Task reliability was estimated through cross-task correlations of each score. Convergent validity coefficients were based on correlations with human scores (both within- and cross-task) and with a score of verbal ability from the same assessment. Human scores constitute the most important convergent validity measure for machine scores as they are designed as an alternative measure of the same student responses. Verbal ability measures (mostly reading comprehension) are theoretically related to writing ability as measures of different language skills. Divergent validity coefficients were based on correlations with EL and with a score of mathematical ability from the same assessment. Although EL was explicitly used as a feature in this study due to its high correlation with human essay scores (.65 and .69

Table 1. Features Used in This Study.

Feature	Description
Vocabulary	Based on the frequencies of essay words in a large corpus of text
Word length	Average word length
Grammar	Based on the rates of errors such as fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, missing possessives, and wrong or missing words
Usage	Based on the rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors
Mechanics	Based on the rates of spelling, capitalization, and punctuation errors
Col/prep	Evaluates correct usage of collocations and prepositions
Organization	Based on the detection of discourse elements (i.e., introduction, thesis, main points, supporting ideas, conclusion)
Essay length	Based on the number of words in essay
Style	Based on the rates of cases such as overly repetitious words, inappropriate use of words and phrases, sentences beginning with coordinated conjunctions, very long and short sentences, and passive voice sentences
Semantic analysis	Based on the similarity of the essay vocabulary to vocabulary of high- versus low-scored essays across topics.

in this study, see the following) and interpretation as a fluency measure, very high correlations with EL are not desirable from a face validity and coachability perspective (Powers, 2005). In fact, criticisms of AES (Ericsson & Haswell, 2006) have often focused on the use of surface features of the text (EL being the most prominent example) that are easily computed but do not reflect true analysis of writing skills.

The questions that were examined were as follows:

Research Question 1: Are there differences in the patterns of feature weights between human-prediction and reliability-based machine scores?

Research Question 2: How do the different weighting schemes compare with respect to convergent and divergent validity coefficients, as well as their reliability coefficients?

Research Question 3: When human and machine scores are combined (averaged), how do the different weighting schemes compare with respect to convergent and divergent validity coefficients, as well as their reliability coefficients?

In evaluating the importance of differences in correlation coefficients for Research Questions 2 and 3, one should remember that the scores obtained from different weighting schemes are not independent of each other (and in fact are highly correlated). An appropriate effect size measure for evaluating differences in correlations of these scores with other measures is Cohen's (1988) f^2 , which is based on incremental variance explained in the context of multiple-regression analyses. By convention, f^2 effect sizes of 0.02, 0.15, and 0.35 are termed small, medium, and large, respectively (Cohen, 1988). For an initial correlation of .70 (typical in AES studies), small, medium, and large effects would mean an increase to correlations of .71, .75, and .79, respectively. As an illustrative example of the magnitude of these effects in AES, consider the results of Attali (2011), who compared the effects of adding semantic analysis (or "content" analysis) measures to a base set of non-semantic features (all but the last feature in Table 1) on the prediction of human essay scores. Across four writing tasks, the largest increases in multiple correlations ranged from .02 to .15, with f^2 effect sizes of 0.13 (GRE issue, from a multiple correlation

Table 2. Relative Feature Weights (in percentage values).

Feature	Human prediction		Optimized reliability		Simple reliability		Principal components	
	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
Vocabulary	0	2	5	6	7	7	8	9
Word length	6	5	3	4	7	7	6	7
Grammar	4	4	9	9	7	7	12	12
Usage	5	8	8	8	7	7	11	10
Mechanics	3	7	11	12	12	12	11	11
Col/prep	8	6	10	10	8	8	13	13
Organization	5	2	11	10	13	13	8	6
Essay length	45	49	26	23	22	22	12	10
Style	−1	−1	5	5	4	4	9	9
Semantic an.	25	19	13	14	13	13	12	13

Note. SR weights are identical for the two tasks. Col/prep = collocation and preposition use. SR = simple reliability.

of .81 to .83), 0.20 (GRE argument, from .75 to .80), 0.07 (TOEFL independent, from .76 to .78), and 0.50 (TOEFL integrated, from .62 to .77).

Results

All coefficients and results are based on the full sample (around 380,000 examinees) as the coefficients and results from a twofold cross-validation design were virtually identical to the full sample results and are more cumbersome to report. Table 2 presents the relative feature weights for all weighting schemes (see the Appendix for technical details). For each of the weighting schemes, the weights for Tasks 1 and 2 are quite similar. The most notable difference across weighting schemes are the high weights for EL and semantic analysis for HP. As a consequence, all other features receive a low weight under HP scoring, and style even shows a negative weight. The variability in weights across features is highest for HP (*SD* of 14% for each task), followed by OR (5% and 6%), SR (5% and 4%), and PC (2% for both tasks).

Tables 3 to 6 present a comparison of the four weighting schemes in terms of validity and reliability coefficients. Human essay scores constitute the principal convergent validity measure and Table 3 presents correlations of human scores with the different machine scores. As expected, same-task human-machine correlations were highest for HP (.75 and .79 for Tasks 1 and 2, respectively), as by definition these weights are optimized to maximize this type of correlation. These correlations decreased the more divergent the weights were from HP, and consequently were lowest for PC (.69 and .71). However, same-task correlations are limited in their generalizability. When cross-task human-machine correlations are compared, almost identical results can be observed for HP (.68 and .63), OR (.67 and .64), and SR (.66 and .63) and slightly lower results for PC (.65 and .62).

Table 4 presents a comparison of additional convergent and divergent validity coefficients. The score from the verbal ability test was used as an additional convergent validity measure. For this measure, the pattern of correlations was reversed, with the highest (best) correlations for PC (.69 and .68) and lowest for HP (.65 and .60). This pattern of results continued with the two divergent validity measures—the score from the mathematical measure and EL—with the best (lowest) correlations for PC and worst for HP. The differences were especially prominent for EL, with very high correlations for HP (.87 and .88) and much lower correlations for PC

Table 3. Correlations With Human Scores.

	HP	OR	SR	PC
Same task				
Task 1	.75	.73	.72	.69
Task 2	.79	.76	.76	.71
Cross-task				
Task 1/Human Task 2	.68	.67	.66	.65
Task 2/Human Task 1	.63	.64	.63	.62

Note. HP = human prediction; OR = optimized reliability; SR = simple reliability; PC = principal component.

Table 4. Correlations With Other Scores.

	Human	HP	OR	SR	PC
Task 1					
Verbal	.59	.65	.66	.67	.69
Math	.23	.22	.22	.24	.20
Essay length	.65	.87	.81	.77	.67
Task 2					
Verbal	.54	.60	.65	.65	.68
Math	.14	.17	.17	.19	.15
Essay length	.69	.88	.75	.75	.60

Note. HP = human prediction; OR = optimized reliability; SR = simple reliability; PC = principal component.

Table 5. Cross-Task Correlations (Reliability Coefficients).

Human	HP	OR	SR	PC
.56	.82	.83	.83	.82

Note. HP = human prediction; OR = optimized reliability; SR = simple reliability; PC = principal component.

(.67 and .60), correlations that are comparable or even lower than those for human scores (.65 and .69).

Because the writing assessment is composed of two writing tasks, it is possible to define reliability coefficients as cross-task correlations. This operationalization is limited in that the instructions for the two tasks are not identical, which is likely to result in some unique task variability. However, given the similarity of the scoring rubrics for the two tasks and the assessment's reporting of a single combined score, cross-task correlations are a reasonable (if not perfect) measure of score reliability. Table 5 shows similar reliability coefficients for all four weighting schemes, with (as expected) slightly higher results for the reliability maximization schemes OR and SR (.83) and .82 for HP and PC. The reliability of human scores (.56) was significantly lower than the reliability of machine scores.

Table 6 presents results for combined (average) human-machine scores. The first line shows that reliability coefficients are highest for combined human-PC scores (.78) and lowest for combined human-HP scores (.76). Similarly, convergent validity coefficients for the verbal measure are highest for combined human-PC scores (.69 and .66) and lowest for combined human-HP

Table 6. Correlations for Human-Machine Average Scores.

	HP	OR	SR	PC
Task 1/Task 2	.76	.77	.77	.78
Task 1/Verbal	.66	.67	.68	.69
Task 2/Verbal	.60	.63	.63	.66
Task 1/Math	.24	.24	.25	.23
Task 2/Math	.16	.16	.17	.16

Note. HP = human prediction; OR = optimized reliability; SR = simple reliability; PC = principal component.

scores (.66 and .60). Finally, divergent validity coefficients for the math measure are almost identical across the different machine scores, but are still lowest (best) for combined human-PC scores (.23 and .16).

In summary, the HP weighting scheme shows superior results only when same-essay human-machine correlations are considered. When we consider other convergent validity measures (cross-task human-machine correlations and correlations with the verbal ability score), divergent validity measures (the math ability score and EL), and reliability coefficients, the reliability-based weighing schemes show similar or superior results. An added benefit of the reliability-based scores is their more homogeneous weights compared with HP weighting, which entails broader construct coverage.

Discussion

This article compared the human-prediction weighting scheme with several reliability-optimization weighting schemes. For both tasks, the HP weights were notably different from the reliability weights, with large weights for EL (close to 50%) and semantic analysis (around 20%) and a negative weight for style (−1%). The reliability weights were more homogeneous, especially for PC scores. The article examined how these differences in weights were reflected in reliability, convergent validity, and divergent validity coefficients.

In summary, the only advantage HP scores seem to have over reliability-based weighting is with respect to same-task human-machine correlations, which by definition they were designed to optimize. However, same-task correlations only measure interrater agreement that reflect just one source of error affecting the reliability or scoring accuracy (Haertel, 2006). Test-retest or parallel-form reliability estimates must consider performance across more than a single task. In addition, validity coefficients should also be considered in the evaluation of different scoring methods. As the results of this study show, optimizing same-task correlation may not extend to psychometric advantages in terms of such reliability or validity coefficients. In this study, cross-task human-machine correlations were similar across weighting schemes and reliability-based scores showed an advantage over HP scores for other validity and reliability coefficients.

A limitation of these results in this respect is that the two tasks that form the basis of the writing assessment are not identical. Therefore, although these tasks show similar measurement properties (e.g., their feature variance-covariance matrices are similar, as well as correlations with other scores, see Tables 3 and 4) they cannot be considered true parallel forms. An alternative study that would enable analysis of parallel forms of each of the writing tasks (argument and issue) would examine performance of examinees that repeated the test, an approach taken in Attali (2007a). The limitations of this alternative approach are that repeaters constitute a

relatively small self-selected sample and there is no guarantee that learning had occurred between the tests.

Nevertheless, the results of this study suggest alternative ways of conceptualizing AES. By and large, AES has been conceived (Page, 1966) as a *replacement* for human scoring and the purpose of automated scores has been to *emulate* human scores, with claims that the computer can “simulate” human judgment or “imitate humans” (Page & Petersen, 1995) or even “do better than human raters” (Bereiter, 2003, p. vii). It is not clear whether these claims are directed at human scores or human cognitive processing and decision making. However, as this article shows, this purpose may result in automated scores that are dependent on just one or two features with a poorer representation of the construct. Alternatively, AES could be defined primarily as a *complement* to human scoring (Attali, 2013), realizing that machine processing of text cannot achieve a deep understanding of the essay, and is therefore limited in its measurement of the higher level aspects of the writing construct. The task of mimicking human raters is further complicated by the difficulty to deduce what factors drive human essay evaluations (Eckes, 2008; Lumley & McNamara, 1995).

The results in this article support this proposal. First, feature weights that resulted from a human-prediction process were different from those that emphasized reliability or internal consistency of features, suggesting that the two approaches give rise to different types of machine scores. However, a close examination of the human-prediction weights shows that it places a large weight on EL. It may be tempting to conclude from this fact that human raters too are driven by EL, but this type of reasoning is flawed. Statistical prediction does not entail a causal relation between EL and human scoring. As it is not plausible that human evaluation is driven by EL to that degree, a more reasonable explanation for this result is that it reflects the relative difficulty of the machine to measure higher order writing skills such as organization and development of ideas (which are statistically correlated with EL). In general, as the computer cannot really imitate human evaluation of essays, statistically optimal weights will not necessarily reflect the relative importance of *human*-measured traits. Human-prediction weights were also meant to maximize the validity of machine scores by providing the best prediction of human scores. However, as the results of this article show, this is not the case when human scores are predicted across tasks. Optimal human-prediction weights did not generalize to a different essay students wrote.

How would machine scores that are recognized as complements to human scoring be used in operational settings? In low stakes and instructional applications, machine scores are more likely to be used as the sole evaluation of the essay. In these settings, it is important to recognize the complementary nature of machine evaluation. For example, the Criterion® Online Writing Evaluation service is a web-based writing tool that helps students plan, write, and revise their essays and uses the e-rater technology to provide immediate diagnostic feedback and evaluation of student writing. The purpose of the tool is not, however, to replace teachers, but to help them “by freeing up valuable classroom time to concentrate on higher level writing skills” (Criterion, 2013). In this context, using the human-prediction feature weighting scheme could weaken the purpose of the tool as it might overemphasize features that are indirectly related to the higher level writing skills that teachers tend to. An alternative that could be implemented in this context is to let instructors determine weights (Attali, 2007b). In the absence of such expert-based weights or other external criteria, weights that maximize internal consistency or alternate-form reliability are a reasonable alternative.

In high-stakes applications, machine scores are invariably used in conjunction with human evaluation, either by averaging the machine score with a single human score (e.g., the TOEFL and GMAT assessments) or by using the machine score as a “check” on the human score (the GRE assessment). On the face of it, this context calls for human-prediction weights to maximize

the similarity of the machine and human scores. However, the results of this article challenge this conventional wisdom from a broad measurement perspective. When combined human-machine scores were considered, human-prediction scores showed the lowest reliability and validity coefficients. This seems like an inconsistent result. PC scores have the same cross-task reliability as HP scores (.82) but show lower cross-task correlations with human scores (.65 vs. .68 for Task 1 and .62 vs. .63 for Task 2). Nevertheless, the combined human-PC scores show higher reliability and higher correlations with verbal scores than the combined human-HP scores.

This seeming paradox is resolved when we realize that the lower correlations between human and PC scores actually reflect true differences in the constructs measured. These differences lead to a more robust combined measure that shows higher reliability and validity coefficients. In other words, a broader assessment perspective should lead us to favor greater separation of the evaluation roles that are assigned to humans and to the machine. A “division of labor” approach, where each method focuses on the aspects of writing quality it measures best, could lead to better and broader measurement of the construct. Since with AES it is possible to control construct representation to some degree, we should not suppress these inherent differences between machine and human measurement but instead accept and capitalize on them.

Appendix

Technical Details on the Computation of Feature Weights

This Appendix provides technical details about the computation of the different weighting schemes. The human-prediction (HP) weights are based on the standardized regression coefficients of a single multiple-regression analysis of the 10 computed features of the essay on the human scores of the essay, separately for each writing task. The HP weights in Table 2 are proportional to the standardized regression weight of each feature (i.e., equal to the standardized regression weight divided by the sum of all 10 standardized regression weights of the appropriate task).

The simple reliability (SR) weights are based on the reliability coefficients of the features, correlations between feature scores on the two tasks. For the data analyzed in this article, these reliability coefficients (r) were .47, .47, .47, .48, .62, .49, .64, .77, .31, and .65 (in the same order presented in Table 2), and the SR weights in Table 2 are proportional to $\sqrt{r}/(1 - r)$.

The principal component (PC) weights are based on the feature loadings from an analysis specifying a single factor (in SAS, this would be achieved with the command *proc factor method = prin priors = one nfactors = 1*;). These loadings are also the correlations between the variables and the factor. For Task 1, these loadings were .44, .32, .70, .64, .61, .73, .44, .67, .50, and .72 (in the same order presented in Table 2), and the PC weights in Table 2 are proportional to these loadings.

To illustrate the computation of optimal reliability (OR) weights, a simple example with two components will be used. Assume that the correlation between the two components is .43 and their reliabilities are .85 and .60. Using the notation from McDonald (1968), the standardized component scores are $x' = [x_1, x_2]$, and the component correlation matrix is given by

$$C = \begin{pmatrix} 1 & .43 \\ .43 & 1 \end{pmatrix}. \quad (1)$$

Our goal is to find a set of weights $w' = [w_1, w_2]$, such that a weighted linear combination of the components

$$s = \mathbf{w}'\mathbf{x} \quad (2)$$

will maximize the reliability of s . Note that the variance of s is equal to $\mathbf{w}'\mathbf{C}\mathbf{w}$ and that $\mathbf{C}$ can be decomposed into the sum of true ($\mathbf{C}_1$) and error ($\mathbf{C}_2$) score matrices:

$$\mathbf{C}_1 = \begin{pmatrix} .85 & .43 \\ .43 & .60 \end{pmatrix}, \mathbf{C}_2 = \begin{pmatrix} .15 & 0 \\ 0 & .40 \end{pmatrix}. \quad (3)$$

If we define $\mathbf{M}$ as the square root of $\mathbf{C}_2$, finding the weights that maximize reliability can be achieved (Method III in McDonald, 1968) by solving the following eigenvalue problem:

$$(\mathbf{M}^{-1}\mathbf{C}_1\mathbf{M}'^{-1} - \mathbf{I})\mathbf{M}'\mathbf{w} = 0, \quad (4)$$

where λ is the largest eigenvalue of the (left-hand side) matrix. For the example at hand, λ is 6.31 and the weights derived from solving (4) for $\mathbf{w}$ are proportional to .82 and .18. For these weights, the reliability of s is given by $\lambda/(\lambda + 1)$, or .86 for this example.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Attali, Y. (2007a). *Construct validity of e-rater in scoring TOEFL essays* (Research Report No. 07-21). Princeton, NJ: Educational Testing Service.
- Attali, Y. (2007b). *On-the-fly customization of automated essay scoring* (Research Report No. 07-42). Princeton, NJ: Educational Testing Service.
- Attali, Y. (2011). *A differential word use measure for content analysis in automated essay scoring* (Research Report No. 11-36). Princeton, NJ: Educational Testing Service.
- Attali, Y. (2013). Validity and reliability of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of Automated essay evaluation* (pp. 181-198). New York, NY: Routledge.
- Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment*, 4(3). Retrieved from <http://www.jtla.org>
- Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, 17, 9-17.
- Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment*, 6(2). Retrieved from <http://www.jtla.org>
- Bentler, P. M. (1968). Alpha-maximized factor analysis (Alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika*, 33, 335-345.
- Bereiter, C. (2003). Foreword. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. vii-ix). Mahwah, NJ: Lawrence Erlbaum.
- Burstein, J., Tetreault, J., & Madnani, N. (2013). The E-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 55-67). New York, NY: Routledge.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Criterion. (2013). *About the criterion online writing evaluation service*. Retrieved from www.ets.org/criterion/about

- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25, 155-185.
- Ericsson, P. F., & Haswell, R. H. (2006). *Machine scoring of student essays: Truth and consequences*. Logan: Utah State University Press.
- Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 68-88). New York, NY: Routledge.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education.
- Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, 1, 98-107.
- Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, 12, 54-71.
- McDonald, R. P. (1968). A unified treatment of the weighting problem. *Psychometrika*, 33, 351-381.
- Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, 48, 238-243.
- Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan*, 76, 561-565.
- Powers, D. E. (2005). "Wordiness": A selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses (Research Memorandum No. 04-08). Princeton, NJ: Educational Testing Service.
- Schultz, M. T. (2013). The intellimetric™ automated essay scoring engine: A review and an application to Chinese essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation* (pp. 89-98). New York, NY: Routledge.
- Wang, M., & Stanley, J. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research*, 40, 663-705.
- Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, 93, 179-197.