ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2024; 15(Special issue); 348-360

Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness

Mo ZHANG* Matthew JOHNSON** Chunyi RUAN***

Abstract

AI scoring capabilities are commonly implemented in educational assessments as a supplement or replacement to human scoring, with significant interest in leveraging large language models for scoring. In order to use AI scoring capability responsibly, the AI scores should be accurate and fair. In this study, we explored one approach to potentially mitigate bias in AI scoring by using equal-allocation stratified sampling for AI model training. The data set included 13 open-ended short-response items in a K-12 state science assessment. Empirical results suggested that stratification did not improve or worsen fairness evaluations on the AI models. BERT based AI scoring models resulting from the stratified sampling method but trained on much less data performed comparably to models resulting from simple random sampling in terms of overall prediction accuracy and fairness on the subgroup level. Limitations and future research are also discussed.

Keywords: AI scoring, educational assessment, large language model, sampling, prediction accuracy, fairness

Introduction

AI scoring capabilities are commonly implemented in educational assessments as a supplement or replacement to human scoring. For example, AI scoring has been used to score open-ended text responses in various content domains (e.g., math, reading, writing, science, speaking) and assessments with varying levels of scale and stakes, including PTE English, TOEFL iBT, GMAT, GRE, LSAT, and certification/licensure tests such as Praxis, as well as many K-12 state-level assessments (e.g., Kentucky Summative Assessment). The literature on AI scoring has grown substantially in the past 10 to 20 years. Bennett and Zhang (2016) considered AI (or automated) scoring as "machine grading of constructed responses that are generally not amenable to exact-matching approaches because the specific form(s) and/or content of the correct answer(s) are not known in advance." An AI scoring algorithm is a computational procedure used in educational testing to predict or determine scores for test items or responses automatically. These algorithms typically use natural language processing and statistical or machine learning techniques to generate the predicted scores based on patterns or associations found in the data.

In early examples of AI scoring such as automated essay scoring, the AI score is usually a weighted combination of a small set of well-defined linguistic features, such as grammatical accuracy, vocabulary sophistication, sentence structure, and so forth, and these features are carefully evaluated by content experts to closely align to and cover the construct of measurement. The scoring algorithms tend to be white-box or gray-box models such as decision trees, linear regressions, and k-means. For these earlier approaches to AI scoring, the features used in the model are construct-relevant, the weights given to

* Senior Research Scientist, Educational Testing Service, New Jersey-USA, mzhang@ets.org, ORCID ID: 0000-0003-2689-2089
** Principal Research Director, Educational Testing Service, New Jersey-USA, msjohnson@ets.org, ORCID ID: 0000-0003-3157-4165
*** Principal Research Data Analyst, Educational Testing Service, New Jersey-USA, cruan@ets.org, ORCID ID: 0009-0009-3073-229X

To cite this article:
Zhang, M., Johnson, M. & Ruan , C. (2024). Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness*, Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 348-360. https://doi.org/10.21031/epod.1561580

Received: 4.10.2024
Accepted: 12.11.2024

each feature can be extracted, and the reasoning from the features can be tracked. In this case, the scores are highly explainable and interpretable.

As generative AI has surged in popularity and revolutionized various sectors in the society, interest has increased in leveraging large language models (LLMs) for scoring (Chamieh, Zesch, & Giebermann, 2024; Lee, Latif, Wu, Liu, & Zhai, 2024; Lubis, Putri, et al., 2021; Kortemeyer, 2024; Oka, Kusumi, & Utsumi, 2024; Whitmer et al., 2021). Using LLMs for scoring is particularly relevant to assessing content and reasoning in areas in which traditional approaches have fallen short. Even though white- or gray-box models have great interpretability, their prediction accuracy is usually lower compared to black-box models such as transformer-based models (e.g., GPT, BART), deep learning, and neural networks (Ali, Abuhmed, El-Sappagh, et al., 2023; Kumar, Dikshit, & de Albuquerque, 2021). However, as models increase in complexity, interpretability diminishes substantially because millions of parameters are estimated to generate a score. For example, LlaMa 3.1 (released on 06/23/2024 by Meta AI) has 405 billion parameters. Although significantly smaller, the $\text{BERT}_{\text{BASE}}$ model (by Google AI) used in this study still has about 110 millions parameters.

In order to use AI scoring capability responsibly, the scores and the scoring process should follow standards in educational testing. There are several entries in the testing standards jointly published by APA, AREA, and NCME that are specifically about AI scoring. For example, Standard 3.8 states that "(AI) scoring algorithms need to be reviewed for potential sources of bias. The precision of scores and validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups of the intended population" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). This standard highlights two core principles in responsible use of AI in educational assessment: AI scores should be accurate and AI scores should be fair. Most of published research to date on AI scoring using LLMs has emphasized prediction accuracy of the models with little discussion of fairness. In Johnson and Zhang (2024), the authors argued that the accuracy of AI is only one component of its responsible use in education and demonstrated that there may be inherent or implicit biases in LLMs that will lead to unfairness in AI scoring. In this study, we conducted an exploratory analysis to investigate whether choices of sampling methods can help mitigate biases in LLM-based AI models.

Statement of Research Problem

Experts from various disciplines have identified, examined, and discussed social, cultural, and gender biases present in pretrained LLMs; see Ayoub et al. (2024); Ma, Scheible, Wang, and Veeramachaneni (2023); Manvi, Khanna, Burke, Lobell, and Ermon (2024); Navigli, Conia, and Ross (2023); Bai, Wang, Sucholutsky, and Griffiths (2024), and Caton and Haas (2024), to name a few. Inherent biases in LLM models are deeply rooted in the data used for their training. These models absorb, internalize, and propagate any biases and stereotypes present in their training data sets, thereby making this issue rather complex. In their recent work, Johnson and Zhang (2024) found that GPT-4o can predict the racial/ethnic group membership of a writer of an essay response better than GPT-4o can score using a zero-shot approach. In order to improve prediction accuracy, a common practice is to fine-tune pretrained LLMs for downstream tasks. The fine-tuning process involves a selection of a pretrained model, preparation of the data, (iterative) model training, and evaluation of operational deployment in which preprocessing of the data is a critical step. Chu, Wang, and Zhang (2024) summarized four stages in the AI model development process that can be adjusted to mitigate inherent bias: (a) preprocessing, (b) in-training, (c) intraprocessing, and (d) postprocessing (in which the authors suggested "data augmentation" as one way to mitigate bias in the preprocessing stage). The goal of data augmentation is to ensure a balanced representation of training data across various subgroups (social, cultural, gender, age, religion, etc.) in the target population. In the field of machine learning, data augmentation, artificially increasing the size of a data set by applying transformations to the train data (Chhabra, Singla, & Mohapatra, 2022), is a common technique. In image recognition and computer vision, transformation techniques include rotating, flipping, or changing the contrast or brightness of images. In text classification, transformation techniques include random deletion or insertion (of words or characters),

sentence shuffling, synonym replacement, and so forth. Another approach to achieve a balanced training data is the equal-allocation stratified sampling technique, which effectively down-weights the larger subgroups by oversampling smaller subgroups in the population. Specifically, given a population P that can be divided into G nonoverlapping subpopulations or strata G_1, $G_{2\ldots}G_g$, a sample s_g of size n_g is taken within each stratum g independently from one stratum to another. Let $n = \sum_{g=1}^{G} n_g$ be the total sample size. In equal-allocation stratified random sampling, n_g is constant for each stratum, that is, $\forall h \quad n_g^{eq} = \frac{n}{G}$. In this study, we examined this equal-allocation sampling approach in fine-tuning LLMs for scoring. Our premise is that if an AI model training data set is imbalanced, meaning a subgroup of test takers is underrepresented, the model may struggle to make accurate predictions for the underrepresented subgroup. In survey sampling, proportional stratification leads to mean estimators (which may be thought of as human mean scores) that are more accurate than those obtained under simple random sampling given the same sample size, while equal-allocation stratified sampling ensures a minimum level of precision in each stratum but does not lead to the best global mean estimates, particularly when the variabilities (or human-score standard deviations) are different between strata (Lohr, 2021). In our current AI scoring scenario, we are not only interested in a model's overall performance, but also in its performance within specific subgroups to ensure fairness. Therefore we still prioritized the equal-allocation stratified random sampling technique and compared it to simple random sampling when constructing the AI model training samples. Given there were implicit biases in pretraining LLMs that we fine-tuned for our scoring tasks, equal-allocation sampling was arguably one method to strike a balance between prediction accuracy and fairness in the case of AI scoring. Finally, we note the lack of systematic analysis of the impact of sampling when applying an LLM-based AI scoring approach in the field of educational assessment. For instance, earlier work on sample-size requirements for automated scoring were mostly conducted prior to the era of LLMs. The amount of data required to fine-tune a pretrained LLM sufficiently for scoring purposes remains uncertain, and, to our best knowledge, there are no published studies addressing this issue. Generally speaking, the literature has indicated that effectiveness of fine-tuning is highly task-specific and is dependent on the model size and data quality. However, we believe it is still worthwhile to fill the gap in the literature and explore this aspect by using the same data source, which includes the same assessment task, test-taker population, and pretrained LLM. Specifically, we addressed two research questions in this study:

1. How well do AI models resulting from different sampling methods predict human scores?
2. To what extent are the AI models resulting from different sampling methods fair? Does stratified sampling help improve fairness?

Methods

Data Set

We used a data set collected from a standardized state science assessment in the United States between 2020 and 2021. There are 13 open-ended questions (or prompts) included in this analysis. All the prompts were graded by trained human raters on a 2-point integer scale: 0, 1, and 2. About 30% of the responses in each prompt were graded by a second human rater to monitor the reliability of the human scores. The response length in characters across prompts are shown in Table 1. By design, the responses are relatively short; on average, the number of characters are between 120 to 200 characters across prompts (about 20 to 40 words). The total number of responses in each prompt ranged from 2,458 to 2,531.

Table 1

Response Length by Item (Character Count Means and Standard Errors)

	Item												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Mean	179.7	196.5	190.6	137.1	150.9	174.8	172.9	122.6	119.7	100.9	132.2	144.8	187.1
S.E.	3.20	3.02	3.72	6.53	2.47	8.27	2.69	2.24	2.29	1.72	2.26	8.62	3.42

To investigate fairness, we focused on race/ethnicity in this study because previous research mostly raised concerns about AI models' performance across different racial/ethnic groups. Primarily due to the geographic location of the state assessment, the test takers were predominately identified in one of the following three race/ethnicity groups: White, Asian, or Hispanic/Latino, accounting for about 25%, 10%, and 50%, respectively, of the test-taker population. The remaining racial/ethnic groups (including Black/African American, American Indian or Alaskan Native, Native Hawaiian or other Pacific Islander, two or more races, or other) each accounted for less than 4% of the test-taker population; altogether they accounted for around 15% of the test-taker population. Due to the sample size of the smaller racial/ethnic groups, they were combined into a single group for sampling purposes. As seen in Table 2, the sample size distribution of the racial/ethnic groups was similar across prompts.

Table 2 also highlights the difference in performance across the groups. The test takers identified as Asian (denoted as Subgroup 3) had, on average, higher human mean scores than the test takers identified as White (denoted as Subgroup 1). The Hispanic/Latino test takers (denoted as Subgroup 2) received, on average, much lower human mean scores. Subgroup 4, which consisted of a mix of test takers from many racial/ethnic groups, had similar human mean scores, on average, as Subgroup 1 across prompts. This difference in performance might be due to differences in writing style, use of vocabulary, or even test-taking strategy and cultural background, among other factors. In the stratified sampling approach, which is described in the next section, the AI models were trained using samples with equal representation from all racial/ethnic groups.

Table 2

Human Mean Scores and Standard Deviations by Subgroup (Test Set)

	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4	
Item	*N*	Mean(S.D.)	*N*	Mean(S.D.)	*N*	Mean(S.D.)	*N*	Mean(S.D.)
1	241	0.81(0.81)	477	0.74(0.75)	115	1.24(0.82)	119	0.92(0.78)
2	233	1.00(0.92)	501	0.72(0.86)	94	1.03(0.90)	113	0.98(0.91)
3	251	0.57(0.69)	477	0.33(0.58)	116	0.86(0.81)	110	0.62(0.75)
4	260	0.53(0.76)	490	0.42(0.70)	97	0.86(0.85)	101	0.60(0.79)
5	251	0.86(0.85)	502	0.60(0.76)	110	1.17(0.83)	103	0.75(0.85)
6	264	0.84(0.83)	491	0.69(0.73)	100	1.26(0.77)	128	0.87(0.85)
7	267	0.65(0.75)	488	0.56(0.66)	91	0.93(0.81)	125	0.77(0.80)
8	251	0.59(0.74)	475	0.39(0.63)	114	0.96(0.80)	132	0.71(0.80)
9	253	0.52(0.75)	466	0.26(0.57)	97	0.82(0.88)	140	0.58(0.78)
10	233	0.47(0.69)	504	0.27(0.55)	110	0.65(0.72)	124	0.52(0.73)
11	254	0.42(0.62)	521	0.33(0.58)	75	0.84(0.84)	127	0.41(0.67)
12	244	0.71(0.81)	483	0.42(0.65)	109	0.97(0.87)	131	0.74(0.85)
13	233	0.70(0.82)	496	0.39(0.66)	102	1.01(0.92)	128	0.69(0.88)

Sampling and AI Model Building

Due to the content-specific nature of the items (that is, one item may be about global warming and another item may be about playing poker game), we built and evaluated AI models on an item basis (i.e., item-specific models). For each item or prompt, we first randomly selected and put aside 40% of the responses as the test set. The percentage of responses for the test set was meant to strike a balance so that even the smallest subgroup under investigation would have at least 100 responses independent from the model-building process for model evaluation. The test-set responses were untouched until the final model evaluation. The remaining 60% of the responses were used for model building and were further split into a training sample and a validation sample. Based on our research question, we compared two sampling approaches to construct the training sample: (a) simple random sampling and (b) equal-allocation stratified random sampling by race/ethnicity (each racial/ethnic group contributed equally to model training). For each prompt, we then used the training and validation samples to fine-tune a pretrained uncased $BERT_{BASE}$ model – one of the transformer-based pretrained LLMs – to predict human scores using deep learning neural networks (NN). AdamW was used as the optimizer in fine-tuning the hyperparameters of the NN models, with a learning rate set at *1e-5*. The batch size was set at 128 and training epoch was set at 25. The script was written in Python and was run on Amazon Web Services (AWS). The statistical analyses of the model performance were conducted on the author's local machine using Python. The model performance resulting from all sampling methods was compared and evaluated using the same test set.

Specifically, for the simple random sampling (denoted as "r" in the paper), two-thirds (66.7%) of the model-building data were used for training and the rest for validation. Of note is that the situation for model validation was slightly complex under stratified sampling due to the fact that (a) the sizes of the racial/ethnic groups were quite unbalanced and (b) after selecting the same number of responses from each racial/ethnic category for model training, the distribution of both human score and race/ethnicity in the remaining validation sample became rather different from the original sample. Therefore we investigated two variations on the validation sample: one simply using what was left after stratification (denoted as s1), knowing that this validation sample drastically differed from both the training sample and the original data, and the other resampling after stratification to match the subgroup (hence also score) distribution to the total sample (denoted as s2). As a result, under the s2 condition, the validation sample would have essentially the same score and subgroup distributions as the test set (which, as a reminder, is 40% of the whole sample). Specifically, we set a total sample of 560, or 140 per subgroup, in constructing the training sample in the s1 method to ensure that there were some responses left for validation in each subgroup. In implementing the s2 method, the (equal) sample size for each racial/ethnic group in the training sample was determined by 90% of the smallest racial/ethnic group. To construct the validation samples, all the remaining responses from the smallest racial/ethnic group were used while the sample sizes for other subgroups were determined according to their proportions in the test set. Because we forced the validation sample to emulate the test set, the larger subgroups for any prompt in the s2 method could be inevitably down-sampled quite a bit, resulting in a much smaller validation set overall.

To provide a full picture of the sampling result, Table 3 lists the final sample size for the training, validation, and test sets in each prompt. A few observations are worth noting. The sample size for the r training set was nearly twice the size of the training sets under the s1 and s2 methods. The training set sample size was similar between the s1 and s2 methods, but the validation sample was drastically reduced under the s2 method, ranging from only 121 to 182 across prompts, compared to 906 to 949 across prompts for the s1 method.

Table 3

Number of Responses in Training, Validation, and Test Sets

	r		s1		s2		
Item	training	validation	training	validation	training	validation	test
1	985	486	560	911	560	140	981
2	984	485	560	909	560	132	980
3	1,011	498	560	949	592	181	1,006
4	1,002	495	560	937	584	137	998
5	1,002	494	560	936	596	167	998
6	1,007	496	560	943	564	141	1,002
7	1,010	498	560	948	516	182	1,006
8	1,003	495	560	938	552	141	1,000

Table 3

Number of Responses in Training, Validation, and Test Sets (Continued)

	r		s1		s2		
Item	training	validation	training	validation	training	validation	test
9	1,007	496	560	943	528	133	1,003
10	978	482	554	906	484	130	974
11	988	488	557	919	540	121	985
12	981	484	555	910	488	128	978
13	1,006	496	560	942	568	157	1,002

Model Evaluation Metrics

To evaluate the accuracy and fairness of the AI scoring model, we followed the best practice suggested by ETS (McCaffrey et al., 2022). Specifically, for scoring accuracy, we examined quadratically weighted kappa (Cohen, 1968), disattenuated correlation, and standardized mean score differences (SMD) between human and AI scores on the test set. Additionally, we examined how well AI could predict the human true score using the proportional reduction in mean squared error (PRMSE) metric (Haberman, 2019; Loukina et al., 2020). The PRMSE is calculated as follows: $PRMSE = 1 - \frac{E(T-M)^2}{V(T)}$, where T is the human true score and M is the AI score. In the case of human scoring, true scores involve expected human ratings given the responses observed. But the variance of human true score cannot be directly estimated. But according to classical test theory, $V(T) = V(O) - V(e)$, where O is the observed score and e is the measurement error. Assuming measurement errors of the human ratings on the same essay are uncorrelated, we can use the agreement samples (responses with two human ratings) to estimate the variance of the measurement error of each prompt: $\hat{V}(e_k) = \sum_{i=1}^{r_k}(X_{ik} - X'_{ik})^2 / 2r_k$, where k is the prompt and r is the number of raters. Disattenuated correlations are calculated as: $d.R = R_{H,M}/\sqrt{R_{H,H}}$, where the numerator is the correlation of human score H and AI score M and the denominator is the correlation of the two human scores. Similar to PRMSE, disattenuated correlation attempts to evaluate prediction accuracy after removing noise in human ratings. Worth noting is that there is a fine distinction between prediction accuracy and agreement: According to Haberman (2019), kappa or QWK is a form of agreement metric and PRMSE is a metric of prediction accuracy. In the context of this study, we evaluated AI model performance on both metrics. The SMD is calculated as $SMD = (\bar{H} - \bar{M})/\sqrt{(s_H^2 + s_M^2)/2}$, where the mean differences between human score H and AI score M is divided by the pooled standard deviation of H and M. While SMD has been commonly suggested in the literature for evaluating the bias of AI models, one issue with SMD is that it can be sensitive to the differences in scales between human and AI scores. For fairness evaluation on the subgroup level, we

used the mean difference in standardized scores (MDSS) metric: $MDSS = \bar{H}' - \bar{M}'$, where H' and M' are standardized scores. The MDSS metric compares the human and AI mean scores by first removing their scales differences. MDSS is also the metric that we operationally use in practice for subgroup evaluation at the authors' organization. Additionally, for subgroup evaluation, we computed an adjusted mean score difference that was conditioned on human true score for each subgroup. The concept of this more recently developed metric is closely aligned to the concept of differential item functioning in psychometrics. That is, people with the same latent ability should have equal probability of getting a machine score, regardless of their group membership. Furthermore, this concept of predicted score M being conditionally independent of group membership G given the human true score $T(M \perp G \,|T)$ is termed "separation" fairness in the machine learning community. Hence for brevity, we denote this metric as the separation metric. The larger the separation is, the more the potential bias is for a given subgroup. The technical details of this metric can be found in Johnson, Liu, and McCaffrey (2022) and Johnson and McCaffrey (2023).

Results

Prediction Accuracy Results

Table 4 gives the means and standard deviations of the human score and AI score resulting from different sampling methods on the same test set. On the raw mean differences between human and AI scores, all differences are within a magnitude of 0.15. It is obvious that all the AI scores resulting from any sampling method have a slightly smaller standard deviation than the human scores. This minor scale shrinkage, however, does not appear to affect systematically the scoring accuracy and fairness of the AI scoring models.

Table 4

Human and AI Mean and Standard Deviations by Sampling Method (Test Set)

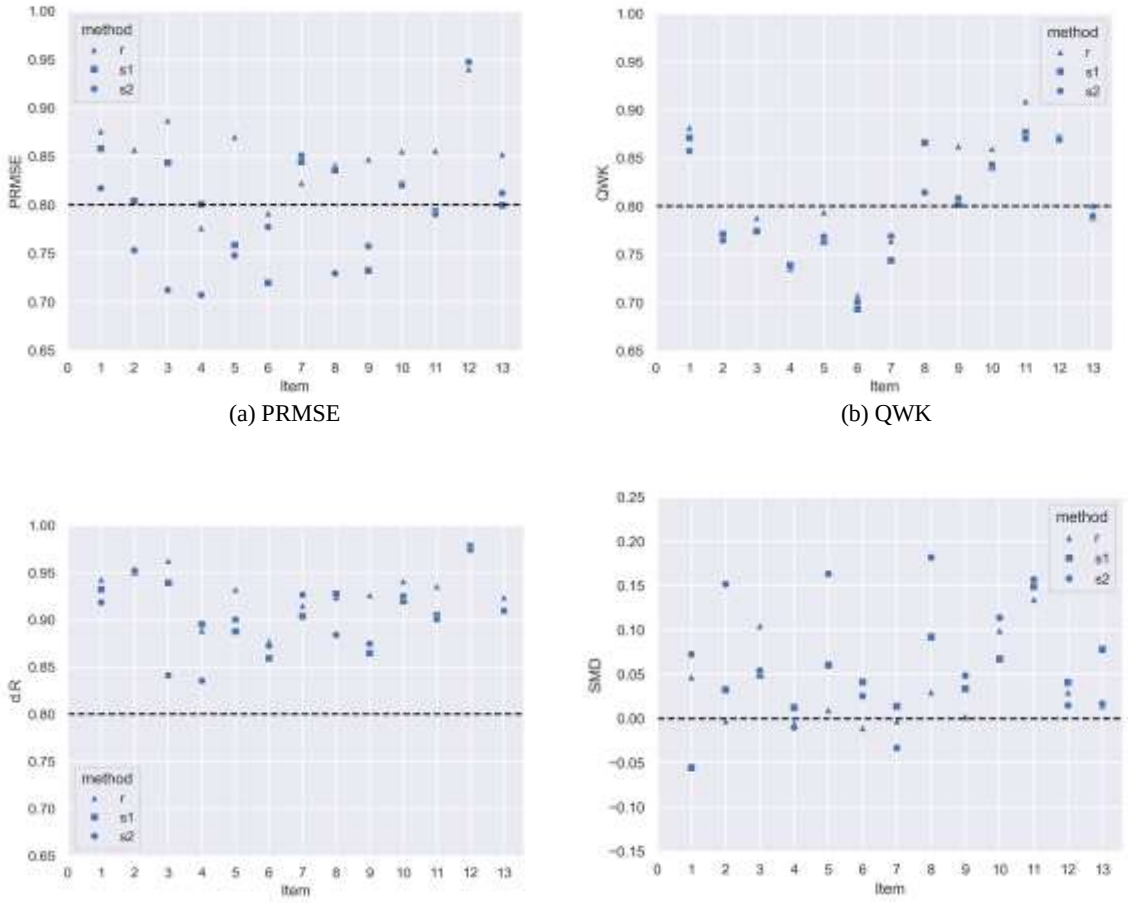
			Sampling method		
Item	Test set *N*	Human score	r	s1	s2
1	981	0.74(0.82)	0.78(0.82)	0.69(0.80)	0.80(0.82)
2	980	0.48(0.68)	0.48(0.59)	0.50(0.65)	0.58(0.68)
3	1,006	0.62(0.72)	0.70(0.66)	0.66(0.67)	0.66(0.47)
4	998	0.54(0.72)	0.54(0.64)	0.55(0.63)	0.53(0.47)
5	998	0.42(0.70)	0.42(0.57)	0.46(0.61)	0.53(0.64)
6	1,002	0.38(0.64)	0.37(0.57)	0.41(0.58)	0.40(0.56)
7	1,006	0.39(0.63)	0.39(0.58)	0.40(0.52)	0.37(0.55)
8	1,000	0.55(0.78)	0.57(0.75)	0.62(0.77)	0.69(0.75)
9	1,003	0.58(0.76)	0.58(0.71)	0.60(0.68)	0.61(0.67)
10	974	0.83(0.89)	0.91(0.87)	0.89(0.82)	0.93(0.85)
11	985	0.82(0.79)	0.93(0.79)	0.94(0.77)	0.94(0.75)
12	978	0.50(0.75)	0.52(0.68)	0.53(0.68)	0.51(0.69)
13	1,002	0.80(0.80)	0.81(0.70)	0.86(0.72)	0.81(0.71)

The standardized mean score differences (SMD) between human and AI scores, shown in Figure 1(d), suggest that all SMDs resulting from r and s1 methods are within the magnitude of 0.15 – a threshold value suggested in the literature (McCaffrey et al., 2022; Williamson, Xi, & Breyer, 2012). However,

the s2 method showed slightly larger SMDs on four items (i.e., Items 2 and 11 on the borderline of 0.15 and Items 5 and 8 in between 0.15 and 0.20), where the AI scores have overall higher means than human scores. This result indicates that the smaller validation sample in the implementation of the s2 method, even though "matched" to the subgroup distributions in the test set, seemed to have an impact on model performance, in particular the AI score means. Even though the validation samples for the s1 method are "not matched" to the test set, there are a much greater number of responses representing each racial/ethnic group. In other words, prioritizing a larger validation sample may be more crucial than to achieve a distributional "match" by sacrificing sample size to AI model performance.

Figure 1

Results of Prediction Accuracy



(a) PRMSE

(b) QWK

(c) Disattenuated Correlation

(d) SMD

Included in Figure 1 are the results for other evaluations on prediction accuracy, that is, PRMSE, QWK, and disattenuated correlation (denoted as "d.R"), between human and AI scores resulting from different sampling methods. All PRMSE statistics were greater than 0.7, which is considered a minimum performance threshold for AI scoring models (McCaffrey et al., 2022). Among the lower PRMSEs, such as those in between 0.7 and 0.8, most resulted from the s2 method and some resulted from s1. Previous research suggested $QWK >= 0.7$ when evaluating automated scoring models (Williamson et al., 2012). In this analysis, all QWK values were greater than 0.7 with the exception of one instance: the $QWK = 0.694$ on Item 6 resulting from the s2 method. The QWKs were also on the borderline of 0.7 for the other two sampling methods. One speculation for why this happened is that the standard deviation of the human scores on this item is small (S.D. = 0.64), which could have an impact on the AI model

building and evaluation. The d.Rs were all above 0.83, with $d.R^2$ all larger than the threshold of 0.7 suggested in McCaffrey et al. (2022).

Overall, the AI scoring models based on all three sampling methods demonstrated reasonably good performance. While the intention to match the validation sample to the test set in the s2 method was to enhance the AI model performance, empirical evidence did not support that decision. AI models based on the simple random sampling (s1) showed the best performance in many cases. Interestingly, in s1, it worked well to use a much smaller (i.e., about half of the size) model training sample with equal subgroup representations but with a much larger validation sample that was different from the test set in terms of score and subgroup distributions. Even though the s2 method did not outperform the r method, the model performance, in general, was in fact quite acceptable.

Fairness Results

Figures 2 shows the results of the separation metric, which evaluated the human-AI mean score differences conditional on true score for each racial/ethnic group. All of the values in Figure 5 were within 0.2, with the majority of the values within 0.1. This result means that, for a given subgroup conditional on the true score, the AI mean scores only differed from human mean scores by less than one tenth of a score point on the 2-point scale. These differences could be considered negligible. The only notably larger separation between human and AI scores was for Item 7, which all the AI models underscored for Subgroup 3 (the Asian test-taker group) on average. In this case, the s1 method notably outperformed the r and s2 methods by better predicting the means of Subgroup 3.

Figure 2

Separation Results by Subgroup

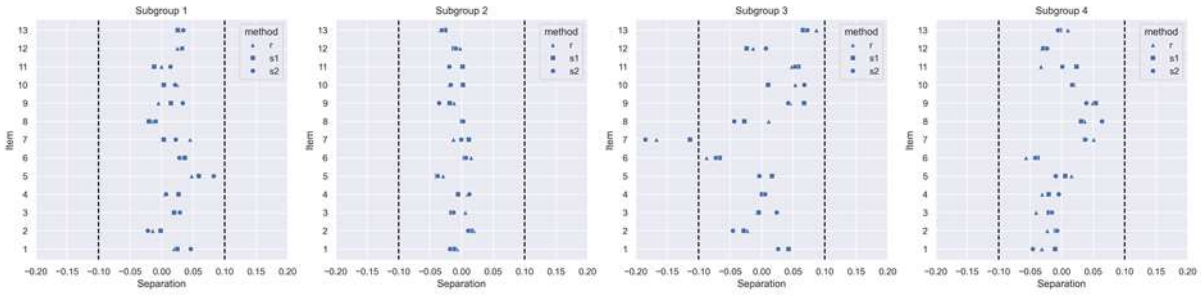
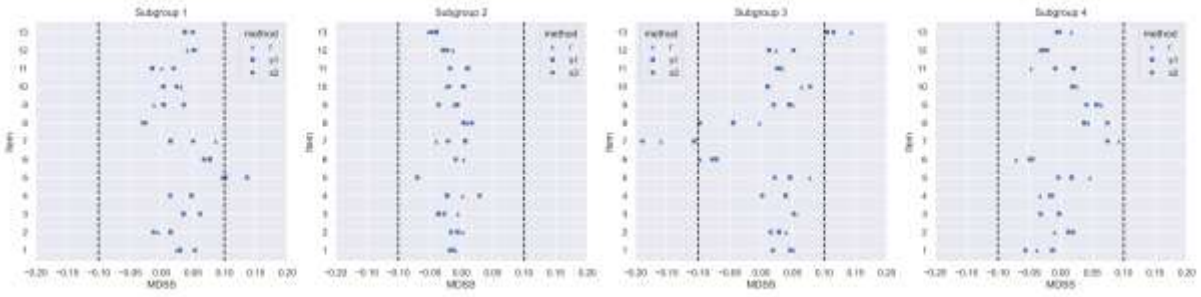


Figure 3

Mean Differences in Standardized Scores (MDSS) by Subgroup



The results on the MDSS for the racial/ethnic groups are shown in Figure 3. The findings are similar to the separation metric. All the mean differences were within the magnitude of 0.2. While there is no established or recommended evaluation threshold for this metric, we applied 0.20 as a common in-house threshold for test operations. As with the separation metric results, the largest MDSS values were

associated with Subgroup 3 on Item 7. It is interesting to observe that the s1 method tended to outperform simple random sampling, especially on cases that had larger MDSS values (e.g., Subgroup 1 on Item 5, Subgroup 3 on Item 7, and Subgroup 3 on Item 13).

Discussion

In this exploratory study, we empirically evaluated the impact of sampling on AI scoring model performance. Simple random sampling (r) and equal-allocation stratification random sampling were compared in constructing the AI model training samples. Two variations of the sampling strategy (s1 and s2) in constructing the validation sample in the AI modeling building process were further examined under the equal-allocation stratification random sampling. Fine-tuned LLMs were trained and evaluated, and all the AI models were prompt-specific for the 13 items included in this analysis. We summarized the characteristics of the samples under each sampling method in Table 5.

Table 5

Sample Characteristics

Method	Training	Validation	Test (same across methods)
r	$N \approx 1000$; representative of whole population; dominated by large subgroups.	$N \approx 500$; representative of whole population; dominated by large subgroups.	$N \approx 1000$; representative of whole population; dominated by large subgroups.
s1	$N \approx 550$; smaller in size; equal contribution from each subgroup	$N \approx 900$; very large in size; very different from whole population even more dominated by large subgroups.	$N \approx 1000$; representative of whole population; dominated by large subgroups.
s2	$N \approx 550$; smaller in size; equal contribution from each subgroup	$N \approx 140$; very small in size; representative of whole population; dominated by large subgroups.	$N \approx 1000$; representative of whole population; dominated by large subgroups.

In response to the research questions, the models were evaluated from two perspectives: overall prediction accuracy and fairness. For RQ1, we found that, in general, the AI scoring models predicted human scores reasonably well regardless of the sampling method. Even when the training sample size was relatively small as in s1 and in s2 compared to r, or when the validation sample was extremely small (as in s2) or relatively large (as in s1), the model performance was marginally affected and was comparable across methods. Even though the AI models appeared to perform slightly worse using the s2 method, the observation was only on the SMD index for three out of the 13 prompts while the evaluations did not reveal other obvious issues for the s2 method. In addressing RQ2, we found that using model training samples with equal representation from subgroups of test takers (s1 and s2) did not systematically improve the fairness of the AI scoring models. In almost all cases, models based on simple random sampling were fair across the different racial/ethnic groups. In a couple of rare cases where models resulting from the r method did not work as well (Subgroup 3 on Items 7 and 13), stratification appeared to have improved fairness. From a big picture point of view, however, equal-allocation stratification did not improve, or worsen, fairness. One could argue, though, that stratification is a critical and early treatment to mitigate bias in the data preprocessing step during model development

process (Chu et al., 2024) in the sense that all subgroups of interest contributed equally in the model training process. The model is not dominated by inherent biases associated with any specific subgroup of interest. The results may also arguably favor stratification given that both the training and validation samples can be relatively small and the validation sample does necessarily need to resemble the test-taker population (s1). These potential advantages on sample requirements seem especially useful when only small data set is available for AI model building.

There is relatively little prior research that specifically focused on sample size requirements for AI scoring with few exceptions such as Haberman and Sinharay (2008), Zhang (2013), and Heilman and Madnani (2015). To our best knowledge, most of the former work was conducted with the earlier generation of AI scoring practice (pre-LLM era) when well-defined features were used in less complex, but more explainable AI scoring models such as logistic regression or multiple linear regression. The authors in investigated the training sample sizes in AI scoring of short-response items using support vector regression models where the predictors included various word n-grams and a proxy of response length. Their findings (Figure 1 in the refereed article) showed that from small training sample size of 100 to larger training sample sizes of 200, 400, 800, 1600, and up to 3200, the scoring model performance as evaluated by QWK steadily and considerably improved. This study found different results related to sample sizes from the prior work, mostly likely due to the use of fine-tuned LLMs. In Heilman and Madnani (2015), about half of the items achieved a human-AI QWK of 0.7 or greater and required at least 800 model training responses. Even when the training sample sizes were as large as 1600 or 3200, it appeared difficult to achieve a QWK of 0.8 and above. In the current study, LLM-based AI scoring models achieved QWK of 0.75 or above on 11 of 13 prompts, regardless of the sampling method. This result aligns well with the literature in the AI community in that complex AI models such as NN tended to achieve greater accuracy in prediction tasks than simpler models such as SVM or decision trees.

Even though the most of the LLM-based AI scoring models demonstrated high prediction accuracy and an acceptable degree of fairness, there still seems to be room for improvement. The top performing models reported in Whitmer et al. (2021) achieved average human-AI QWK ranging from 0.860 to 0.888 across NAEP Reading items, about 0.05 points higher on average than the QWKs reported in this study. Most of the top performing models in the NAEP study were either ensembles of multiple models or leveraged information in the prompt and source text. The total samples in the NAEP study ranged from 19,934 to 28,307 across items (Whitmer et al., 2021), which are much larger than the samples per item available in this study. So it is highly likely that we can further improve the current AI model performance on these items when we collect more responses in test operations. By improving the overall model prediction accuracy, fairness will likely be improved accordingly. In this analysis, we did not customize the model fine-tuning process for each item; instead, we applied the same setting for all items. Customizing the fine-tuning process will most likely improve model performance on the item level as well.

Overall, this study offers some empirical evidence on the choice of sampling methods in building LLM-based AI scoring models for short-response assessment items. For the items investigated in this study, a training sample size of 1000 from simple random sampling was generally sufficient. We found the models based on stratified samples performed comparably to models based on simple random samples. However, it is worth noting the stratified training samples were only half of the size. For testing programs that intend to prioritize fairness in the AI model training process, stratified sampling can be seriously considered.

This study has several limitations. One is that the $BERT_{BASE}$ LLMs were fine-tuned with minimum effort. The same settings were using across prompts. It is possible that differences in prediction accuracy and fairness may emerge along with more optimal fine-tuning such as adjusting the learning rate in each item. The results may not generalize to LLMs beyond $BERT_{BASE}$, making a comparative analysis worthwhile in the future. We also did not consider intersections of demographic variables (e.g., gender by race/ethnicity, language skill by gender), which future research is encouraged to explore. Additionally, as a follow-up of the current analysis, stratified sampling by race/ethnicity *and* score levels can provide more useful results on improving fairness in the model-training process. For example,

selecting the equal number of responses from each racial/ethnic group at each score level essentially makes the (human) scores orthogonal to race/ethnicity and, as a result, any detected biases in the machine scores would be due to other reasons than one's race or ethnicity alone. This is a natural next step once more responses are collected. Due to the limited responses in some racial/ethnic groups, equal-allocation stratified random sampling on student-written responses could only utilize a relatively small sample. Future research can consider augmenting the data set with synthetic data (e.g., using GPT, for underrepresented subgroups). Alternatively, future research may also apply techniques that algorithmically mitigate bias in the training data, such as sample reweighting and adopting fairness-aware machine learning models (Ferrara, Sellitto, Ferrucci, et al., 2024; Haberman, 1984). Finally explaining detected biases is challenging with complex AI scoring models. Johnson and McCaffrey (2023) proposed one method to weight AI features differently to reduce subgroup biases in simpler models; future research is encouraged to generalize the method in Johnson and McCaffrey (2023) to LLM-based AI scoring.

Declarations

Gen-AI Use : The authors of this article declare (Declaration Form #: 2611241949) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Author Contribution: M. Zhang and M. Johnson conceptulized the study and wrote the manuscript. M. Zhang and C. Ruan conducted the modeling and statistical analysis.

Conflict of Interest: None

Ethical Approval: Not applicable.

References

Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*(C). Retrieved from https://doi.org/10.1016/j.inffus.2023.101805

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ayoub, N. F., Balakrishnan, K., Ayoub, M. S., Barrett, T. F., David, A. P., & Gray, S. T. (2024). Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proceedings: Digital Health*, *2*, 186–191. Retrieved from https://doi.org/10.1016/j.mcpdig.2024.03.003

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). *Measuring implicit bias in explicitly unbiased large language models*. arXiv. Retrieved from https://arxiv.org/pdf/2402.04105

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology in testing: Measurement issues* (pp. 142–173). Taylor & Francis.

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), Article 166. Retrieved from https://doi.org/10.1145/3616865

Chamieh, I., Zesch, T., & Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In E. Kochmar et al. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 309–315). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.bea-1.25.pdf

Chhabra, A., Singla, A., & Mohapatra, P. (2022). Fair clustering using antidote data. In J. Schrouff, A. Dieng, M. Rateike, K. Kwegyir-Aggrey, & G. Farnadi (Eds.), *Proceedings of the algorithmic fairness through the lens of causality and robustness* (Vol. 171, pp. 19–39). PMLR. Retrieved from https://proceedings.mlr.press/v171/chhabra22a.html

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter*, *26*(1), 34–48. Retrieved from https://doi.org/10.1145/3682112.3682117

Cohen, J. (1968). Weighted kappa: Normal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. http://dx.doi.org/10.1037/h0026256

Ferrara, C., Sellitto, G., Ferrucci, F., et al. (2024). Fairness-aware machine learning engineering: How far are we? *Empirical Software Engineering*, *29*(9). Retrieved from https://doi.org/10.1007/s10664-023-10402-y

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*(3), 971–988. Retrieved from https://www.jstor.org/stable/2240973

Haberman, S. J. (2019). *Measures of agreement versus measures of prediction accuracy* (Research Report No. RR-19-20). Retrieved from https://doi.org/10.1002/ets2.12258

Haberman, S. J., & Sinharay, S. (2008). *Sample-size requirements for automated essay scoring* (Research Report No. RR-08-32). Retrieved from https://doi.org/10.1002/j.2333-8504.2008.tb02118.x

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 81–85). Retrieved from https://doi.org/10.3115/v1/W15-0610

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, *59*, 338–361. Retrieved from https://doi.org/10.1111/jedm.12335

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment.* Routledge.

Johnson, M. S., & Zhang, M. (2024). *Examining the responsible use of zero-shot AI approaches to scoring essays.* Manuscript submitted for publication.

Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, *4*(47). Retrieved from https://doi.org/10.1007/s44163-024-00147-y

Kumar, A., Dikshit, S., & de Albuquerque, V. (2021). Explainable artificial intelligence for sarcasm detection in dialogues. *Wireless Communications and Mobile Computing*, *1*, 1–13. Retrieved from https://doi.org/10.1155/2021/2939334

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lohr, S. L. (2021). *Sampling: Design and analysis* (3rd ed.). Chapman and Hall/CRC. Retrieved from https://doi.org/10.1201/9780429298899

Loukina, A., Madnani, N., Cahill, A., Yao, L., Johnson, M. S., Riordan, B., & McCaffrey, D. F. (2020). Using PRMSEs to evaluate automated scoring systems in the presence of label noise. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, H. Y. Ildikó Pilán, & T. Zesch (Eds.), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 18–29). Retrieved from https://doi.org/10.18653/v1/2020.bea-1.2

Lubis, F. F. M., Putri, A. W. D., et al. (2021). Automated short-answer grading using semantic similarity based on word embedding. *International Journal of Technology*, *12*(3), 571–581. Retrieved from https://doi.org/10.14716/ijtech.v12i3.4651

Ma, W., Scheible, H., Wang, B., & Veeramachaneni, G. (2023). Deciphering stereotypes in pre-trained language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 11328–11345). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2023.emnlp-main.697

Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). *Large language models are geographically biased*. arXiv. Retrieved from https://arxiv.org/abs/2402.02680

McCaffrey, D. F., Casabianca, J., Ricker-Pedley, K. L., Lawless, R., & Wendler, C. (2022). *Best practices for constructed-response scoring* (Research Report No. RR-22-17). Retrieved from https://doi.org/10.1002/ets2.12358

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, *15*(2), 1–21. Retrieved from https://doi.org/10.1145/3597307

Oka, R., Kusumi, T., & Utsumi, A. (2024). Performance evaluation of automated scoring for the descriptive similarity response task. *Nature Scientific Reports*, *14*, Article 6228. Retrieved from https://doi.org/10.1038/s41598-024-56743-6

Whitmer, J., Deng, E. Y., Blankenship, C., Beiting-Parrish, M., Zhang, T., & Bailey, P. (2021). *Results of NAEP reading item automated scoring data challenge (fall 2021)*. EdArXiv. Retrieved from https://osf.io/preprints/edarxiv/2hevq

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. Retrieved from https://doi.org/10.1111/j.1745-3992.2011.00223.x

Zhang, M. (2013). *The impact of sampling approach on population invariance in automated scoring of essays* (Research Report No. RR-13-18). https://doi.org/10.1002/j.2333-8504.2013.tb02325.x