Quadratic Weighted Kappa is Not Enough for Evaluating Automated Essay Scoring Models

Salam Albatarni and Tamer Elsayed
Computer Science and Engineering Department, Qatar University
Doha, Qatar
{sa1800633, telsayed}@qu.edu.qa

Abstract

Quadratic Weighted Kappa (QWK) has been the standard evaluation metric in Automated Essay Scoring (AES) research for over two decades. Despite repeated criticisms highlighting its limitations, the community has largely continued to rely on QWK without adopting alternative metrics. This study aims to encourage a shift toward more suitable evaluation practices by systematically examining QWK's behavior under three key conditions: dataset size, class imbalance, and score range. Using both a publicly available AES dataset and carefully synthesized datasets, we demonstrate scenarios where QWK produces unstable or misleading results. Our findings highlight the need for more robust evaluation practices and point to alternative metrics, particularly variants of Gwet's AC2, that offer greater reliability across a variety of conditions.

Keywords: Automated Essay Scoring, Metric Evaluation, Quadratic Weighted Kappa

1. Introduction

Automated Essay Scoring (AES) has been an active area of research for over six decades, with the goal of automatically assessing writing proficiency (Page, 1966). While AES systems have advanced significantly, one critical aspect has often been overlooked: the choice of evaluation metrics. The ASAP[1] dataset, the most-widely used benchmark for AES, was introduced through a Kaggle competition that adopted Quadratic Weighted Kappa (QWK) (Cohen, 1968) as its official evaluation measure. Since then, QWK has become the *de facto standard* for comparing AES systems. However, this widespread adoption has led to an unintended consequence: QWK is now *routinely* applied across diverse contexts: different datasets, languages, and assessment , without sufficiently studying its suitability for the different setups.

While QWK is a statistically sound metric for agreement, its dominance has overshadowed other meaningful performance indicators, particularly in constrained environments. Indeed, this particular problem has been brought up by multiple studies (Yannakoudakis and Cummins, 2015; Brenner and Kliebsch, 1996). Yet, modern AES research continues relying on QWK as the sole metric.

AES systems are typically evaluated under two settings: prompt-specific (training and testing on the same prompt) and cross-prompt (training and testing on different prompts)[2]. A critical factor in both settings is the number of essays available for evaluation. For instance, the ASAP dataset contains an average of 1,600 essays across 8 prompts. In prompt-specific evaluation, the standard 5-fold cross-validation setup uses roughly 320 essays per fold, per prompt, for testing (Taghipour and Ng, 2016), while cross-prompt evaluation tests on all essays from a held-out prompt (∼1,600 essays) (Ridley et al., 2021). While such large-scale evaluation is feasible for English (thanks to ASAP), it becomes impractical for low-resource languages, where datasets may comprise only 200 essays per prompt (Bashendy et al., 2024). In such cases, a 5-fold prompt-specific evaluation would test on just 40 essays per fold, a scenario where QWK's reliability is questionable.

This limitation motivates the present study to address a central question: How robust are evaluation metrics in AES? Robustness is defined as the stability of a metric's system rankings across varying evaluation conditions (Sakai, 2021). To investigate this, we evaluate seven metrics, including QWK and others less commonly reported, under controlled variations in the data. We divide the main question into three research questions, each targeting a specific condition: **(RQ1)** How do the metrics behave under different dataset sizes? **(RQ2)** How does the score range of essays affect metric behavior? **(RQ3)** How are the metrics influenced by skewed score distributions?

Our contributions are as follows:

- We systematically evaluate the robustness of seven common AES metrics under three conditions: limited dataset size, varying score ranges, and different score distributions.
- We analyze the metrics across these conditions and offer practical recommendations on which metrics to use under specific constraints.

[1] https://www.kaggle.com/c/asap-aes

[2] A prompt refers to a distinct writing task or question.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 4447–4456
11–16 May 2026.

- We make our code and data publicly available to ensure reproducibility.[3]

We examine AES metrics and prior evaluation studies in §2, outline our evaluation methodology in §3, present results addressing the three research questions in §4, discuss the findings and implications in §5, and conclude in §6.

2. Related Work and Background

In this section, we review the evaluation metrics commonly reported in AES research (§2.1) and prior studies evaluating QWK (§2.2).

2.1. AES Metrics Background

QWK is favored in AES as it accounts for both agreement and the ordinal nature of scores. It penalizes larger disagreements more heavily than smaller ones. It also adjusts for chance agreement, making it more informative than raw accuracy. Because of these properties, QWK has been adopted as the standard evaluation metric in many AES shared tasks and benchmarks.

Pearson and **Spearman** correlations are commonly used in AES to measure the relationship between predicted scores and human scores (Yannakoudakis et al., 2011; Alqahtani and Al-Saif, 2020; Yannakoudakis and Briscoe, 2012). Pearson captures linear correlation, while Spearman measures monotonic relationships. These metrics indicate how well the system ranks essays, not whether the predicted scores match true scores. Indeed, a system could achieve a high correlation while consistently over- or under-predicting scores (Yannakoudakis and Cummins, 2015).

Root Mean Square Error (**RMSE**) has also been employed to evaluate AES systems (Hirao et al., 2020; Bashendy et al., 2024). It measures the average squared difference between predicted and true scores, penalizing large errors more strongly. However, it assumes continuous intervals, overlooks the ordinal and discrete nature of essay scores, and is insensitive to score distribution differences.

Accuracy, defined as the percentage of predictions that exactly match human scores, has also been reported in some studies (Alqahtani and Al-Saif, 2020; Attali and Burstein, 2004). Although intuitive, it ignores the magnitude and order of errors and is highly affected by class imbalance, allowing models to achieve high scores by overpredicting the majority class.

Proportional Reduction in Mean-Squared Error (PRMSE) has been proposed as an alternative AES evaluation metric (Casabianca et al., 2023; Loukina et al., 2020). It reflects how automated scores reduce prediction error relative to the average human score. However, it requires scores from two human raters to adjust for measurement error, limiting its applicability when only one rater's scores are available.

2.2. Evaluating Evaluation Measures for AES

Studying evaluation measures has been a long-standing focus in Information Retrieval (IR) (Buckley and Voorhees, 2017), leading to well-established standards (Sakai, 2021). In contrast, AES lacks comparable standardization. Although several studies have questioned the use of QWK as the main evaluation metric (Brenner and Kliebsch, 1996; Yannakoudakis and Cummins, 2015; Doewes et al., 2023), it remains widely adopted, highlighting the need to reexamine its suitability.

Brenner and Kliebsch (1996) demonstrated that QWK is sensitive to the number of scoring categories. They systematically varied the number of categories from 2 to 8 and found that QWK scores consistently increased with more categories. This trend was especially noticeable within the range of 2 to 5 categories, which is commonly used in practical AES applications. Yannakoudakis and Cummins (2015) provided a broader evaluation of different agreement measures. They proposed four essential criteria for assessing AES metrics: robustness to trait prevalence, marginal homogeneity, score range, and sensitivity to the magnitude of misclassification. Their study compared several metrics, including QWK, linearly weighted kappa and Gwet's AC2. They found that QWK performed the worst, while the AC2 measure, particularly AC with quadratic weights, is more reliable. Notably, they explicitly advised against using Cohen's Kappa. More recently, Doewes et al. (2023) revisited the shortcomings of QWK, reinforcing concerns raised in previous work. Although they proposed Krippendorff's alpha as an alternative, they did not systematically show how it overcomes QWK's limitations.

Loukina et al. (2020) highlighted QWK's sensitivity to rater reliability. They demonstrated, through simulation experiments, that QWK is highly influenced by the specific set of human raters used as reference. The same automated system can receive significantly different performance scores depending solely on the variability in human-human agreement, emphasizing the need to consider rater quality when interpreting evaluation results.

The efforts of prior studies, particularly Yannakoudakis and Cummins (2015) and Loukina et al. (2020), provide a solid foundation for questioning the continued use of QWK in AES. However, they

[3] https://github.com/salbatarni/LREC2026_QWK_for_AES

fall short of offering a complete picture. Most research focuses on statistical properties, theoretical limitations, or changes in metric values, but do not assess whether changes in dataset characteristics lead to meaningful shifts in *system rankings*. In this work, we address this gap by examining how system rankings are are influenced by varying dataset properties: size, number of categories, and data skewness through systematic experimentation.

3. Evaluation Methodology

In this work, we examine how specific dataset conditions affect the ranking of AES systems produced by evaluation metrics. This analysis centers on the concept of *robustness*, which refers to the stability of system rankings when the evaluation conditions change (Sakai, 2021). In practice, robustness is critical; it determines whether we can trust the reported ranking of systems. A robust metric should consistently indicate that one system outperforms another, even when the evaluation dataset is slightly altered.

We define the problem as follows: given a set of AES systems $S = \{s_1, s_2, \ldots, s_N\}$, and an evaluation metric $\mathcal{M}$, how robust is the ranking of these systems under different dataset conditions $\mathcal{C}$? Specifically, we ask whether changes in $\mathcal{C}$ will lead to meaningful changes in the system ranking produced by $\mathcal{M}$. Formally, $R^{\mathcal{C}}_{\mathcal{M}}$ denotes the ranking of systems according to metric $\mathcal{M}$ under condition $\mathcal{C}$, and $R^{\mathcal{C}'}_{\mathcal{M}}$ denotes the ranking under a different condition $\mathcal{C}'$. We quantify the consistency between these rankings using Kendall's τ correlation coefficient (Kendall, 1938):

$$\tau = \frac{(n_c - n_d)}{\frac{1}{2}N(N-1)} \tag{1}$$

where n_c and n_d are the number of concordant and discordant pairs across the two rankings, respectively. High τ would indicate that the metric $\mathcal{M}$ is robust regardless of the change in condition $\mathcal{C}$. Hence, metric $\mathcal{M}$ system ranking is trustworthy.

3.1. Synthetic Systems

To evaluate the robustness of evaluation metrics under controlled conditions, we construct a set of 50 synthetic AES systems; each is defined by a target accuracy value, ranging from 0 to 0.98 in increments of 0.02. Accuracy here is measured as the proportion of essays over all prompts that are assigned the correct (i.e., human-annotated) score.

For each system, we generate its predicted scores by randomly sampling a proportion of essays equal to the target accuracy and assign them their true scores. For the remaining essays (i.e., the inaccurate predictions), we randomly assign an incorrect score selected uniformly from the valid score range of the corresponding prompt, ensuring that it differs from the true score. This distinction is essential, as ranking consistency depends on variation in system-level scores.

3.2. Metrics

We establish a set of criteria for selecting evaluation metrics based on their popularity in the literature and recommendations from prior research. Rather than identifying a single *"best"* metric for all scenarios, we aim to determine which metrics are most appropriate to report in specific contexts. Since different metrics capture distinct aspects of system performance, we also investigate whether their behavior is influenced by underlying or hidden factors. With these considerations, we select seven widely used and previously recommended metrics in AES:

- Quadratic Weighted Kappa (**QWK**) (Cohen, 1968) measures the agreement between two raters while penalizing larger disagreements more heavily. It is sensitive to the ordinal nature of the scores, making it suitable for evaluating systems that predict ordered categories.
- **Pearson's r** is a correlation coefficient that quantifies the linear relationship between predicted and true scores. Despite its limitations (Carroll, 1961), it is still used for AES (Alqahtani and Al-Saif, 2020).
- Weighted Agreement Coefficient (**AC2**) (Gwet, 2014) is a chance-corrected agreement measure designed to be less sensitive to prevalence and marginal distributions. Yannakoudakis and Cummins (2015) recommended quadratic AC2, hereafter AC2-q, over QWK.
- Linear AC2 (Gwet, 2014) is a variant of AC2 using linear weights, hereafter AC2-l, instead of quadratic, resulting in less severe penalties for disagreements.
- **Krippendorff** (Krippendorff, 1970) is a general-purpose reliability coefficient that supports different data types, including nominal, ordinal, and interval. Doewes et al. (2023) suggested it as an alternative to QWK. However, they did not evaluate its reliability, which we aim to investigate.
- Root Mean Squared Error (**RMSE**) captures the average magnitude of prediction errors.
- **Accuracy** computes the proportion of exact matches between predicted and true labels. Although limited in AES, it is sometimes used

as a secondary metric (Rupp et al., 2019; Ormerod, 2025).

4. Experimental Evaluation

4.1. Experimental Environment

In our experiments, we use scikit-learn for computing QWK, RMSE, and Accuracy. To compute AC2, we use the Python library irrCAC,[4] which is the official Python implementation corresponding to the author's R library (Gwet, 2014). For Krippendorff's alpha, we use the Fast Krippendorff library (Castro, 2017), with the level of measurement set to interval.

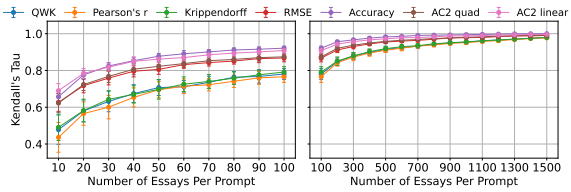
4.2. Effect of Dataset Size (RQ1)

While it is generally accepted that evaluating on larger datasets leads to more reliable and stable results, this consideration is often overlooked in AES research, particularly for languages other than English. The majority of existing studies focus on English, where large-scale datasets are more readily available. As a result, the implications of small dataset sizes on evaluation metrics remain underexplored. Notably, several works have applied cross-validation on datasets with very limited samples, sometimes with test sets containing fewer than 50 essays (Bashendy et al., 2024; Alqahtani and Al-Saif, 2020; Hirao et al., 2020). This raises concerns about the reliability of evaluation results in low-resource settings and highlights the need to systematically examine how dataset size influences metric behavior and system ranking.

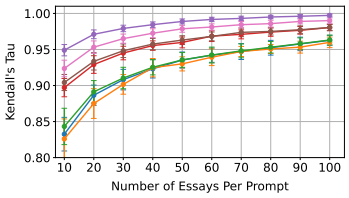
Dataset For this study, we elect to use a real dataset rather than a synthetic one, as our goal is to investigate how variations in dataset size affect evaluation metrics in practical real-world conditions. We selected two widely used datasets: the Automated Student Assessment Prize (ASAP)[5] and ELLIPSE (Crossley et al., 2023). ASAP contains essays from eight prompts with distinct score ranges, each annotated with a holistic score and several trait-specific scores; on average, each prompt includes about 1,600 essays. ELLIPSE consists of 44 prompts with scores ranging from 1 to 5 in 0.5 increments, with roughly 150 essays per prompt. In this work, we focus on the holistic scores for both datasets. The scores of ELLIPSE are multiplied by 2 to obtain integer values; this transformation is applied in all subsequent experiments.

Experimental Settings For each prompt in both ASAP and ELLIPSE, we first evaluate the performance of all synthetic systems using the full set of essays using each evaluation metric $\mathcal{M}$. This yields the baseline system ranking $R^b_{\mathcal{M}}$. Next, we systematically reduce the dataset size by randomly sampling subsets of essays of varying sizes. For ASAP, subset sizes range from 10 to 100 essays in increments of 10, and then from 100 to 1,500 essays in increments of 100. Since Prompt 8 contains only 723 essays, it is omitted from this particular experiment. For ELLIPSE, we limit subsets to a maximum of 100 essays, resulting in 17 prompts used for this analysis. In all cases, the baseline ranking $R^b_{\mathcal{M}}$ is computed using the full dataset for each prompt. For each subset size z, we evaluate all systems again using metric $\mathcal{M}$, obtaining new rankings $R^z_{\mathcal{M}}$ based on the reduced data. For mitigating the effects of random sampling, we repeat it 50 times per subset size z. This overall procedure is repeated independently for each prompt.

We quantify the robustness of metric $\mathcal{M}$ by comparing the baseline ranking $R^b_{\mathcal{M}}$ with each reduced-data ranking $R^z_{\mathcal{M}}$ using Kendall's τ. This allows to assess how system rankings change as a function of dataset size for each evaluation metric. Kendall's τ values of $R^z_{\mathcal{M}}$ are then averaged across repetitions and prompts to provide a robust estimate of *ranking stability* under dataset size variation.



(a) ASAP: Results averaged across 7 prompts.



(b) ELLIPSE: Results averaged across 17 prompts.

Figure 1: Robustness of metrics under varying dataset sizes, measured by Kendall's τ between system rankings using all data and subsets of different sizes.

Results Figure 1 presents the robustness results of seven different evaluation metrics as measured by Kendall's τ across varying dataset sizes.

Generally, metrics are consistent across the two datasets and there are clear differences in metric robustness under reduced data conditions. We note that the most robust metrics for both datasets were Accuracy and AC2-l. These metrics maintained relatively high Kendall's τ values even with

[4] https://github.com/afergadis/irrCAC
[5] https://www.kaggle.com/c/asap-aes

smaller subsets of essays. Conversely, QWK, Pearson's r, and Krippendorff exhibited the lowest robustness, with Kendall's τ values dropping significantly as dataset size decreased. For ASAP, their robustness persisted only above roughly 200 essays, while for ELLIPSE, rankings began to decline after about 50 essays. This finding raises concerns about task-specific AES on prompts of small number of essays. In fact, in many AES studies, researchers employ the common 5-fold cross-validation splits, resulting in test sets of small sizes that might be below the suggested minimum for reliable metric use.

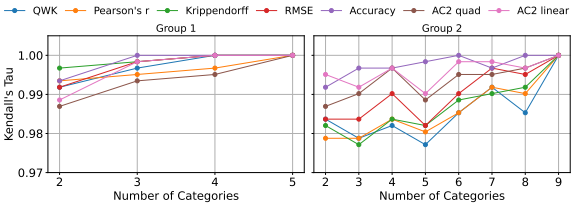
Fortunately, metrics such as AC2-l, AC2-q, Accuracy, and RMSE exhibited greater robustness. While their Kendall's τ values declined as dataset sizes dropped below 200 for ASAP and 50 for ELLIPSE, they generally remained above 0.6 and 0.9, respectively, indicating relatively stable system rankings.

Messages These findings lead to several important takeaways. First, while it is acceptable to use QWK, researchers must be cautious about the size of the evaluation data, whether it is development or test data. Second, accuracy proved to be one of the most robust metrics. It offers a straightforward interpretation of how often a system predicts the correct score. However, it is not designed to capture inter-rater agreement and may overestimate performance when score distributions are skewed. Thus, while it can be informative, it should be used deliberately and in conjunction with other metrics. Third, AC2-l and AC2-q demonstrated strong robustness across dataset sizes, supporting prior recommendations for their use as reliable agreement-based metrics (Yannakoudakis and Cummins, 2015).

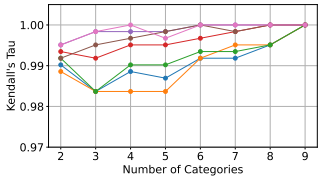
Finally, while collecting large annotated essay sets is resource-intensive, our results highlight the need for careful metric selection rather than relying on popular defaults. As it directly affects the reliability of AES conclusions, especially with the small datasets common in low-resource languages.

4.3. Effect of Score Range (RQ2)

The score range for essays can vary significantly depending on the writing trait being assessed, proficiency level of the students, or the type of assessment. As a result, different writing assessment tasks often use different numbers of scoring categories. Prior studies have noted that QWK is sensitive to the number of score categories, where the increase of the score range results in increase in QWK (Yannakoudakis and Cummins, 2015; Brenner and Kliebsch, 1996; Doewes et al., 2023). However, these findings focus on absolute metric values rather than the consistency of system rankings



(a) ASAP dataset.



(b) ELLIPSE dataset.

Figure 2: Robustness of metrics under varying score ranges, measured by Kendall's τ between system rankings using the original score range and reduced category settings.

across varying score ranges. In this section, we address a different, but practically important, question: *Does the number of scoring categories affect the robustness of evaluation metrics?*

4.3.1. Dataset and Experimental Settings

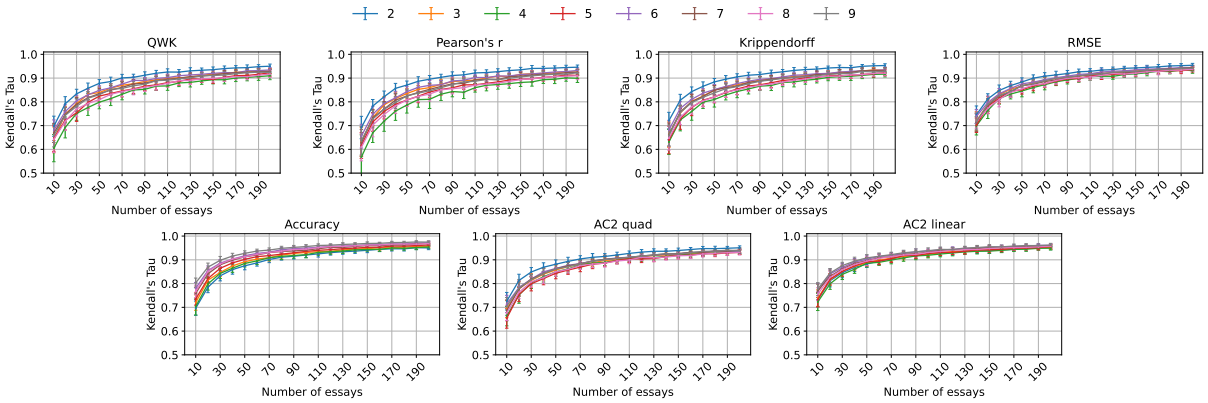
ASAP and ELLIPSE datasets are used for this experiment as well, allowing us to study diverse set of prompts across varying score ranges. To assess the effect of score range, we progressively reduce the number of scoring categories within each dataset, till having 2 categories. This is done by rescaling all (true and predicted) scores. We then observe how these transformations affect system rankings under each evaluation metric.

For ELLIPSE, we include all 44 prompts each with nine categories (score range $[2, 10]$). For ASAP, we organize prompts into two groups based on their original score ranges:

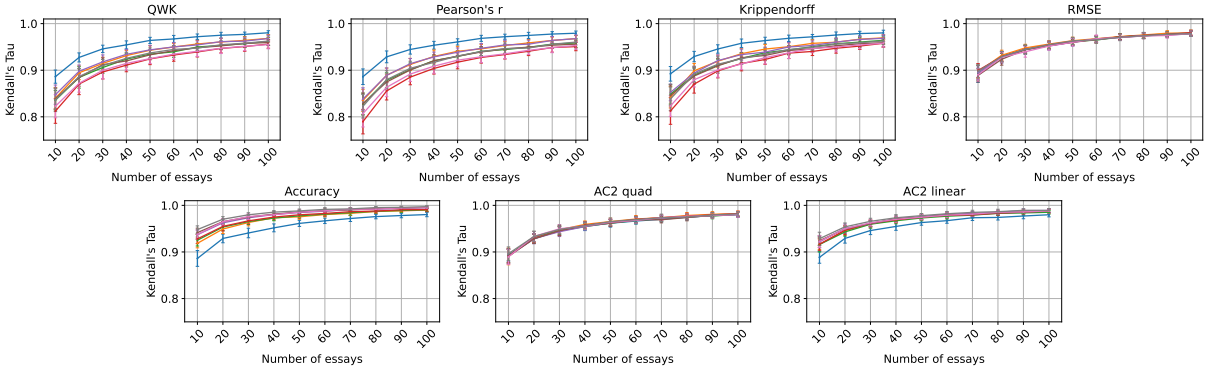
- **Group 1** includes prompts 2, 3, 4, 5, and 6, which we normalize to a range of 5 categories; all scores are scaled to the range $[0, 4]$, except for prompts 3 and 4, which naturally fall within $[0, 3]$ and are left unchanged.

- **Group 2** includes prompts 1, 7, and 8, which we normalize to a wider scale with 9 categories, mapping their scores to the range $[2, 10]$.

Similar to the dataset size experiment, our baseline ranking is defined as the system ranking obtained using the group's original score range. For each reduced category setting, we then compute the new ranking and evaluate its consistency with the baseline using Kendall's τ.



(a) ASAP dataset.



(b) ELLIPSE dataset.

Figure 3: Effect of the interaction between dataset size and number of categories on metric robustness. Kendall's τ is computed between system rankings based on subsets of varying size (x-axis) and the full dataset ranking. Each curve corresponds to a different number of categories.

4.3.2. Varying Score Ranges

We plot Kendall's τ against the number of scoring categories in Figure 2. The most notable observation is that Kendall's τ remains consistently high across all category settings in both datasets, not falling below 0.97. This suggests that the number of categories does not substantially affect system rankings, even though it may influence the absolute values of some metrics, as noted in Yannakoudakis and Cummins (2015). In other words, while metric scores may shift depending on the score range, the relative ranking of systems remains stable.

In terms of metric behavior, Group 1 from ASAP, exhibits minimal ranking variation across all metrics. Most metrics show a drop of only around 0.01 in Kendall's τ, with Krippendorff showing the smallest decline of just 0.3 points. The largest drop is observed with AC2-q, though the decline remains negligible overall. Group 2 from ASAP and ELLIPSE dataset follows a similar trend, although the effects of score range changes are more pronounced due to the wider initial scale. QWK, Pearson's r, and Krippendorff show the largest drops, up to 2 points in Kendall's τ. In contrast, accuracy and AC2-l remain highly stable, with τ values above 0.99. Nevertheless, this experiment indicates that system comparisons are reliable regardless of the number of scoring categories.

4.3.3. Varying Dataset Size & Score Range

While the previous experiments addressed the independent effects of dataset size and scoring range, a natural follow-up question arises: *How do these two factors interact?* To explore this, we extend the dataset size robustness experiment by including variations in the number of categories. We focus on Group 2 prompts from ASAP and ELLIPSE dataset due to their wider original scoring range.

We follow the same setup as before: system rankings are computed using subsets of increasing sizes, ranging from 10 to 200 and 100 essays per prompt (in increments of 10) for ASAP and ELLIPSE, respectively. For each score range, Kendall's τ is computed with respect to the full dataset ranking. For ASAP, based on earlier observations from RQ1, we restrict the analysis to subsets below 200 essays, where most metric instability occurs. Each configuration is repeated

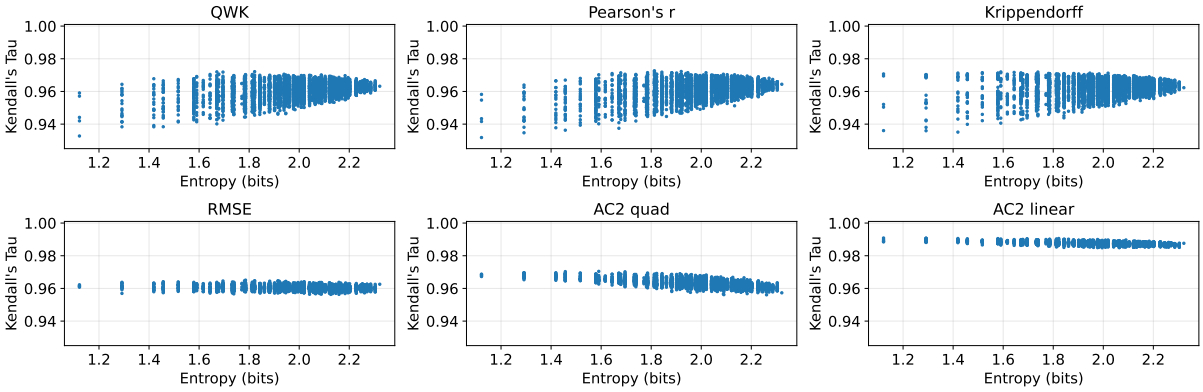


Figure 4: Kendall's τ between system rankings under different score distributions (High, Medium, Low entropy).

50 times to account for variability due to random sampling.

Figure 3 shows the results of the seven metrics. The first clear observation is that both datasets results are consistent. Second, score ranges with only 2 categories achieve stable rankings with relatively few essays compared to other metrics. Specifically, accuracy was the metric needed the least number of essays on 5 out of the 7 metrics, namely all but accuracy and AC2-l for ASAP dataset and ELLIPSE datasets. Beyond this, the metrics exhibit different behaviors with respect to the number of categories. For example, on ASAP, QWK shows higher robustness with nine categories ($\tau = 0.75$ at 20 essays) than with four ($\tau = 0.69$ at 20 essays). More categories amplify small quality differences, leading to wider score gaps and more distinct system rankings which makes the evaluation more stable. However, this trend is not universal, each metric responds differently based on its underlying calculation. Accuracy, in particular, shows the opposite pattern: as the number of categories increases, fewer essays are needed for stable rankings. This is because finer-grained scoring reduces ties, making accuracy more discriminative and allowing even small improvements to translate more reliably into ranking changes.

Another pattern is the varying sensitivity of metrics to number of categories, visible in the spacing between curves in Figure 3. Metrics like AC2-l show tightly grouped curves, indicating consistent robustness across different category numbers. In contrast, Pearson's r and QWK exhibit a wider spread, suggesting that its stability depends more heavily on number of categories. Overall, AC2-l consistently shows the highest robustness. It maintains a Kendall's $\tau > 0.8$ across both datasets in nearly all settings, dropping below this threshold only when fewer than 20 essays are used. RMSE, AC2-q, and accuracy also demonstrate strong stability across different category settings.

4.3.4. Messages

Requiring only a small number of essays to evaluate models on datasets with 2 categories is an important observation. Although such scales are not very common, they can be useful for assessing writing traits like relevance, where essays are simply classified as relevant or not. In such cases, even less reliable metrics may tolerate small dataset sizes. However, AC2-l, still proves the most robust metric.

4.4. Effect of Dataset Distribution (RQ3)

The final question we address is how the dataset distribution affects metric robustness. This issue has been discussed in prior work, notably by Yannakoudakis and Cummins (2015) who showed that the *absolute value* of QWK is highly sensitive to data distribution. However, our focus here is different; rather than examining changes in metric *values*, we consider the *relative performance of systems*. In other words, could a lower-performing system "appear better" than a stronger one purely due to distributional shifts?

4.4.1. Dataset and Experimental Settings

Our goal is to investigate how changes in score distribution, from uniform to highly skewed, affect the robustness of evaluation metrics. We first fix the score range to 1–5, yielding five categories and and set the number of samples to 1,000. We consider the uniform distribution, where all score categories are equally represented, as the *ideal* or reference case, which produces the baseline system ranking $R^b_{\mathcal{M}}$. Starting from this reference, we gradually alter the distribution by lowering its entropy, effectively moving from balanced to increasingly skewed datasets each associated with a ranking $R^z_{\mathcal{M}}$. To achieve this, we discretize the probability space using a fixed 5% step size and enumerate all possi-

ble combinations of category proportions that sum to 100%. This procedure yields a comprehensive set of 3,876 distributions covering a broad range of entropy levels, allowing us to analyze how metric robustness changes as the dataset becomes more or less balanced. For each of these 3,876 distributions, we generate system predictions as described in §3.1, repeating the process 50 times per distribution to account for randomness in system predictions, and report the average Kendall's τ across repetitions.

4.4.2. Results

The Kendall's τ correlations between system rankings across different entropies are shown in Figure 4. We note that accuracy is not part of these because it remains unchanged across all conditions ($\tau = 1$), as it is the fixed parameter used to control system performance.

Overall, AC2 variants and RMSE exhibit very high correlations, consistently above 0.95, indicating that these metrics are highly robust to changes in score distribution. In contrast, QWK, Pearson's r, and Krippendorff show lower correlations in some cases, with τ occasionally dropping below 0.95.

The spread of points at the same entropy reveals additional insights. For a given entropy level, all distributions share the same proportions but differ in which score dominate. AC2 and RMSE produce tightly clustered points, suggesting that their system rankings are largely unaffected by which scores dominate. Conversely, QWK, Pearson, and Krippendorff exhibit wider spreads, indicating that the dominance of particular score levels can influence system rankings even when overall entropy is fixed.

4.4.3. Messages

Despite these variations, kendall's tau remains generally high across all metrics (above 0.9), showing that the relative system rankings are fairly robust to changes in dataset distribution. This complements prior studies that focused on absolute metric values (Yannakoudakis and Cummins, 2015), showing that, even though absolute scores may change under different distributions, the relative system rankings remain largely robust.

5. Discussion and Recommendations

Why does this matter? In recent AES research, system performance is usually compared to baselines under fixed conditions, with modest gains (often under 10 points) and no statistical significance testing. Such improvements should therefore be interpreted cautiously. Our findings indicate that these small differences may not reflect true performance gaps but rather arise from uncontrolled factors like dataset size, score range, or class imbalance. Dataset size is a particularly common issue. In many cases, datasets with around 1,000 essays are small from a machine learning point-of-view, leading researchers to use 10-fold cross-validation. This leaves only about 100 essays per test fold, which our results show is insufficient for stable rankings under several metrics, especially QWK, and more so when score distributions are skewed or mid-range scoring conditions. Based on our three research questions, we suggest the following recommendations:

Avoid QWK with small datasets: When working with small datasets (fewer than 200 essays in evaluation data), avoid using QWK, Pearson's r, and Krippendorff, as these metrics show high instability in system rankings. Instead, prefer metrics like AC2-l, AC2-q or RMSE, which demonstrate stronger robustness under small dataset settings.

Consider the scoring scale: Metrics behave differently depending on the number of score categories. In particular, metrics such as QWK and Pearson's r become less reliable when mid-range scoring scales (e.g., 4 categories) are used in small datasets. Choose metrics that remain consistent across scales, AC2 is a strong candidate.

Account for label distribution: Our experiments show that skewed distributions can affect ranking stability. When working with imbalanced data, use metrics that are less sensitive to distributional shifts, again, AC2-l performs best in this regard.

Report multiple metrics: To gain a more comprehensive view of model performance, report a diverse set of metrics rather than relying on a single value. Including RMSE or accuracy alongside agreement-based metrics provides a more complete picture of system behavior.

6. Conclusion and Future Work

This paper contributes to the growing body of work questioning the reliability of commonly used evaluation metrics, such as QWK, in AES research. Through a series of controlled experiments, we show that QWK, in particular, is not robust under small sample sizes, and its behavior is strongly influenced by score distribution and, to a lesser extent, by the number of score categories. Our findings highlight the need for more careful metric selection and reporting practices in AES research. Metrics such as AC2-l, AC2-q and RMSE consistently demonstrate higher robustness across a wide range of experimental conditions, making them

more suitable for fair and trustworthy evaluation, especially when working with limited or imbalanced data.

For future work, it would be valuable to gather available AES datasets across different languages, assessment types (e.g., argumentative, persuasive), and writing traits (e.g., organization), and examine how their underlying characteristics affect the robustness of evaluation metrics. Moreover, training state-of-the-art AES systems under varying dataset conditions could help determine whether their rankings remain consistent when evaluated using more robust metrics such as AC2.

7. Acknowledgments

We thank Fatima Elnahas for her earlier contributions to this work. This work was made possible by NPRP grant# NPRP14S-0402-210127 from the Qatar Research Development and Innovation (QRDI) Council. The statements made herein are solely the responsibility of the authors.

8. Bibliographical References

Abeer Alqahtani and Amal Al-Saif. 2020. Automated arabic essay evaluation. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 181–190.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2):i–21.

May Bashendy, Salam Albatarni, Sohaila Eltanbouly, Eman Zahran, Hamdo Elhuseyin, Tamer Elsayed, Walid Massoud, and Houda Bouamor. 2024. QAES: First publicly-available trait-specific annotations for automated scoring of Arabic essays. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 337–351, Bangkok, Thailand. Association for Computational Linguistics.

Hermann Brenner and Ulrike Kliebsch. 1996. Dependence of weighted kappa coefficients on the number of categories. *Epidemiology*, pages 199–202.

Chris Buckley and Ellen M Voorhees. 2017. Evaluating evaluation measure stability. In *Acm Sigir Forum*, volume 51, pages 235–242. ACM New York, NY, USA.

John B Carroll. 1961. The nature of the data, or how to choose a correlation coefficient. *Psychometrika*, 26(4):347–372.

Jodi M Casabianca, Daniel F McCaffrey, Matthew S Johnson, Kathryn L Ricker-Pedley, Ourania Rotou, and Joseph Martineau. 2023. Exploration of the proportional reduction in mean-squared error for evaluating automated scores. *Research Memorandum No. RM-23-01). ETS. https://www.ets. org/Media/Research/pdf/RM-23-01. pdf*.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agreement measure. https://github.com/pln-fing-udelar/fast-krippendorff.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Scott Crossley, Yu Tian, Perpetual Baffour, Alex Franklin, Youngmeen Kim, Wesley Morris, Meg Benner, Aigner Picou, and Ulrich Boser. 2023. The english language learner insight, proficiency and skills evaluation (ellipse) corpus. *International Journal of Learner Corpus Research*, 9(2):248–269.

Afrizal Doewes, Nughthoh Kurdhi, and Akrati Saxena. 2023. Evaluating quadratic weighted kappa as the standard performance metric for automated essay scoring. In *16th International Conference on Educational Data Mining, EDM 2023*, pages 103–113. International Educational Data Mining Society (IEDMS).

Kilem Li Gwet. 2014. The definitive guide to measuring the extent of agreement among raters. *Handbook of inter-rater reliability. 4th Edition. Advanced Analytics, LLC*, 429.

Reo Hirao, Mio Arai, Hiroki Shimanaka, Satoru Katsumata, and Mamoru Komachi. 2020. Automated essay scoring system for nonnative japanese learners. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1250–1257.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Anastassia Loukina, Nitin Madnani, Aoife Cahill, Lili Yao, Matthew S Johnson, Brian Riordan, and Daniel F McCaffrey. 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*, pages 18–29.

Christopher Ormerod. 2025. Automated essay scoring incorporating annotations from automated feedback systems. *arXiv e-prints*, pages arXiv–2505.

Ellis B Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Robert Ridley, Liang He, Xin-yu Dai, Shujian Huang, and Jiajun Chen. 2021. Automated cross-prompt scoring of essay traits. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13745–13753.

André A Rupp, Jodi M Casabianca, Maleika Krüger, Stefan Keller, and Olaf Köller. 2019. Automated essay scoring at scale: A case study in switzerland and germany. *ETS Research Report Series*, 2019(1):1–23.

Tetsuya Sakai. 2021. Evaluating evaluation measures for ordinal classification and ordinal quantification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2759–2769.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Helen Yannakoudakis and Ted Briscoe. 2012. Modeling coherence in esol learner texts. In *Proceedings of the seventh workshop on building educational applications using NLP*, pages 33–43.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading esol texts. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 180–189.

Helen Yannakoudakis and Ronan Cummins. 2015. Evaluating the performance of automated text scoring systems. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 213–223.