Research Memorandum
ETS RM–23-01

Exploration of the Proportional Reduction in Mean-Squared Error for Evaluating Automated Scores

Jodi M. Casabianca
Daniel F. McCaffrey
Matthew S. Johnson
Kathryn L. Ricker-Pedley
Ourania Rotou
Joseph Martineau

April 2023

ETS Research Memorandum Series

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 Memorandum 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 Memorandum 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.

Exploration of the Proportional Reduction in Mean-Squared Error for Evaluating Automated Scores

Jodi M. Casabianca, Daniel F. McCaffrey, Matthew S. Johnson,
Kathryn L. Ricker-Pedley, Ourania Rotou, and Joseph Martineau
ETS, Princeton, New Jersey, United States

April 2023

Corresponding author: J. M. Casabianca, Email: JCASABIANCA@ets.org

Suggested citation: Casabianca, J. M., McCaffrey, D. F., Johnson, M. S., Ricker-Pedley, K. L., Rotou, O., & Martineau, J. (2023). *Exploration of the proportional reduction in mean-squared error for evaluating automated scores* (Research Memorandum No. RM-23-01). ETS.

Action Editors: Usama Ali and Sooyeon Kim

Reviewers: Mo Zhang, Brad Moulder, and Jonathan Weeks

Abstract

A key challenge to the use of automated scoring is developing evidence that the resulting scores support the goals and claims of the assessment. The concordance between the automated scores and the human ratings is a critical piece of evidence when the item being scored was designed to be evaluated by a human rater and the automated score is a prediction of a human rating. Current practice for evaluating automated scoring models typically uses the quadratic weighted kappa and the correlation between human and automated scores. This study explores the proportional reduction in mean-squared error (PRMSE) as an alternative to those standard concordance measures. To date, there have been no established rules of thumb or thresholds to apply to the PRMSE in model evaluation. Using empirical data, we explore various conditions by manipulating automated scores to induce changes in the PRMSE and the properties of the final reportable scores. This allows us to link variation in the PRMSE to impact on scores. Based on the results, we establish possible guidelines for using the PRMSE in practice given several factors including the length and stakes of the test.

Keywords: AI scoring, automated scoring, proportional reduction in mean-squared error, evaluation metrics, validity evidence

Acknowledgments

We are grateful to the following colleagues for their helpful reviews of this work: Usama Ali, Sooyeon Kim, Bradley Moulder, Jonathan Weeks, and Mo Zhang.

A key challenge to the use of automated or artificial intelligence (AI) scoring is developing evidence that the resulting scores support the goals and claims of the assessment. For items developed to be scored by human raters and for which the automated scoring model predicts those human ratings, the concordance between the automated scores and the human ratings is the primary source of evidence. Multiple statistics commonly are used to evaluate the concordance. The decision regarding whether the evidence is sufficient to support the operational use of the automated scoring model often depends on whether these statistics exceed thresholds or satisfy rules of thumb. The most widely used statistics are the quadratic weighted kappa (QWK; Fleiss & Cohen, 1973) for the agreement between machine and human scores, the change in the QWK for human and automated scores agreement relative to the QWK for the agreement between two human ratings,[1] and the standardized mean difference (SMD; Cohen, 1988) between the machine and the human scores (Williamson et al., 2012). In practice, based on Williamson et al. (2012), a QWK or correlation of .70 or higher is commonly used as an acceptable level of concordance between the human and automated scores. The logic behind this is that it equates to roughly 50% of the variance in scores being attributed to the automated scoring model versus error. Similarly, Williamson et al. used degradation less than .10 and SMD less than .15, and these criteria also have had wide use in practice.

As an alternative to these traditional concordance statistics, ETS researchers proposed the proportional reduction in mean-squared error (PRMSE) in the prediction of the human rater true scores (Haberman & Qian, 2007; Loukina et al., 2020; Rotou & Rupp, 2020; Wang & Dorans, 2021; Yan & Bridgeman, 2020). The PRMSE has many desirable properties for use in the evaluation of automated scores and scoring model; however, there is no widely accepted criterion for when PRMSE is large enough for an automated scoring model to be used. The purpose of this study is to examine the relationship between PRMSE and impact of item scores to total test scores in the context of an English language proficiency examination. A further goal is to use the results to establish possible rules of thumb for using PRMSE in decisions regarding the implementation of an automated scoring model for operational testing programs.

The Proportional Reduction in Mean-Squared Error

A widely used measure of prediction error is the mean-squared error (MSE). MSE is the expected value of the squared difference between the predicted values (e.g., the automated scores) and the criterion (e.g., the human rating). Let M equal the automated score for a response and H equal a human rating for the same response, then

$$MSE(M:H) = E[(M-H)^2].$$

Let T equal the human rater true score, the expected value of the human ratings for a given response— $T = E[H \mid Response]$. The MSE for predicting the T is $MSE(M:T) = E[(M-T)^2]$. $MSE(M:T) = MSE(M:H) - \sigma^2$, where $\sigma^2 = E[(H-T)^2]$, the error variance in the human ratings. PRMSE equals the proportional reduction to the error in predicting T by using the automated score rather than the overall mean (i.e., using no specific information about the response to predict the true score). The MSE from using the mean, $\mu_T = E[T]$, to predict T is $MSE(\mu_T:T) = E[(\mu_T - T)^2] = var(T)$. Hence,

$$PRMSE(M:T) = \frac{var(T) - MSE(M:T)}{var(T)} = 1 - \frac{MSE(M:T)}{var(T)}.$$

The PRMSE may also be written as

$$PRMSE(M:T) = Cor(M,T)^2 - \frac{(\mu_M - \mu_T)^2}{var(T)} - \frac{[SD(T)Cor(M,T) - SD(M)]^2}{var(T)}, \quad (1)$$

where $\mu_M = E(M)$, $SD(T) = \sqrt{var(T)}$, and $SD(M) = \sqrt{var(M)}$.[2] $Cor(M,T)$ is the correlation between the automated score and the true score. Because $Cor(M,T) = Cor(M,H)/\sqrt{Cor(H_1,H_2)}$, where H_1 and H_2 are ratings from two raters for each response, estimates of $Cor(M,T)$ are often referred to as the disattenuated correlation. From Equation 1 one can see that the PRMSE depends on the correlation between the automated score and the human rater true score. It also depends on the scaling of the automated scores, which does not necessarily equal the scaling of the human ratings. In some applications, scale is unimportant. For example, scaling is unimportant if the automated score is used in an item response theory (IRT) model to produce the score or if automated scores are equated post hoc because in these scenarios the

score is not being combined with human ratings. Also, the scale for automated scores that is optimal for PRMSE is the one that minimizes the squared difference between automated scores and the true score. That scale does not necessarily make the automated scores and true scores closest in other metrics, such as QWK.[3]

The maximum value of PRMSE is 1, which occurs when the automated scores provide perfect predictions of the true scores. PRMSE is on a common scale like QWK. However, unlike the QWK, the PRMSE is not constrained by the quality of the human ratings (Loukina et al., 2019). One downside to using PRMSE to evaluate automated scores is that estimating it requires a sample of responses with multiple human ratings.[4] There are no rules of thumb for values of PRMSE that are large enough to support using an automated scoring model in practice either in place of a human rating or in combination with human ratings. Such rules of thumb exist for other statistics such as QWK or the SMD (the difference between the mean of the automated scores and the mean of the human ratings divided by the standard deviation of the human ratings or the pooled standard deviation). Establishing rules of thumb for PRMSE has proven difficult because there is little or no research on the relationship between the value of PRMSE and impact on test scores and inferences about test takers. Moreover, for some items the PRMSE values conflict with values for other statistics. For example, we observed an item with a low PRMSE (.25), which indicates poor prediction accuracy, but all other statistics had values that would typically indicate that operational use is acceptable (QWK = .86, SMD = .01). Thus, rules of thumb for other concordance statistics do not necessarily apply to PRMSE. To provide information toward the goal of establishing a PRMSE criterion, we conducted a study of the relationship among different subcomponents of the PRMSE, the value of PRMSE, and resulting scale scores.

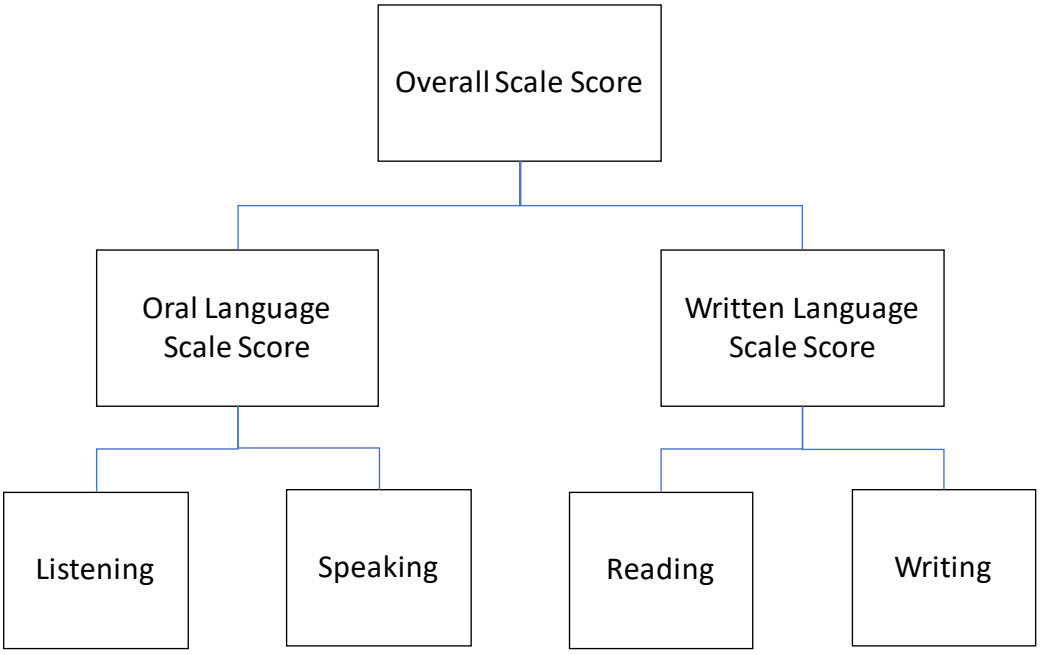
Methods for Exploring the PRMSE

Data

The study used data from an English language assessment. Scores from this assessment are used to evaluate whether students (K–12) need English language development services and determine whether a student's English proficiency has improved. The test has four sections (listening, speaking, reading, writing), which yield an oral language score, a written language

score, an overall scale score, and performance levels for each of the four domains (see Figure 1). We analyzed writing constructed-response (CR) and reading selected-response (SR) data. Scores from both sections are used to compute a composite written language score. The forms analyzed were from a recent administration. There were 26 SR reading items and six CR writing items. Alpha reliability for reading, writing, and the written language composite were .82, .81, and .87, respectively. The CR items consisted of a mixture of extended response formats (scored on a 0- to 4-point scale) and short answer (scored on a 0- to 2-point scale). The data included human ratings and automated scores for the six CR items for 11,532 test takers. A subset of the sample (n = 568) had two human scores because they were in the agreement sample, which contained a random selection of test takers chosen for reliability estimation. Automated scores were from prompt-specific models using either the *e-rater®* (Attali & Burstein, 2006) or *c-rater®* (Leacock & Chodorow, 2003) engine, depending on the task type. Models were built using the first human score.

Figure 1. Test Structure



Note. The test structure includes two selected-response sections (listening and reading) and two constructed-response sections (speaking and writing). Scale scores for oral language, written language, and overall are reported as well as performance levels at these score levels and for each section of the test.

Manipulating the PRMSE Components and Automated Scores

We manipulated the scores for one CR item to degrade the PRMSE in specific ways by changing three subcomponents of the PRMSE: (a) the correlation between the automated scores and the human true score (the disattenuated correlation), (b) the difference in the automated score and human rating means, and (c) the standard deviation of the automated scores relative to the product of the standard deviation of the human true scores and the disattenuated correlation. The correlation between the automated scores and the true scores is determined by the data, the feature inputs algorithm, the statistical or machine learning model used for predicting the human ratings, and the algorithm used to fit the model to the data. This correlation determines the potential for the scores to introduce construct-irrelevant variance into the final scores. Such construct-irrelevant variance could be correlated with other factors such as test-taker group membership. The correlation between automated scores and true scores also determines the potential for the automated scores to underrepresent the construct. The correlation cannot be altered after the model is fitted to the data. The mean and standard deviation of the automated scores can be set post hoc by centering and scaling scores to the desired values. As noted above, in some applications the mean and scale of the scores will not matter. However, in other applications the mean will affect the effective difficulty of the item. Shifting the mean up would make items appear easier, whereas shifting the mean down would make the item appear more difficult assuming no other changes to the scores for the item or the test. Also, when the test scores depend on the sum of the item scores, the scale on the item can effectively change the weight given to the item in the total score. When rescaling by multiplying the item scores by a constant, the constant is effectively a weight. The scaling that maximizes PRMSE might not be optimal in other metrics. For example, post hoc setting of the mean and standard deviation of the automated score to the mean and standard deviation of the human ratings maximizes the QWK.

In the current study, the item scores contribute to the total sum score, which was scaled using the test characteristic curve (TCC). The IRT model combines the two-parameter logistic model (2PL) for dichotomous items and the generalized partial credit (GPC) model for items with more than two score points. Item calibration used automated scores rounded to whole

numbers for the item being manipulated (the only item that used automated scores) and the human ratings for the remaining CR items. The impact of scaling on this model is a priori unclear, so the study explored the impact of different scaling on the total scores as part of the simulation.

To simulate scores with different values for PRMSE, we parameterized PRMSE by the following three factors:

1. $\rho = Cor(M,T)$
2. $\delta = \frac{(\mu_M - \mu_T)}{SD(T)}$
3. $\gamma = \frac{SD(M)}{SD(T)Cor(M,T)}$

By Equation 1, PRMSE equals:

$$PRMSE(M:T) = \rho^2 - \rho^2(1-\gamma)^2 - \delta^2. \quad (2)$$

We can degrade the PRMSE by decreasing ρ or increasing $|\gamma - 1|$ or δ. Our simulations were designed to manipulate all three degradation methods (see Table 1 for details on the conditions). First, to decrease ρ and keep γ and δ fixed, we created simulated automated scores as

$$M^* = [\pi \times \frac{(M-\mu_M)}{SD(M)} + \sqrt{1-\pi^2} \times e] \times S + \mu_M.$$

Here, $0 < \pi < 1$, $S = \pi SD(M)$, and e is a standard normal random variable. The variable manipulated to deflate the correlation is π (we explored values between .7 and .95). The deflated correlation is $\rho^* = \pi \times \rho$ and $100 \times \rho^2(1-\pi^2)[1-(1-\gamma)^2]/PRMSE$ is the percentage change in PRMSE. We set $S = \pi SD(M)$ to keep γ fixed across conditions such that the $\frac{var(M^*)}{[Cor(M^*,T)^2 var(T)]} = \frac{var(M)}{[Cor(M,T)^2 var(T)]}$.

To increase γ, we simulate automated scores as

$$M^* = (M - \mu_M) \times \gamma^* + \mu_M, \text{for } \gamma^* > 0.$$

Table 1. Study Conditions

Condition	PRMSE component(s) manipulated	Description	Equation for manipulated automated score
Human score	None	The first human score (H_1)	n/a
Automated score	None	The automated score (c-rater)	n/a
Reduce *Cor(M, T)*			
pi95	ρ	π = .95; small reduction	$M^* = [\pi \times \frac{(M-\mu_M)}{SD(M)} + \sqrt{(1-\pi^2)} \times e] \times S + \mu_M$
pi90		π = .90	
pi85		π = .85	
pi80		π = .80	
pi75		π = .75	
pi70		π = .70; larger reduction	
Increase mean bias			
delta_x1.10	δ	δ^* = δ x 1.1; 10% increase	$M^* = M + \delta^* \times \sqrt{[Var(H) \times Cor(H_1, H_2)]}$
delta_x1.20		δ^* = δ x 1.2; 20% increase	
delta_x2		δ^* = δ x 2; doubled	
delta_x4		δ^* = δ x 4; quadrupled	
delta_x10		δ^* = δ x 10; times 10	
delta_x20		δ^* = δ x 20; times 20	
delta_x30		δ^* = δ x 30; times 30	
Change scale of automated score			
gamma_plus_0.5	γ	γ^* = γ + .5	$M^* = (M-\mu_M) \times \gamma^* + \mu_M$, for $\gamma^* > 0$
invgamma		γ^* = 1/γ	
Match moments of human raters or true score			
SD match	γ & δ	*SD(M) = SD(H_1)*	$M^* = (M-\mu_M) \times \frac{SD(H_1)}{SD(M)} + \mu_M$
Mean & SD match		*SD(M) = SD(H_1)* and *Mean(M) = Mean(H_1)*	$M^* = (M-\mu_M) \times \frac{SD(H_1)}{SD(M)} + \mu_{H_1}$
Mean & SD match true		*SD(M) = SD(T)* and *Mean(M) = Mean(T) = Mean(H_1)*	$M^* = (M-\mu_M) \times \frac{SD(H_1)\sqrt{Cor(H_1, H_2)}}{SD(M)} + \mu_{H_1}$

Note. PRMSE = proportional reduction in mean-squared error.

One value of interest for γ^* is $1/\gamma$ because it sets the middle term in Equation 2 to zero. Finally, to study the impact of *increasing* δ , we simulate automated scores as

$$M^* = M + \delta^* \times SD(T), \text{ for } \delta^* > 0,$$

and use various factors ranging from 1.1 to 30 to increase δ. The $var(T)$ and $SD(T)$ can be estimated using the fact that $var(T) = var(H) \times Cor(H_1, H_2)$.

In addition to simulating automated scores by individually manipulating the three PRMSE subcomponents, we explored three additional cases to evaluate the impact of scaling on the PRMSE and on scores:

1. The factor γ was changed to set the standard deviation of the automated scores to equal the standard deviation of the human ratings by $M^* = (M - \mu_M) \times \frac{SD(H_1)}{SD(M)} + \mu_M$.
2. Both γ and δ were changed to force the mean and standard deviation of the automated scores to equal the mean and standard deviation of the human ratings. This is a popular scaling method. We simulated automated scores as $M^* = (M - \mu_M) \times \frac{SD(H_1)}{SD(M)} + \mu_{H_1}$.
3. Both γ and δ were changed to force the mean and standard deviation of the automated scores to equal the mean and standard deviation of the human true scores. The mean of the human true scores equals the mean of the human ratings, and the standard deviation of the true scores equals the standard deviation of the human ratings times the square root of the interrater correlation, $Cor(H_1, H_2)$. We simulated automated scores as $M^* = (M - \mu_M) \times \frac{SD(H_1)\sqrt{Cor(H_1,H_2)}}{SD(M)} + \mu_{H_1}$.

Study Design

We explored the relationship between the value of PRMSE for one item and the final reported scores to evaluate the robustness of the reported scores to the degraded PRMSE. We selected a writing item that had a high PRMSE in our data (.986, based on unrounded automated scores), a correlation between human ratings of .823 (based on the agreement sample), and a high disattenuated correlation between the automated score and human ratings

(.993, based on the agreement sample). The item prompted test takers to describe a picture, was scored on a 0–2 scale, and had mean (standard deviation) human scores of 1.43 (0.63) and 1.40 (0.65) by the first human and the second human, respectively. The mean (unrounded) automated score from the c-rater engine was 1.42 (0.57).

Table 1 provides a list of conditions explored in this study. The first condition uses human scores as a baseline condition. The second condition uses the actual automated score as another baseline condition. The next set of conditions separated by horizontal lines (pi95–pi70) are conditions that manipulate the disattenuated correlation ρ by varying π. Following this set are sets of conditions (separated by horizontal lines) that manipulate δ or γ and then three cases that manipulate the scaling as described above. For each of the 18 conditions, we produced scale scores based on the full test, including the SR items. In addition, we also created subtests to study the impact of the manipulated CR item's relative contribution to the overall score. Thus, we have three sets of results for 18 conditions: (a) based on the actual test composition (six CRs, 26 SRs), (b) based on a shortened CR section (three CRs, 26 SRs), and (c) based on shortened CR and SR sections (three CRs, 13 CRs). In the full-length test, the possible score range was 0 to 43 (twenty-six 1-point SR items, two 4-point CR items, one 3-point CR item, and three 2-point CR items) where the CR item being manipulated contributes 4.65% to the total possible score. In the test with the shortened CR section, the possible score range was 0 to 36 (twenty-six 1-point SR items, two 4-point CR items, and one 2-point CR item) and the manipulated CR item contributes 5.56% to the total possible score. In the test with the shortened SR and CR sections, the possible score range was 0 to 23 (thirteen 1-point SR items, two 4-point CR items, and one 2-point CR item) and the manipulated CR item contributes 8.70% to the total possible score.

Score Estimation

We computed a written language score (see Figure 1) and determined test takers' overall performance levels based on various altered versions of the original automated score by manipulating the components as described in Table 1. All other item scores were kept the same. We approximated the operational scoring procedures as outlined in the test's technical report to compute the final scale scores for written language. That is, we implemented the

following steps: (a) calibrated item scores from the reading and writing sections using the 2PL/GPC model, (b) performed a Stocking and Lord linking (Stocking & Lord, 1983) to transform the item parameter estimates to the base scale,[5] (c) used the item scores and transformed item parameter estimates from the IRT calibration to find each test taker's latent trait estimate via the TCC scoring method, (d) applied scaling constants to the latent traits (intercept = 512.12, slope = 38.87), and (e) added 1,000 to the rescaled latent trait value to place the final score on a 1,150 to 1,950 scale. The final score is the scale score for written language. We classified each test taker into one of four performance levels using the established cut points for the scale scores (1,529 for Level I/II, 1,595 for Level II/III, and 1,646 for Level III/IV). We performed these analyses for each of the conditions in Table 1, calibrating each time to estimate a new set of item parameters to define the TCC. The automated scores were rounded before being used in the IRT calibration. Note that while we approximated the operational scoring procedures, we make no claims that the scores are representative of an operational evaluation.

Outcome Measures

To evaluate the impact of manipulating different components of the PRMSE, we computed mean differences and root mean-squared differences per condition, comparing the scale scores from using the baseline (human or automated score) and the manipulated automated score. At the item level, we computed summary statistics, PRMSE, and QWK and examined score distributions. At the scale score level, we estimated reliability (using alpha) and conditional standardized errors of measurement (CSEM). Finally, we examined the agreement between the performance levels based on human score and performance levels based on the manipulated automated scores.

Results

Table 2 provides item-level summary statistics for the manipulated item. To understand how study conditions impacted prediction accuracy, Table 2 also provides the PRMSE, QWK, and the three components making up PRMSE (ρ, δ, and γ). The π conditions relate to a degradation in the disattenuated correlation ρ, with π being the factor by which the correlation is reduced.

Table 2. Item Score Distributions and PRMSE Components for Each Study Condition

Condition	Item score distribution			Item statistics		ρ	δ	γ	PRMSE	QWK
	0	1	2	Mean	SD					
Human score	.07	.42	.51	1.44	0.63				n/a	1.000
Automated score	.07	.42	.51	1.44	0.62	.993	0.022	1.014	.986	.896
Reduce *Cor*(*M*, *T*)										
pi95	.06	.43	.51	1.45	0.60	.940	0.022	1.019	.882	.843
pi90	.05	.46	.49	1.44	0.59	.889	0.035	1.025	.788	.791
pi85	.04	.48	.48	1.44	0.57	.838	0.030	1.027	.701	.737
pi80	.03	.51	.46	1.43	0.55	.778	0.038	1.037	.603	.672
pi75	.02	.54	.44	1.42	0.54	.729	0.042	1.045	.528	.617
pi70	.01	.56	.42	1.41	0.52	.684	0.054	1.046	.464	.565
Increase mean bias										
delta_x1.10	.07	.42	.51	1.45	0.62	.993	0.000	1.014	.986	.897
delta_x1.20	.07	.42	.51	1.45	0.62	.993	-0.002	1.014	.986	.897
delta_x2	.07	.42	.51	1.45	0.62	.993	-0.017	1.014	.986	.897
delta_x4	.06	.42	.52	1.45	0.61	.993	-0.057	1.014	.983	.895
delta_x10	.06	.42	.52	1.46	0.60	.993	-0.175	1.014	.955	.884
delta_x20	.04	.42	.53	1.49	0.58	.993	-0.373	1.014	.847	.844
delta_x30	.03	.42	.56	1.53	0.55	.993	-0.570	1.014	.661	.783
Change scale of automated score										
gamma_plus_0.5	.11	.37	.52	1.40	0.68	.993	-0.005	1.559	.678	.851
invgamma	.07	.42	.51	1.45	0.62	.993	0.024	0.977	.985	.893
Match moments of human raters or true score										
SD match	.08	.41	.51	1.44	0.63	.993	0.019	1.082	.979	.900
Mean & SD match	.08	.41	.51	1.44	0.63	.993	-0.007	1.082	.980	.900
Mean & SD match true	.07	.42	.51	1.45	0.62	.993	-0.003	0.988	.986	.894

Note. The PRMSE, PRMSE components, and QWK are based on unrounded automated scores. PRMSE = proportional reduction in mean-squared error; QWK = quadratic weighted kappa.

The δ conditions here relate to large increases in the SMD between the simulated automated scores and the human score. The γ conditions are a function of the ratio of the automated score standard deviation to the product of the true score standard deviation and the disattenuated correlation (between the machine and human scores). Looking down each of the columns for ρ, δ, and γ shows the different values that manifested from the manipulation of the automated scores. This item is scored on a 0–2 scale, with most responses getting a 1 or 2. There were noticeable differences in the score distributions for the different π conditions, especially when π dropped from .80 to .70. The standard deviation of scores was also lower in these conditions when compared to the human score statistics. Note that the γ values change slightly with changes to π—this is due to the sampling error. Similarly, because of sampling error, the estimates of the correlation between the automated scores and true scores are lower than what we would expect theoretically (ρ x π). For example, for π = .9, the estimate of ρ^* = .940, where theory suggests $\rho^* = \rho \times \pi$ = .993 x .95 = .9433.

The actual PRMSE for this item was .986. There were a few conditions for which the PRMSE was much lower: delta_x20, delta_x30, gamma_plus_0.5, and all six of the π conditions. There were not many differences among the δ conditions, but when sufficiently increasing δ (see delta_x20, delta_x30) both the PRMSE and QWK were noticeably degraded and the summary statistics were distinctly different from the other δ conditions and the human/machine conditions. In terms of γ, the condition with the large γ^* ($\gamma^* = \gamma + 0.5$; gamma_plus_0.5) resulted in more variable automated scores by design. The PRMSE was reduced to .678 in this condition.

Table 3 provides the mean differences and the RMSD between scale scores based on the human scores and simulated automated scores and based on the actual automated scores and the simulated automated scores, for each condition and test length combination. To help with interpreting the results, results for conditions that yield PRMSE less than .70 are bolded.[6] Treating the statistics for the actual automated score condition as a baseline (second row), we compared the extent to which the other conditions led to much larger differences and also determined the impact of test length. For all the test lengths, especially the shorter tests, when

PRMSE was below .70, using the automated scores instead of human ratings had a notable impact on the mean difference of the scores, the RMSD, or both. For the full test, the π = .70 to .90 conditions had RMSDs ranging from about 5.38 to 7.39 (as shown in the column labeled with "H" for human score). These are substantially larger than the RMSDs for all the other conditions, which were all around 2.50 (including the RMSD for the actual automated score). In the conditions with the reduced CR section, the RMSD for the actual automated score was 4.00, and most conditions were similar except for π = .70 to .90 (RMSD ranged from 6.60 to 9.65) and delta_x30 (RMSD equals 4.57). Finally, in the reduced test length, the RMSD for the actual machine was 7.92, and it was only the π conditions (specifically, .70 to .90) that resulted in much larger differences. These results support the conclusion that the disattenuated correlation is very important in its influence on the PRMSE and how that translates to impact at the scale score level. The largest RMSD of 16.72 (in the reduced test length, π = .70 condition) is greater than 20% of the scale-score standard deviation of 72. This is a nontrivial difference in scores. There were even large RMSDs in the π = .85 and π = .90 conditions, which had PRMSEs of .701 and .788, respectively. These differences were exacerbated in the shorter tests. However, for the delta_x30 and gamma_plus_0.5, which had PRMSE < .70, their RMSDs were still similar to the original machine.

The results also show that the impact increases as the item's contribution to the scale score increases (as test length decreases). For the shorter test, these differences were more than double the differences based on the actual automated score. When PRMSE was below .70, differences between scale scores based on simulated scores and the human score ranged from about 11% to 23% of the scale-score standard deviation (standard deviations ranged from 60 to 61 for the full test, 62 to 65 for the test with the reduced CR section, and 65 to 72 for the reduced test).

We estimated scale reliability of the written language section using Cronbach's alpha reliability and CSEM. Cronbach's reliability estimates as well as the average CSEMs were mostly comparable within test length (see Table 4). For example, alpha was around .875 for almost all conditions for the full-length test. As the test length decreased, the differences in alpha across conditions increased to about .015 in the π conditions. The same was true for the average

CSEMs—they were between 32.19 and 32.96 for the full test and between 37.24 and 37.84 and 45.99 and 47.43 for the two reduced test length conditions.

Finally, in a comparison of the performance levels for the written language section based on the human score and on the (manipulated) automated scores, we found that rates of exact agreement were very high for all conditions and test lengths (see Table 5). There was a marked impact on the performance levels on the reduced length test when PRMSE was below .70. While some of these conditions (with PRMSE < .70) yielded acceptable levels of agreement with scores based on human ratings (for example, the gamma_plus_0.5 condition where PRMSE=.678 had almost 99% exact agreement with performance levels) others with similar values of PRMSE did not. For example, the delta_x30 condition where PRMSE = .661 had 92.17% exact agreement with performance levels—this would impact many students in the school system.

Discussion and Implications

The current practices for the evaluation of automated scoring models in large-scale assessment contexts are heavily based on Williamson et al. (2012). The PRMSE is different from the QWK in that it is a direct measure of prediction accuracy, with respect to the true score. However, given the widely accepted threshold of .70 for the QWK and correlation, there has been a natural inclination to apply the same threshold to the PRMSE. Indeed, as we have established, there is a relationship between the PRMSE and the correlation between the human and automated scores. To what extent using the same threshold for the PRMSE is reasonable and effective is unclear. It is also unclear how the PRMSE should be used in different use contexts. Furthermore, different conditions, such as test length, impact the reported scores based on automated scores, thereby possibly necessitating adaptation in the standards for how we use the PRMSE in evaluation. This is especially critical to consider as the proposed applications for automated scoring expand with different engines, contexts, and assumptions.

In this exploration, we used a single item with a very high PRMSE to explore factors that can impact PRMSE and the relationship between the value of PRMSE and impact on test scores.

Table 3. Mean Differences and Root Mean-Squared Differences in Written Language Scale Scores Comparing the Simulated Automated Scores to Both the Human Score and Actual Automated Score

Condition	Full test length: 26 SRs, 6 CRs				Reduced CR section: 26 SRs, 3 CRs				Reduced test length: 13 SRs, 3 CRs			
	Mean difference		RMSD		Mean difference		RMSD		Mean difference		RMSD	
	M	H	M	H	M	H	M	H	M	H	M	H
Human score	-0.02	0.00	2.50	0.00	0.05	0.00	4.00	0.00	0.12	0.00	7.92	0.00
Automated score	0.00	0.02	0.00	2.50	0.00	-0.05	0.00	4.00	0.00	-0.12	0.00	7.92
Reduce *Cor(M, T)*												
pi95	-0.20	-0.19	1.51	2.69	-0.19	-0.24	1.83	4.18	-0.21	-0.33	2.94	8.14
pi90	-0.35	-0.33	4.95	5.38	-0.10	-0.15	5.53	6.60	0.07	-0.05	9.56	11.88
pi85	-0.59	-0.57	6.72	7.01	-0.13	-0.19	8.47	9.16	0.06	-0.06	12.30	14.17
pi80 [a]	**-0.81**	**-0.79**	**7.13**	**7.39**	**-0.11**	**-0.16**	**9.02**	**9.65**	**0.28**	**0.16**	**14.69**	**16.08**
pi75 [a]	**-1.07**	**-1.05**	**6.83**	**7.07**	**-0.25**	**-0.30**	**7.88**	**8.55**	**0.17**	**0.05**	**14.72**	**15.22**
pi70 [a]	**-1.44**	**-1.41**	**6.53**	**6.65**	**-0.43**	**-0.48**	**8.05**	**7.79**	**0.03**	**-0.10**	**16.74**	**16.72**
Increase mean bias												
delta_x1.10	-0.03	-0.01	0.32	2.50	-0.04	-0.09	0.38	4.00	-0.05	-0.17	0.61	7.93
delta_x1.20	-0.03	-0.01	0.32	2.50	-0.04	-0.09	0.38	4.00	-0.05	-0.17	0.62	7.93
delta_x2	-0.05	-0.03	0.43	2.50	-0.07	-0.12	0.52	4.01	-0.09	-0.21	0.83	7.93
delta_x4	-0.10	-0.08	0.62	2.51	-0.14	-0.19	0.76	4.02	-0.18	-0.30	1.16	7.93
delta_x10	-0.25	-0.23	0.98	2.54	-0.33	-0.38	1.23	4.06	-0.42	-0.55	1.83	7.95
delta_x20	-0.59	-0.57	1.55	2.66	-0.79	-0.84	2.01	4.23	-1.00	-1.12	3.03	8.08
delta_x30 [a]	**-1.06**	**-1.04**	**2.16**	**2.90**	**-1.43**	**-1.48**	**2.89**	**4.57**	**-1.82**	**-1.94**	**5.22**	**8.04**
Change scale of automated score												
gamma_plus_0.5 [a]	**0.51**	**0.52**	**1.36**	**2.58**	**0.62**	**0.57**	**1.80**	**4.14**	**0.78**	**0.66**	**2.68**	**8.19**
invgamma	-0.06	-0.04	0.44	2.50	-0.07	-0.12	0.52	4.00	-0.09	-0.21	0.85	7.92
Match moments of human raters or true score												
SD match	0.09	0.11	0.54	2.51	0.10	0.05	0.67	4.01	0.12	-0.00	0.99	7.95
Mean & SD match	0.06	0.08	0.52	2.51	0.06	0.01	0.62	4.01	0.07	-0.05	0.93	7.95
Mean & SD match true	-0.08	-0.06	0.50	2.50	-0.10	-0.15	0.60	4.01	-0.13	-0.25	0.94	7.92

Note. The mean difference and RMSD (root mean-squared difference) are based on the differences in scale scores for either the human score (H) or actual automated score (M) and the simulated automated scores based on the different study conditions. CR = constructed response; SR = selected response.

[a] Indicates rows, which are in bold, that have PRMSE < .70.

Table 4. Cronbach's Alpha Reliability and Conditional Standardized Errors of Measurement for Written Language Score Scale Statistics by Test Length

Condition	PRMSE	Full test length: 26 SRs, 6 CRs		Reduced CR section: 26 SRs, 3 CRs		Reduced test length: 13 SRs, 3 CRs	
		Alpha	Average CSEM	Alpha	Average CSEM	Alpha	Average CSEM
Human score	n/a	.875	32.78	.841	37.71	.800	47.30
Automated score	.986	.875	32.79	.841	37.71	.801	47.33
Reduce *Cor(M, T)*							
pi95	.882	.875	32.70	.840	37.62	.800	47.23
pi90	.788	.874	32.53	.838	37.48	.797	46.96
pi85	.701	.873	32.38	.836	37.37	.794	46.64
pi80 [a]	**.603**	**.872**	**32.26**	**.835**	**37.27**	**.791**	**46.28**
pi75 [a]	**.528**	**.872**	**32.19**	**.834**	**37.24**	**.790**	**45.99**
pi70 [a]	**.464**	**.871**	**32.31**	**.832**	**37.51**	**.787**	**46.08**
Increase mean bias							
delta_x1.10	.986	.875	32.79	.841	37.70	.801	47.33
delta_x1.20	.986	.875	32.79	.841	37.70	.801	47.33
delta_x2	.986	.875	32.78	.841	37.70	.801	47.32
delta_x4	.983	.875	32.77	.841	37.69	.800	47.32
delta_x10	.955	.875	32.73	.841	37.66	.800	47.29
delta_x20	.847	.875	32.62	.840	37.57	.798	47.19
delta_x30 [a]	**.661**	**.874**	**32.47**	**.839**	**37.44**	**.797**	**47.01**
Change scale of automated score							
gamma_plus_0.5 [a]	**.678**	**.876**	**32.96**	**.843**	**37.84**	**.803**	**47.43**
invgamma	.985	.875	32.77	.841	37.69	.801	47.31
Match moments of human raters or true score							
SD match	.979	.876	32.83	.842	37.74	.801	47.36
Mean & SD match	.980	.876	32.82	.842	37.73	.801	47.35
Mean & SD match true	.986	.875	32.77	.841	37.69	.800	47.31

Note. The alpha is based on the rounded automated scores, but values are almost equivalent to the alpha based on unrounded scores.
CR = constructed response; CSEM = conditional standardized errors of measurement; PRMSE = proportional reduction in mean-squared error; SR = selected response.

[a] Indicates rows, which are in bold, that have PRMSE < .70.

Table 5. Written Language Performance Level Classifications and Exact Agreement With Performance Level Classifications Based on Human Score Condition, by Test Length

Condition	Full test length: 26 SRs, 6 CRs					Reduced CR section: 26 SRs, 3 CRs					Reduced test length: 13 SRs, 3 CRs				
	1	2	3	4	Total exact	1	2	3	4	Total exact	1	2	3	4	Total exact
Human score	33.53	40.89	17.99	7.59	100.00	36.21	40.97	15.54	7.28	100.00	36.72	38.05	18.87	6.37	100.00
Automated score	33.61	40.84	17.90	7.65	99.32	36.17	41.03	15.44	7.37	99.24	36.68	38.03	18.84	6.45	98.81
Reduce *Cor(M, T)*															
pi95	33.59	40.83	17.95	7.63	99.15	36.13	41.09	15.48	7.30	98.96	36.64	38.10	18.85	6.41	98.58
pi90	33.70	40.89	17.95	7.46	98.50	36.21	41.13	15.46	7.21	98.53	36.78	38.27	18.74	6.21	97.59
pi85	33.71	40.98	17.90	7.40	98.19	36.37	41.15	15.40	7.08	97.98	36.67	38.44	18.98	5.91	96.74
pi80 [a]	**33.65**	**41.13**	**18.03**	**7.19**	**97.51**	**36.16**	**41.39**	**15.53**	**6.91**	**97.45**	**36.53**	**39.05**	**18.65**	**5.76**	**95.84**
pi75 [a]	**33.85**	**40.99**	**18.11**	**7.06**	**97.29**	**36.40**	**41.32**	**15.52**	**6.76**	**96.90**	**30.31**	**45.53**	**18.45**	**5.71**	**90.45**
pi70 [a]	**33.85**	**41.13**	**18.07**	**6.94**	**96.80**	**36.40**	**41.38**	**15.63**	**6.59**	**96.43**	**30.22**	**45.84**	**18.56**	**5.38**	**90.03**
Increase mean bias															
delta_x1.10	33.60	40.85	17.90	7.65	99.33	36.16	41.03	15.44	7.37	99.25	36.68	38.01	18.87	6.45	98.81
delta_x1.20	33.60	40.85	17.90	7.65	99.33	36.16	41.03	15.44	7.37	99.25	36.68	38.01	18.87	6.45	98.81
delta_x2	33.59	40.86	17.90	7.65	99.34	36.14	41.06	15.44	7.37	99.24	36.66	38.03	18.87	6.45	98.82
delta_x4	33.56	40.88	17.90	7.65	99.35	36.12	41.06	15.45	7.37	99.21	36.64	38.04	18.88	6.45	98.82
delta_x10	33.53	40.89	17.92	7.66	99.36	36.08	41.09	15.45	7.38	99.21	36.55	38.08	18.91	6.45	98.82
delta_x20	33.41	40.95	17.93	7.71	99.26	35.97	41.15	15.47	7.41	99.14	36.30	38.24	18.99	6.46	98.63
delta_x30 [a]	**33.24**	**40.98**	**18.05**	**7.74**	**99.14**	**35.73**	**41.28**	**15.53**	**7.45**	**98.97**	**29.53**	**44.87**	**19.12**	**6.47**	**92.17**
Change scale of automated score															
gamma_plus_0.5 [a]	**33.66**	**40.82**	**17.88**	**7.64**	**99.29**	**36.25**	**40.95**	**15.43**	**7.37**	**99.19**	**36.79**	**37.91**	**18.85**	**6.45**	**98.73**
invgamma	33.60	40.85	17.90	7.65	99.33	36.16	41.03	15.44	7.37	99.23	36.66	38.05	18.84	6.45	98.82
Match moments of human raters or true score															
SD match	33.62	40.84	17.90	7.65	99.32	36.20	41.01	15.43	7.37	99.22	36.70	38.00	18.85	6.45	98.78
Mean & SD match	33.62	40.84	17.90	7.65	99.32	36.18	41.03	15.43	7.37	99.22	36.69	38.00	18.86	6.45	98.80
Mean & SD match true	33.60	40.85	17.90	7.65	99.33	36.15	41.04	15.44	7.37	99.24	36.65	38.04	18.87	6.45	98.82

Note. The percentages under the columns labeled 1 to 4 are from the score distributions from the marginal table for each simulated condition. The "Total exact" column is the sum of the diagonal in the performance level crosstabulation (comparing the simulated condition to the human scoring condition). CR = constructed response; SR = selected response.

[a] Indicates rows, which are in bold, that have PRMSE < .70.

We added noise or changed the mean and scale of the scores as to reduce the PRMSE from the very high value for the original, unaltered scores for the item. We could then observe how the final test scores, which equaled the sum of the scores on all the items, rescaled using the TCC, changed as a function of PRMSE. Impact on the final scores does not fully assess the meaning of PRMSE for evaluating automated scores. For one reason, final scores should be mostly insensitive to the score on a single item, even if the scores for that item fail to meet the claims of the item or the test. We explored this issue by using subsets of items to simulate shorter tests, giving the single item greater weight in the final score. More generally, little impact on the final score does not directly provide evidence related to the validity claims about content nor demonstrate that the scores capture the construct rather than other factors. However, limited impact can suggest that, for purposes of interpreting the final scores, the scores based on automated scores will have similar results to scores using human ratings. Hence, inferences and interpretations appropriate for the scores based on the human rating should be similar to those from scores derived using automated scoring.

Although the study is limited by the use of a single item, there are important patterns in the results that may provide lessons about PRMSE. These patterns can help with more general interpretations of this statistic when used to evaluate automated scores.

First, the results show that PRMSE is relatively insensitive to deviations in the mean and variance of the automated scores from the values that maximize PRMSE (the mean equal to the mean of the human ratings and the variance equal to the variance of the human rating times the squared correlation between the human ratings and the automated scores). Even though PRMSE is a quadratic function of mean difference and scale differences, PRMSE changes very little until the differences are large relative to what occurs with observed scores. For example, it is not until the SMD (delta) is 20 times the value for the original scores that the PRMSE shows any appreciable decline. This is due, in part, to the fact that automated score means are typically close to the mean of the human rating relative to the variance in the human true scores. Similarly, rescaling the scores to change the variance (gamma) has very limited effect.

Second, relatedly, rescaling so the automated scores match both the mean and standard deviation of the human ratings or the human true scores had almost no effect on the final

scores. The PRMSE remained very high. Scaling the scores to match the mean and standard deviation of the human ratings results in a PRMSE of .980, very close to the original value of .986 and very close to the maximum value of .987 (which equals the square of the disattenuated correlation between human ratings and the automated scores). Moreover, this sort of rescaling would have almost no effect on the item statistics and would have almost no effect on IRT model item parameters for the studied item. Note, if the automated scores had a lower correlation with the human true scores, the variance of the automated scores would most likely be more compressed relative to the variance of the human ratings or true scores than what is observed with our data. Consequently, in examples where the automated scores have a relatively low correlation with the human ratings, changing the scale to match the variance of the human ratings or true scores would tend to have greater impact on the total score because the implicit weight of the item would be increased more than in this example.

Third, PRMSE is sensitive to changes in correlation between the automated score and the human true score. Although, holding γ and δ constant, PRMSE is a quadratic function of the correlation, PRMSE changes almost linearly with the correlation in the range of values considered in this exploration. Importantly, the range of values in this exploration is similar to the range of values found in many operational applications of automated scoring. Hence, these patterns demonstrate that the disattenuated correlation drives PRMSE. Unlike the mean and variance of the automated scores, which can be tuned post hoc to values that will optimize PRMSE, the correlation cannot be changed post hoc. Instead, the correlation depends on the information contained in the inputs to the prediction model (the human ratings and the engine features).

The fourth notable pattern in the results is that at very high values of PRMSE, such as 0.98 or even 0.90, using the automated scores rather than human ratings had very little impact on scale scores. The value of PRMSE where the impact on scores changes from trivial to something more notable depended on the overall length of the test. This is intuitive because the item has greater contribution to the overall score when the test is shorter and any changes to the item score distribution will have a larger impact on the total score distribution. With shorter tests, the threshold for PRMSE needed to be higher to keep the score effectively

exchangeable with the corresponding scores based on human rating for the item. With the very short test, there was nontrivial impact with PRMSE around .90 and even some impact with PRMSE as high as 0.98. The impact of a single item on the total score for a short test is similar to the impact of multiple items on a longer test. Hence, if multiple items are scored automatically, a higher value of PRMSE for each item might be needed to keep impact on scores trivial.

As mentioned earlier, there is a tendency to appeal to the guidelines of Williamson et al. (2012) and apply a threshold of .70 for the PRMSE. In addition, .70 is often considered a minimally desirable value for human rater reliability, which is roughly equal to the square of the correlation between the human ratings and the true score. Because the squared correlation between the automated scores and the true score is a key component of PRMSE, .70 is a threshold worthy of consideration when interpreting the results. Further, it might be justifiable to use .70 as threshold for PRMSE. As discussed in the Results section, use of automated scores with PRMSE below .70 results in notable impact on the scale scores. Consequently, when PRMSE is below .70, there is the potential for inferences and outcomes based on scale scores that use the automated scores to differ from scale scores using the human ratings. Evidence in support of human ratings might not apply to the automated scores. Claims supported by the human ratings might not hold for the automated scores because the two rating methods might yield notably different results for some test takers. Additional evidence would be needed to support the use of automated scores such as correlations with some other external criteria linked to the construct. Thus, .70 might be considered a minimum level: We might avoid using scores with PRMSE below .70 unless there is additional evidence to support their use.

The interpretation of PRMSE above .70 is less clear from this analysis. The deviations from human scores are sensitive to the length of the test. This risk for deviations from scores based on human ratings increases with the share of the total score contributed by the automated scores. In addition, unless the PRMSE is very high (above .95), there is opportunity for automated scores to introduce construct-irrelevant variance that is not in the human ratings but is related to other factors, leading to unintended outcomes or unfair scores. Lower values of PRMSE allow for more variance that is unassociated with the human ratings. When the PRMSE is below .70, the potential for construct-irrelevant variance specific to the automated

scores is very great and, as noted previously, evidence in support of the use of automated scores beyond their prediction accuracy is needed. For values between .70 and roughly .95, other considerations such as the contribution to the total score, the use context, consequences for the test taker, and other evidence in support of the claims must be part of the evaluation of the automated scores in addition to prediction accuracy. For more specific guidelines and a deeper discussion of the PRMSE and selected thresholds for evaluating automated scoring models, consult *Best Practices for Constructed-Response Scoring* (ETS, 2021).

Taken together, these results suggest that if the use of automated scores is to rely heavily on both evidence in support of human ratings and the fact that automated scores are good predictors of human ratings, then the PRMSE would need to be very high. This is especially true if the automated scores have a large contribution to the total score. In the range of many applications, PRMSE is most sensitive to the correlation between automated scores and human true scores, so this correlation needs to be high. Scaling the mean and variance to maximize PRMSE will often make little difference on PRMSE, based on the results of this study (and other explorations), but it could matter more for the final scores.

Of note is the method in which we performed the scaling and linking of the test scores. Our goals were not to reproduce the true operational process fully. For convenience, there were some deviations that may have yielded summaries of scale scores and performance levels that appear different from what is seen operationally. For example, although we performed a Stocking and Lord linking (Stocking & Lord, 1983) to place the item parameter estimates on the base scale, the process by which this was conducted for the study did not exactly match what is done operationally. There were differences in the linking designs. Because we linked a single manipulated item, all of the other items could all be anchors; however, operationally, field test items are linked using a smaller set of anchor items. Keeping the scoring procedures consistent across conditions permits a comparison of scores within the study. The results are relevant for understanding the impact of different conditions on the PRMSE for both this particular test and other applications of automated scoring applications.

References

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

ETS. (2021). *Best practices for constructed-response scoring*. https://www.ets.org/content/dam/ets-org/pdfs/about/cr_best_practices.pdf

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, *33*(3), 613–619. https://doi.org/10.1177/001316447303300309

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

Leacock, C., & Chodorow, M. (2003). *C-rater*®: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389–405. https://doi.org/10.1023/A:1025779619903

Loukina, A., Madnani, N., Cahill, A., Yao, L., Johnson, M. S., Riordan, B., & McCaffrey, D. F. (2020). Using PRMSE to evaluate automated scoring systems in the presence of label noise. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 18–29). Association for Computational Linguistics. https://aclanthology.org/2020.bea-1.2

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1–10). Association for Computational Linguistics. https://aclanthology.org/W19-4401

Rotou, O., & Rupp, A. A. (2020). *Evaluations of automated scoring systems in practice* (Research Report No. RR-20-10). ETS. https://doi.org/10.1002/ets2.12293

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement*, *7*(2), 201–210. https://doi.org/10.1177/014662168300700208

Wang, W., & Dorans, N. J. (2021). *Impact of categorization and scaling on classification agreement and prediction accuracy statistics* (Research Report No. RR-21-26). ETS. https://doi.org/10.1002/ets2.12339

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Yan, D., & Bridgeman, B. (2020). Validation of automated scoring systems. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 297–318). Chapman and Hall/CRC.

Notes

[1] This difference in QWK values is referred to as the "degradation in the QWK," even though the QWK for human and machine scores is not always lower than the QWK for two human raters.

[2] Because $Cor(M,H) = Cor(M,T) \times Cor(H,T)$ and $var(T) = var(H) \times Cor(H,T)^2$, Equation 1 can be used to easily show that $PRMSE(M:H) = PRMSE(M:T) \times Cor(H,T)^2$.

[3] QWK is maximized when $\mu_M = \mu_T$ and $var(M) = var(T)$.

[4] PRMSE can be estimated using the mean of the squared differences between the automated scores and the human ratings to estimate $MSE(M:H)$ and a sample of responses with multiple human rating to estimate σ^2 and $var(T)$. See Haberman and Qian (2007) for details.

[5] To perform the scaling, we used a nonequivalent groups with an anchor test design to link the two tests. The anchor in this case was the set of all items on each test excluding the manipulated CR item. The transformed IRT parameters established by the program were used as the base parameters to put the newly calibrated "item" on the base scale.

[6] The cutoff of .70 was chosen because it is used as a threshold for QWK, so it is important to explore whether it appears to be a useful threshold for PRMSE.