Using PRMSE to evaluate automated scoring systems in the presence of label noise

Anastassia Loukina, Nitin Madnani, Aoife Cahill
Lili Yao, Matthew S. Johnson, Brian Riordan, Daniel F. McCaffrey
{aloukina,nmadnani,acahill}@ets.org
lili.yao@gmail.com, {msjohnson,briordan,dmccaffrey}@ets.org
Educational Testing Service, NJ, USA

Abstract

The effect of noisy labels on the performance of NLP systems has been studied extensively for system *training*. In this paper, we focus on the effect that noisy labels have on system *evaluation*. Using automated scoring as an example, we demonstrate that the quality of human ratings used for system evaluation have a substantial impact on traditional performance metrics, making it impossible to compare system evaluations on labels with different quality. We propose that a new metric, proportional reduction in mean squared error (PRMSE), developed within the educational measurement community, can help address this issue, and provide practical guidelines on using PRMSE.

1 Introduction

NLP systems are usually trained and evaluated using human labels. For automated scoring systems, these would be scores assigned by human raters. However, human raters do not always agree on the scores they assign (Eckes, 2008; Ling et al., 2014; Davis, 2016; Carey et al., 2011) and the inter-rater agreement can vary substantially across prompts as well as across applications. For example, in the ASAP-AES data (Shermis, 2014), the agreement varies from Pearson's r=0.63 to r=0.85 across "essay sets" (writing prompts) .

In many automated scoring studies, the data for training and evaluating the system are randomly sampled from the same dataset, which means that the quality of human labels may affect both system training and evaluation. Notably, the effect of label quality on training and evaluation may not be the same. Previous studies (Reidsma and Carletta, 2008; Loukina et al., 2018) suggest that when annotation noise is relatively random, a system trained on noisier annotations may perform as well as a system trained on clean annotations. On the other hand, noise in the human labels used for evaluation can have a substantial effect on the estimates of system performance even if the noise is random.

In this paper, our focus is the effect of noise in human labels on system *evaluation*. How do we compare two systems evaluated on datasets with different quality of human labels? While there exist several public data sets that can be used to benchmark and compare automated scoring systems, in many practical and research applications the scoring systems are customized for a particular task and, thus, cannot be evaluated appropriately on a public dataset. As a result, the research community has to rely on estimates of system performance to judge the effectiveness of the proposed approach. In an industry context, the decision to deploy a system is often contingent on system performance meeting certain thresholds which may even be codified as company- or industry-wide standards.

A typical solution to the problem of different human-human agreement across evaluation datasets is to use human-human agreement itself as a baseline when evaluating a system (Shermis, 2014). In this case, the system can be evaluated either via a binary distinction (did its performance reach human-human agreement?) or by looking at the differences in agreement metrics as measured between two humans and between a single human and the machine, known as "degradation" (Williamson et al., 2012). Yet how do we interpret these numbers? Is a system that exceeds a human-human agreement of r=0.4 on one dataset better than another that performs just below a human-human agreement of r=0.9 on a *different* dataset?

In this paper, we use simulated data to demonstrate that the rate of human-human agreement has a substantial effect on estimates of system performance, making it difficult to compare systems that are evaluated on different datasets. We also show that this problem cannot be resolved by simply looking at the difference between human-human

Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, pages 18–29

and machine-human agreement. We then show that one possible solution is to use proportional reduction in mean squared error (PRMSE) (Haberman, 2008), a metric developed in the educational measurement community, which relies on classical test theory and can adjust for human error when computing estimates of system performance.

2 Related work

The effect of noisy labels on machine learning algorithms has been extensively studied in terms of their effect on system training in both general machine learning literature (see, for example, Frénay and Verleysen (2014) for a comprehensive review), NLP (Reidsma and Carletta, 2008; Beigman Klebanov and Beigman, 2009; Schwartz et al., 2011; Plank et al., 2014; Martínez Alonso et al., 2015; Jamison and Gurevych, 2015) and automated scoring (Horbach et al., 2014; Zesch et al., 2015).

One key insight that emerged from such work is that the nature of the noise is extremely important for the system performance. Machine learning algorithms are greatly affected by systematic noise but are less sensitive to random noise (Reidsma and Carletta, 2008; Reidsma and op den Akker, 2008). A typical case of random noise is when the labeling is done by multiple annotators which minimizes the individual bias introduced by any single annotator. For example, in a study on crowdsourcing NLP tasks, Snow et al. (2008) showed that a system trained on a set of non-expert annotations obtained from multiple annotators outperformed a system trained with labels from one expert, on average.

The studies discussed so far vary the model training set, or training regime, or both while keeping the evaluation set constant. Fewer studies have considered how inter-annotator agreement may affect system *evaluation* when the training set is held constant. These studies have shown that in the case of evaluation, the label quality is likely to have a substantial impact on the estimates of system performance even if the annotation noise is random.

Reidsma and Carletta (2008) used simulated data to explore the effect of noisy labels on classifier performance. They showed that the performance of the model, measured using Cohen's Kappa, when evaluated against the 'real' (or gold-standard) labels was higher than the performance when evaluated against the 'observed' labels with added random noise. This is because for some instances, the classifier's predictions were correct, but the 'observed' labels contained errors.

Loukina et al. (2018) used two different datasets to train and evaluate an automated system for scoring spoken language proficiency. They showed that training an automated system on perfect labels did not give any advantage over training the system on noisier labels, confirming previous findings that automated scoring systems are likely to be robust to random noise in the data. At the same time, the choice of evaluation set led to very different estimates of system performance *regardless of what data was used to train the system*.

Metrics such as Pearson's correlation or quadratically-weighted kappa, commonly used to evaluate automated scoring systems (Williamson et al., 2012; Yannakoudakis and Cummins, 2015; Haberman, 2019), compare automated scores to observed human scores without correcting for any errors in human scores. In order to account for differences in human-human agreement, these are then compared to the same metrics computed for the human raters using measures such as "degradation": the difference between human-human and human-machine agreement (Williamson et al., 2012).

In this paper, we build on findings from the educational measurement community to explore an alternative approach where estimates of system performance are corrected for measurement error in the human labels. Classical test theory (Lord and Novick, 1968) assumes that the human holistic score is composed of the test's true score and some measurement error. A "true" score is defined as the expected score over an infinite number of independent administrations of the test. While such true scores are latent variables, unobservable in real life, their underlying distribution and measurement error can be estimated if a subset of responses is scored by two independently and randomly chosen raters. Haberman (2008); Haberman et al. (2015); Haberman and Yao (2015); Yao et al. (2019a,b); Zhang et al. (2019) proposed a new metric called proportional reduction in mean squared error (PRMSE) which evaluates how well the machine scores predict the *true* score, after adjusting for the measurement error. The main contribution of this paper is a further demonstration of the utility of this metric in the context of automated scoring. Outside of educational measurement, a similar approach has been been explored in pattern recognition by Lam and Stork (2003), for example, who used estimated error rates in human labels to adjust

performance estimates.

We further explore how agreement between human raters affects the *evaluation* of automated scoring systems. We focus on a specific case where the human rating process is organized in such a way that annotator bias is minimized. In other words, the label noise can be considered *random*. We also assume that the scores produced by an automated scoring system are on a continuous scale. This is typical for many automated scoring contexts including essay scoring (Shermis, 2014), speech scoring (Zechner et al., 2009) and, to some extent, content scoring (Madnani et al., 2017a; Riordan et al., 2019) but, of course, not for all possible contexts: for example, some of the SemEval 2013 shared tasks on short answer scoring (Dzikovska et al., 2016) use a different scoring approach.

3 Simulated data

In this paper, we use simulated gold-standard (or "true") scores, human scores and system scores for a set of 10,000 responses. Since "true" scores are not available for real data, using simulated data allows us to compare multiple raters and systems to the known ground-truth.[1] We focus on evaluation only and make no assumptions about the quality of the labels in the training set or any other aspects of system training. The only thing we know is that different human raters and different systems in our data set assign different scores and have different performances when evaluated against true scores.

As our gold-standard, we use a set of *continuous* scores simulated for each response and consider these to be the correct "true" score for the response. Note that the continuous nature of gold-standard scores allows us to capture the intuition that some responses fall between the ordinal score points usually assigned by human raters. To create such gold-standard scores, we randomly sampled 10,000 values from a normal distribution using the mean and standard deviation of human scores observed in a large-scale assessment (mean=3.844, std=0.74). Since the scores in the large-scale assessment we use as reference varied from 1 to 6, the gold-standard scores below 1 and above 6 were also truncated to 1 and 6 respectively.

Next, we simulated scores from 200 human raters for each of these 10,000 "responses". For each rater, its score for a response was modeled as the gold-standard score for the response plus a random error. We model different groups of raters: with low (inter-rater correlation r=0.4), moderate (r=0.55), average (r=0.65) and high (r=0.8) agreement. The correlations for different categories were informed by correlations we have observed in empirical data from various studies. The errors for each rater were drawn from a normal distribution with a mean of 0. We chose the standard deviation values used to sample the errors in order to create 4 categories of 50 raters, each defined by a specific average inter-rater correlation. Since in most operational scenarios, human raters assign an integer score, all our simulated human scores were rounded to integers and truncated to lie in $[1, 6]$, if necessary. Table 1 shows the correlations between the simulated human rater scores within each category.

Category	# raters	HH-corr	mean	std
Low	50	0.40	3.83	1.14
Moderate	50	0.55	3.83	0.99
Average	50	0.65	3.83	0.91
High	50	0.80	3.83	0.83

Table 1: A description of the 4 categories of simulated human raters used in this study. The table shows the label of each category, the number of raters in the category, the average correlation between pairs of raters within the category, and the mean and standard deviation of the scores assigned by raters in the category.

For each response, we also simulated 25 automated scores. Like human scores, automated scores were simulated as gold-standard scores plus random error. We chose the standard deviation values used to sample the random errors so as to obtain specific levels of performance against the gold-standard scores: the worst system had a Root Mean Squared Error (RMSE) of 0.74 score points while the best system had an error of 0.07 score points. Since the interpretation of RMSE depends on the score scale, we chose these values as the percentage of gold-standard score variance.

Table 2 summarizes different automated systems simulated for this study. We created 5 categories of systems with 5 systems in each category. For the worst systems ("poor"), the mean squared error was equal to the variance of gold-standard scores (R^2=0). In other words, in terms of scoring error, a system from the "poor" category performed no

[1] cf. Reidsma and Carletta (2008); Yannakoudakis and Cummins (2015) who also used simulated data to model system evaluation.

better than a constant.[2] For the best system (from the "perfect" category), the mean squared error was only 0.1% of gold-standard score variance with the system achieving an R^2 of 0.99. The systems *within* each category were very close in terms of performance as measured by mean squared error but the actual simulated scores for each system were different. These simulated systems will help evaluate whether performance metrics can both differentiate systems with different performance and correctly determine when two systems have similar performance.

Category	N	R^2 (GS)	r (GS)	r ('Average')
Poor	5	0.01	0.71	0.57
Low	5	0.40	0.79	0.64
Medium	5	0.65	0.86	0.69
High	5	0.80	0.91	0.74
Perfect	5	0.99	1.00	0.80

Table 2: A description of the 5 categories of simulated systems used in this study. The table shows the label of each category, the number of systems in the category, the average R^2 of the systems within the category, and the r when evaluating the systems in the category against the gold-standard scores ("GS"). The last column shows the average correlation of the systems' scores with simulated rater scores from the "Average" category.

To summarize, the final simulated dataset consisted of 10,000 "responses". Each response had 1 "gold-standard" score, 200 "human" scores and 25 "system" scores. [3]

4 Problems with traditional metrics

4.1 Rating quality and performance

We first considered how the quality of human labels affects the estimates of the metrics that are typically used to evaluate automated scoring engines. For the analyses in this section, we used the scores from one of our simulated systems from the "High" system category (R^2 with gold-standard scores = 0.8). We then randomly sampled 50 pairs of simulated raters from each rater category and evaluated the human-machine agreement for each pair. We used both the score from the first rater in the pair as well as the average of the the two rater scores in the pair as our reference score and computed four metrics: Pearson's r[4], quadratically-weighted kappa (QWK)[5], R^2, and degradation (correlation between the scores of the two humans minus the correlation between scores of our chosen system and the reference human score). Figure 1 shows how these metrics *for the same system* vary depending on the human agreement in the evaluation dataset.

As the figure shows, the estimates of performance *for the same set of scores* vary drastically depending on the quality of human ratings whether we use the score from the first human rater or the average of the two scores. For example, estimates of correlation vary from mean r = 0.69 when computed against the average scores of two raters with low agreement to r = 0.86 when computed against the average score of two raters with high agreement. The difference between r = 0.69 and r = 0.86 is considerable and, at face value, could influence both deployment decisions in an industry context as well as conclusions in a research context. Yet all it actually reflects is the amount of noise in human labels: both correlations were computed using the *same* set of automated scores. Looking at degradation does not resolve the issue: the degradation in our simulation varied from -0.05 to -0.30. It is obvious that the metrics improve when the human-human agreement goes from low to high, regardless of which metric is used, and do not provide a stable estimate of model performance. This pattern is consistent across different sets of automated scores.

4.2 Rating quality and ranking

Given how much the estimates of system performance vary depending on the quality of human ratings, it is clear that the quality of human ratings will also affect the comparison between different systems if they are evaluated on different datasets.

To demonstrate this, we randomly sampled 25 pairs of simulated raters with different levels of human-human agreement, the same as the number of simulated systems in our data, and "assigned" a different pair to each system. Each pair of raters

[2] $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$ where y_i are the observed values (human scores), $\hat{y}_i$ are the predicted values and $\bar{y}$ is the mean of observed score. R^2 standardizes the MSE by the total variance of the observed values leading to a more interpretable metric that generally varies from 0 to 1, where 1 corresponds to perfect prediction and 0 indicates that the model is no more accurate than simply using mean value as the prediction.

[3] The data and the code are publicly available at https://github.com/EducationalTestingService/prmse-simulations. We encourage the readers to use this code to run further simulations with varying input parameters.

[4] We use raw correlation coefficients, not z-transforms, as is the norm in automated scoring literature.

[5] QWK for continuous scores was computed cf. Haberman (2019) as implemented in RSMTool (Madnani et al., 2017b)

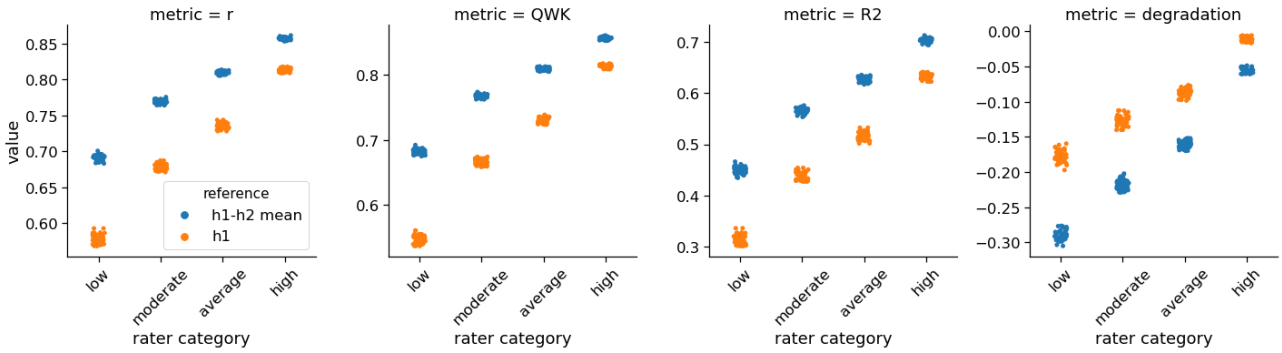

Figure 1: The effect of human-human agreement on the evaluation results for the same set of automated scores against either the first human rater or the average of two human raters. Note that the metrics are on different scales.

is always sampled from the same rater category but different systems are evaluated on pairs from different rater categories. Thus, for example, 3 of 5 systems in the "low" system category were evaluated against rater pairs with "high" agreement, while the remaining two systems in that category were evaluated against rater pairs with "average" agreement. At the same time, for "medium" category systems, 3 out of 5 systems were evaluated on raters with "low" agreement (see also Table 1 in the Appendix). This simulation was designed to mimic, in a simplified fashion, a situation where different research studies might evaluate their systems on datasets with *different* quality of human ratings [6].

We then evaluated each system against their assigned rater pairs using the standard agreement metrics and ranked the systems based on each of the metrics. The results are presented in the first four subplots in Figure 2.[7] For comparison, we also evaluated the systems against a single pair of raters from the "average" rater category, i.e., using the *same* rater pair for each system. The system ranking when systems are evaluated against this same rater pair are shown as red dots. The figure shows that when different systems are evaluated against the same pair of raters, their ranking is consistent with what we know to be the correct ranking in our simulated dataset. However, when different systems are evaluated against *different* pairs of raters, their ranking can vary depending on the quality of the ratings and the chosen metric. All metrics - except degradation - correctly ranked the worst performing systems (in the "poor" system category), but they could **not** reliably differentiate between the other categories. In our simulated dataset, we see substantial overlaps in R^2 between systems in the "medium", "high", and "perfect" system categories, with even larger overlaps for other metrics.

Notably, when rater quality differs across the datasets used to evaluate a system, the degradation between human-human and system-human agreement, a common way to control for differences in said rater quality, does not always provide accurate system rankings. In our simulated dataset, based on degradation, some of the systems from the "perfect" system category ranked lower than some of the systems from the "medium" system category.

4.3 What if we had more than two raters?

Figure 1 showed that evaluating system scores against the average of two raters leads to higher estimates of agreement than when the system is evaluated against a single rater. This is not surprising: in our simulated dataset, the rater error is modeled as random and averaging across several simulated raters means that errors can cancel out when the number of raters is sufficiently large. In fact, we expect that evaluating the system against the average of *multiple* raters should provide performance estimates close to the known performance against the gold-standard scores. In this section, we simulated a situation where each response is scored by up to 50 raters.

For each category of raters, we randomly ordered the raters within this category and computed the cumulative average score of an increasing number of raters. We then evaluated the same system from the "high" system category used in §4.1 against this cumulative average score. The results are presented in Figure 3. The red lines indicate the values when evaluating the system's performance against the

[6] Note that the random assignment between rater categories and systems is a key aspect of this simulation since we are exploring a situation where the system performance is *independent* of the quality of human labels used to evaluate such systems.

[7] The last subplot will be explained in §5.2.

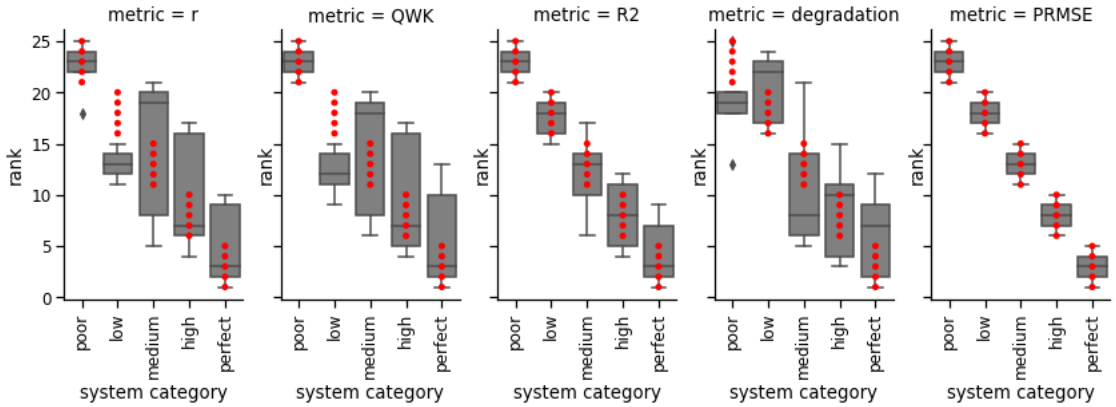

Figure 2: The ranking of systems from different categories when evaluated against randomly selected pairs of raters with different human-human agreement levels. The X axis shows the known ranking of the simulated systems in terms of their performance measured against the gold-standard scores. The red dots show the ranking when the systems are evaluated against the *same* pair of raters.

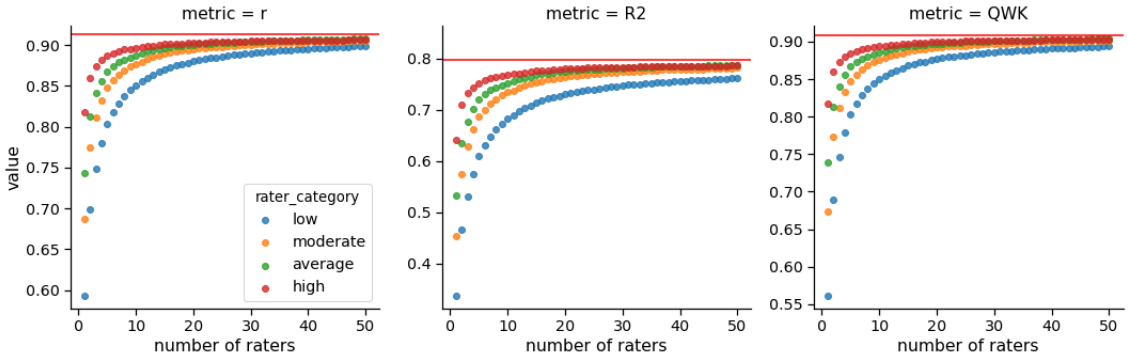

Figure 3: The effect of number of raters on several common metrics. Each plot shows a different metric computed for a randomly chosen system in our dataset against an increasing number of human raters. The red line indicates the metric value computed against the gold-standard scores & different colors indicate different rater categories.

gold-standard scores. As expected, for all rater categories, the performance estimates for the system approach the known gold-standard performance as the number of raters increases.

5 PRMSE with reference to true scores

The simulations in the previous sections demonstrate that the values of metrics usually used to evaluate automated scoring systems are directly dependent on the quality of human ratings used to evaluate the system. In fact, the effect of human label quality can be so large such that two identical systems may appear drastically different while the performance of two very different systems may appear very similar. One possible solution is to collect additional ratings for each response from multiple raters as we showed in §4.3. This solution is likely to be too expensive to be feasible: for example, in our simulated dataset, we would need to collect at least 10 additional ratings for each response in order to obtain stable estimates of system performance, more if the rater agreement is low.

The solution we propose comes from the educational measurement community and draws on test theory methods to adjust the system performance estimates for measurement error.

5.1 The definition of PRMSE

The main idea behind PRMSE is to evaluate the automated scores against the true scores rather than the observed human scores. Classical test theory assumes that the human label H consists of the true score T and a measurement error E and $\mathrm{Var}(H) = \mathrm{Var}(T) + \mathrm{Var}(E)$. While it is impossible to compare system scores to the latent true scores *for each individual response*, it is possible to use the variability in human ratings to estimate the rater error and to compute an overall measure of agreement between automated scores and true scores after subtracting the rater error from the vari-

ance of the human labels.

Just like R^2, PRMSE relies on the concepts of mean squared error (MSE) and proportional reduction in mean squared error (hence PRMSE), but in this case, these measures are computed between the automated score M and *the true score* T instead of the human label H, where $\text{MSE} = E(M-T)^2$ and $\text{PRMSE} = 1 - \frac{\text{MSE}}{\text{Var}(T)}$.

Also similar to R^2, PRMSE is expected to fall between 0 and 1. A value of 0 indicates that system scores explain none of the variance of the true scores, while a value of 1 implies that system scores explains all the variance of true scores. In general, the higher the PRMSE, the better the system scores are at predicting the true scores.

We provide a detailed derivation for PRMSE in the Appendix. A Python implementation of PRMSE is available in RSMTool in the `rsmtool.utils.prmse` module[8].

5.2 PRMSE and human-human agreement

In this section, we show how PRMSE can help address the issues discussed in §4. We first considered the case where the same system is evaluated against ratings of different quality. As shown in §4.1, *all* traditional metrics of system performance are affected by human-human agreement and, therefore, estimates for these metrics vary depending on which pair of raters is used to evaluate the system. Therefore, in this section, we only compare PRMSE to R^2.

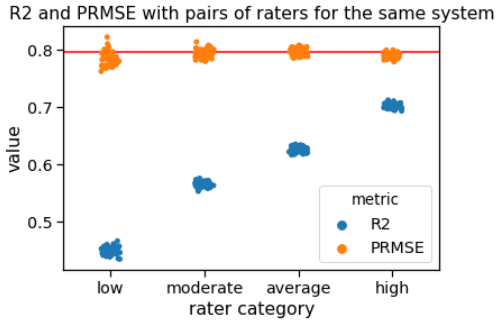

Figure 4: R^2 with average human score and PRMSE for the same system when evaluated against human ratings with different levels of agreement. The red line shows the value of R^2 when evaluating system performance against gold-standard scores.

We used the same pairs of raters and the same systems as in §4.1 to compute PRMSE and then compared its values to the values of R^2 for the same pair of raters. Both these metrics rely on comparing the mean prediction error to the variance of gold-standard scores. For R^2, the gold-standards scores are the *observed* human-assigned scores that are available and can be used for computation. The gold-standard scores for PRMSE are the *latent* true scores that cannot be used directly: the metric is instead computed using the observed human scores and the estimates of rater variance as explained in the previous section.[9] Figure 4 shows the values of R^2 when evaluating the same system against different categories of human raters and the values of PRMSE for the same evaluations. While R^2, as we have already seen, varies between 0.43 and 0.71 depending on the quality of human ratings, PRMSE remains relatively stable between 0.76 and 0.82. We also note that the values of PRMSE are centered around the R^2 between system scores and gold-standard scores (0.8 in this case), as expected.

Next, we considered whether PRMSE can help obtain stable system rankings when systems are evaluated against human ratings with different qualities. We used the same combinations of simulated rater pairs and systems as in §4.2 and computed PRMSE for each system and rater pair. We then ranked the systems based on their PRMSE values. The results are presented in the last subplot in Figure 2. The figure shows that even though different systems were evaluated against human ratings of different quality, their final ranking based on PRMSE was consistent with the *known* correct ranking based on the gold-standard scores.

In summary, PRMSE is more robust to the quality of human ratings used for system evaluation and can reliably rank systems regardless of the quality of human labels used to evaluate them.

5.3 PRMSE and double-scoring

In §5.2, we considered a situation where all responses are double-scored. In reality, often only a subset of responses has several scores available to compute inter-rater agreement. The formula for PRMSE presented in the Appendix also allows us to compute PRMSE in such a situation: in this case, the variance of human errors is computed using *only* the double-scored responses. The prediction error

[8] https://rsmtool.readthedocs.io/en/stable/api.html#prmse-api

[9] Although the true scores are known in our simulation, the values of PRMSE in this and the following sections are computed using *observed* human scores only following the formulas in the Appendix, *without* using the simulated true scores.

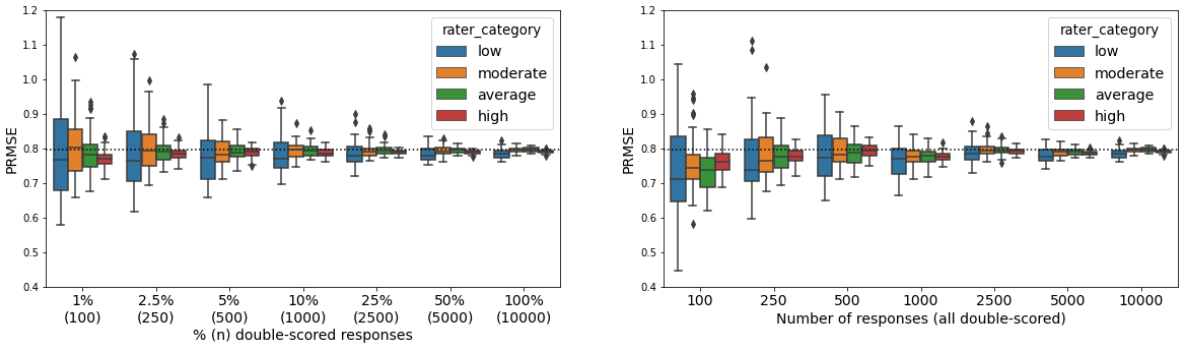

Figure 5: The distribution of PRMSE values depending on the percentage (left) or number (right) of double-scored responses. Different colors indicate levels of inter-rater agreement, i.e, rater category. The dotted line shows the known R^2 against gold-standard scores. Some PRMSE values for N=100 and "low" agreement were around 1.6 and are omitted for clarity. PRMSE values > 1 indicate that sample size is too small to reliably estimate error variance.

and variance are computed using *all* responses in the sample and either the average of two scores when available or the single available score. The numbers are adjusted for the percentage of the total number of ratings available for each response.

To test how PRMSE values depend on the percentage of double scored responses, we randomly sampled 50 pairs of raters from each rater category and created, for each of these 200 pairs, 7 new datasets each with a different percentage of double-scored responses. We then computed PRMSE for a randomly selected system from the "high" category for each of these 1,400 datasets. To check whether it is the *percentage* of double-scored responses that matters or the *number* of double-scored responses, we also computed a second PRMSE value over *only* the double-scored responses available in each case. For example, when simulating the scenario where we only have 10% of the responses double-scored, we compute two PRMSE values: (a) over the full dataset (10,000 responses) with 10% (1,000) double-scored and 90% (9,000) single-scored responses and (b) over a smaller dataset that *only* includes the 1,000 double-scored responses. The results are shown in Figure 5 (see also Table 2 in the Appendix). These results show that PRMSE values are much more stable with a larger number of double-scored responses and what matters is the total *number* of double-scored responses, not their percentage in the sample. There is substantial variability in PRMSE values when the number of double-scored responses is low, especially when computed on human ratings with low inter-rater agreement. In our simulated experiments, consistent values of PRMSE (to the first decimal) were achieved with 1,000 responses if the quality of human ratings was moderate-to-high. More responses would be necessary to reliably estimate PRMSE with low inter-rater agreement.

6 Discussion

The performance of automated systems is often lower on data with lower human-human agreement. While this may mean that responses harder to score for humans are also harder to score for machines, our analyses show that this is not always true. Furthermore, since subsets of the same dataset are often used for both system training and evaluation, separating the effect of noisy labels on *training* from that on *evaluation* may be impossible.

In this paper, we showed that even for the *same* set of automated scores, estimates of system performance depend directly on the the quality of the human labels used to compute the agreement metrics. We also showed that using standard performance metrics to compare two systems may be misleading if the systems are evaluated against human scores with different inter-rater agreements. Comparing system performance to human-human agreement using degradation does not resolve this issue.

We proposed that a new metric, PRMSE, developed within the educational measurement community for evaluation is an effective way to obtain estimates of system performance that are adjusted for human-human agreement. PRMSE provides system evaluation against 'true' scores, thus making it possible to compare different systems on the same scale and offering a performance metric that is robust to the quality of human labels.

We emphasize that PRMSE does not affect the evaluation results when the systems are evaluated on the *same* set of human labels, for example, in

the context of a shared task or a benchmark dataset. However, it can help compare system performance across studies as well as within studies, for example, when the dataset includes multiple items with varying levels of human-human agreement in their respective human scores.

The theory behind PRMSE makes certain assumptions about the nature of the rater error: it is assumed to be random with a mean of 0 and finite variance. Furthermore, the rater error is assumed to be independent of the item and its true score. There are several steps one can take to make sure the data meets these assumptions. For example, a standard way to randomize rater error is to set up the scoring process in a way such that multiple raters each score a different set of responses. Furthermore, one should additionally check whether human ratings have similar mean and variance. We note that other models discussed in the NLP literature (see §2), made other assumptions, for example that noisier labeling is more likely for some items ("hard" cases) than others. The performance of PRMSE under such conditions remains subject for future studies.

Finally, while PRMSE can adjust estimates of system performance for human error, it does not fully address the issue of different datasets. Users of automated scoring still need to use their judgement – or additional extrinsic criteria – to decide whether two systems can be deemed comparable.

7 Practical guidelines for PRMSE

We conclude with guidelines for using PRMSE.

- PRMSE estimates of system performance are robust to human-human agreement and can be used to compare systems across datasets.
- PRMSE computation assumes that the rating process is set up to randomize rater error: e.g. even if most responses only have a single score, the scoring process should involve multiple raters each scoring a different set of responses to minimize individual rater bias.
- Both sets of human ratings used to estimate PRMSE should have similar mean and variance and similar agreement with system scores.
- Responses selected for double-scoring must be a random sample of all responses.
- We recommend a total of at least 1000 double-scored responses to reliably estimate the human error. For human-human correlations > 0.65, a smaller sample (such as 500) might suffice. PRMSE values above 1 indicate that the double-scored sample is too small.
- PRMSE should be used in combination with other metrics of human-machine agreement.

Acknowledgments

We thank Beata Beigman Klebanov, Oren Livne, Paul Deane and the three anonymous BEA reviewers for their comments and suggestions that greatly improved this paper.

References

Beata Beigman Klebanov and Eyal Beigman. 2009. From Annotator Agreement to Noise Models. *Computational Linguistics*, 35(4):495–503.

Michael D. Carey, Robert H. Mannell, and Peter K. Dunn. 2011. Does a Rater's Familiarity with a Candidate's Pronunciation Affect the Rating in Oral Proficiency Interviews? *Language Testing*, 28(2):201–219.

Larry Davis. 2016. The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1):117–135.

Myroslava O. Dzikovska, Rodney D. Nielsen, and Claudia Leacock. 2016. The joint student response analysis and recognizing textual entailment challenge: making sense of student responses in educational applications. *Language Resources and Evaluation*, 50(1):67–93.

Thomas Eckes. 2008. Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2):155–185.

Benoît Frénay and Michel Verleysen. 2014. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869.

Shelby J. Haberman. 2008. When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33:204–229.

Shelby J. Haberman. 2019. Measures of Agreement Versus Measures of Prediction Accuracy. *ETS Research Report Series*, 2019(1):1–23.

Shelby J. Haberman and L. Yao. 2015. Repeater analysis for combining information from different assessments. *Journal of Educational Measurement*, 52:223–251.

Shelby J. Haberman, L. Yao, and S. Sinharay. 2015. Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology*, 68:363–385.

Andrea Horbach, Alexis Palmer, and Magdalena Wolska. 2014. Finding a Tradeoff between Accuracy and Rater's Workload in Grading Clustered Short Answers. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 588–595.

Emily K. Jamison and Iryna Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of EMNLP 2015*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

Chuck P. Lam and David G. Stork. 2003. Evaluating classifiers by means of test data with noisy labels. *IJCAI International Joint Conference on Artificial Intelligence*, pages 513–518.

Guangming Ling, Pamela Mollaun, and Xiaoming Xi. 2014. A Study on the Impact of Fatigue on Human Raters when Scoring Speaking Responses. *Language Testing*, 31:479–499.

Frederic M. Lord and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Addison Wesley, Reading, MA.

Anastassia Loukina, Klaus Zechner, James Bruno, and Beata Beigman Klebanov. 2018. Using exemplar responses for training and evaluating automated speech scoring systems. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nitin Madnani, Anastassia Loukina, and Aoife Cahill. 2017a. A Large Scale Quantitative Exploration of Modeling Strategies for Content Scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 457–467, Copenhagen, Denmark. Association for Computational Linguistics.

Nitin Madnani, Anastassia Loukina, Alina Von Davier, Jill Burstein, and Aoife Cahill. 2017b. Building Better Open-Source Tools to Support Fairness in Automated Scoring. In *Proceedings of the First Workshop on ethics in Natural Language Processing, Valencia, Spain, April 4th, 2017*, pages 41–52, Valencia. Association for Computational Linguistics.

Héctor Martínez Alonso, Barbara Plank, Arne Skjærholt, and Anders Søgaard. 2015. Learning to parse with IAA-weighted loss. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1357–1361, Denver, Colorado. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting subjective annotations. In *COLING 2008 workshop on Human Judgments in Computational Linguistics*, pages 8–16, Manchester, UK.

Dennis Reidsma and Jean Carletta. 2008. Reliability Measurement without Limits. *Computational Linguistics*, 34(3):319–326.

Brian Riordan, Michael Flor, and Robert Pugh. 2019. How to account for mispellings: Quantifying the benefit of character representations in neural content scoring models. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 663–672, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark D. Shermis. 2014. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20:53–76.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 254–263, Stroudsburg, PA, USA. Association for Computational Linguistics.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. 2012. A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of Automated Text Scoring systems. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223.

Lili Yao, Shelby J. Haberman, and Mo Zhang. 2019a. Penalized best linear prediction of true test scores. *Psychometrika*, 84 (1):186–211.

Lili Yao, Shelby J. Haberman, and Mo Zhang. 2019b. Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors. ETS Research Report RR-19-13, ETS, Princeton, NJ.

Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. *Speech Communication*, 51(10):883–895.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015. Reducing Annotation Efforts in Supervised Short Answer Scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132, Denver, Colorado.

Mo Zhang, Lili Yao, Shelby J. Haberman, and Neil J. Dorans. 2019. Assessing scoring accuracy and assessment accuracy for spoken responses. In *Automated Speaking Assessment*, pages 32–58. Routledge.

A The distribution between system and rater categories

The table below shows how systems from different categories were assigned to different pairs of raters.

	Human-human agreement			
System	Low	Moderate	Average	High
Poor	1	3	0	1
Low	0	0	2	3
Medium	3	0	1	1
High	2	1	1	1
Perfect	2	0	2	1

Table 3: The distribution between different systems and different pairs of raters. The table shows how many systems from each system category were evaluated using pairs of raters from different rater categories.

B Deriving the PRMSE formula

Let

- N denote the total number of responses in the evaluation set
- c_i denote the number of human ratings for response i,
- H_{ij} denote human rating $j = 1, \ldots, c_i$ for response i, and
- $\bar{\text{H}}_i = \frac{1}{c_i} \sum_{j=1}^{c_i} H_{ij}$ denote the average human rating for response i.
- $\bar{\text{H}} = \frac{\sum_i c_i \bar{\text{H}}_i}{\sum_i c_i}$ denote the average of all human ratings.
- Let M_i denote the predicted score for response i.

The true human score model assumes a hypothetical infinite population/sequence of human raters that could score responses and assumes that the raters a response actually receives are an unbiased sample from this population. The raters H_{ij} are assumed to have the same error variance and the errors e_{ij} are uncorrelated. The model defines the true human score by

$$T_i = \lim_{c_i \to \infty} \frac{1}{c_i} \sum_{j=1}^{c_i} Y_{ij} = E[H_{ij}] \tag{1}$$

and the error ϵ_{ij} as $\epsilon_{ij} = H_{ij} - T_i$, or stated differently $H_{ij} = T_i + \epsilon_{ij}$.

B.1 Estimating the error variance

If we have only two ratings per response then we estimate the error variance by recognizing

$$V_\epsilon = \frac{1}{2} E[(H_{i2} - H_{i1})^2] \tag{2}$$

which can easily be estimated with the unbiased estimator

$$\hat{V}_\epsilon = \frac{1}{2N} \sum_{i=1}^{N} (H_{i2} - H_{i1})^2 \tag{3}$$

When we have more than two raters, the variance of rater errors is computed as a pooled variance estimator. We first calculate the within-subject variance of human ratings V_i for each response i using denominator $c_i - 1$:

$$V_i = \frac{\sum_{j=1}^{c} (H_{i,j} - \bar{H}_i)^2}{c_i - 1} \tag{4}$$

We then take a weighted average of those within-responses variances:

$$\hat{V}_\epsilon = \frac{\sum_{i=1}^{N} V_i * (c_i - 1)}{\sum_{i=1}^{N} (c_i - 1)} \tag{5}$$

B.2 Estimating true score variance

An unbiased estimator of the true score variance is

$$\hat{V}_T \equiv \widehat{\text{Var}}(T) = \frac{\sum_{i=1}^{N} c_i (\bar{\text{H}}_i - \bar{\text{H}})^2 - (N-1)\hat{V}_\epsilon}{c_\cdot - \frac{\sum_{i=1}^{N} c_i^2}{c_\cdot}} \tag{6}$$

where $c_\cdot = \sum_{i=1}^{N} c_i$ is the total number of observed human scores.

B.3 Estimating mean squared error

We estimate the mean squared error of the automated scores M_i with the following unbiased estimator.

$$\widehat{\text{MSE}}(T|M) = \frac{1}{c_\cdot} \left(\sum_{i=1}^{N} c_i (\bar{H}_i - M_i)^2 - N\hat{V}_\epsilon \right) \tag{7}$$

B.4 Estimating PRMSE

With estimators for the MSE and the variance of the true score available, estimation of PRMSE is simple.

$$\widehat{\text{PRMSE}} = 1 - \frac{\widehat{\text{MSE}}(T|M)}{\hat{V}_T} \tag{8}$$

C Impact of double-scoring

Table 4 shows the range of PRMSE values we observed for different number of double-scored responses and human-human agreement.

	Human-human agreement			
N	Low	Moderate	Average	High
100	1.01	0.41	0.26	0.12
250	0.46	0.30	0.15	0.09
500	0.33	0.17	0.12	0.07
1,000	0.24	0.13	0.08	0.06
2,500	0.18	0.09	0.07	0.03
5,000	0.08	0.07	0.04	0.02
10,000	0.06	0.03	0.02	0.02

Table 4: The range of observed PRMSE values for different number double-scored responses and different levels of human-human agreement.