

Effects of Demographic Representations and Model Training Methods on Automated Scoring Engines

Yangmeng Xu, Martha Bellows and Edward W. Wolfe

Pearson

Abstract

This study evaluated whether automated scoring engines maintain stable performance when training data composition and training methods vary. We manipulated demographic representation (gender, English language learner, race, student with disabilities) and compared feature-based versus transformer-based models. Results showed performances were stable across subgroup-representation densities and transformer models exhibited greater stability.

1 Introduction

Automated scoring (AS) engines increasingly support large-scale educational assessments, yet concerns persist regarding their performances across demographic subgroups. While previous research has documented AS performance, limited evidence exists regarding how training data composition and model architecture influence scoring stability across populations. This study addresses this gap by systematically examining whether demographic representation in training data and choice of modeling approach during training affect the performance of AS engines.

We evaluated the Intelligent Essay Assessor™ (IEA), Pearson's proprietary automated scoring technology, across four demographic variables: gender, English language learner (ELL) status, race/ethnicity, and students with disabilities (SWD) status. We used two statistical indices to quantify the engine performance: Quadratic Weighted Kappa (QWK) measured agreement between automated scores and human scores (used as reference scores), while Standardized Mean Difference (SMD) assessed differences in their mean scores. These metrics enable comprehensive evaluation of both scoring stability and potential bias across subgroups.

2 Research Questions

- 1) To what extent does variation in subgroup representation within training data affect the performance of AS engines?
- 2) How does the choice of training method (feature-based vs. transformer-based) affect the stability of subgroup results?

3 Method

3.1 Data

Data comprised 508,305 student responses from operational English Language Arts (ELA) administrations (2022–2025) across five prompts spanning grades 5 and 8. Response types included Research Simulation (RST), Response to Literature (LAT), and Narrative (NWT) tasks, scored on two traits: Reading Comprehension/Written Expression and Conventions. The scores range from 0-4 for RC/WE on RST and LAT tasks and from 0-3 on Con tasks and NWT RC/WE tasks.

3.2 Design

We conducted an experiment in which scores assigned by human raters (i.e., scorers) to student responses to constructed-response prompts were sampled purposively to create engine training data files composed of different concentrations of demographic subgroups. These data files were used to train an AS engine, and the predicted scores from the resulting AS models were evaluated for differences attributable to training sample composition in the context of standard AS subgroup analyses. The human scores and the IEA predicted scores that came from each model were used to conduct standard subgroup analyses. The values of the resulting statistics were compared across training methods and levels of subgroup representation in the training data to evaluate the invariance of automated scoring results and to determine whether observed

patterns differed systematically across these conditions.

3.3 Variables

Independent Variables

Demographic Representation: Four demographic variables were examined with two or three groups each. For binary variables (Gender, ELL, SWD), training data concentrations (i.e., densities) were created at 10/90, 30/70, 50/50, 70/30, and 90/10 splits (%). For race/ethnicity (three groups: Black, Hispanic, White), concentrations were 30/55/15, 55/15/30, and 15/30/55 (%).

Training Approaches: We trained the IEA engine using one of two approaches, where the “feature-based” training method refers to predictive modeling such as regression analysis in which features of the written response are differentially weighted and combined to predict a human score, and the transformer (i.e., deep neural network) training method refers to a complex computational model that consist of multiple linear layers (each composed of multiple nodes) that are interconnected by non-linear functions which take as inputs the outputs of the previous layer resulting in increasing complexity as information propagates through subsequent layers. Specifically, for the feature-based approach, content features were selected for Reading Comprehension/Written Expression and grammar/mechanics features for Conventions, with all features fed into an ensemble of Random Forest, XGBoost, and SVM models for score prediction. For the transformer approach, a DeBERTa V3 model (He et al., 2023) was trained using 48 randomly selected hyperparameter sets within specified bounds for learning rate, weight decay, epochs, scheduler, and warm-up ratio.

Dependent Variables

Each of the following indices was computed and compared by subgroups using a standard subgroup analysis with industry standard flagging thresholds (Williamson et al., 2012).

QWK quantified agreement between IEA and human scores (accepted threshold: ≥ 0.70); ΔQWK measured agreement differences between IEA-human and human-human comparisons (accepted threshold: $QWK_{IH} - QWK_{HH} \geq -0.10$);

$|SMD|$ assessed standardized mean differences between IEA and human scores (accepted threshold: ≤ 0.10).

We used the “best” human scores for all training and test datasets and for the human scores in IEA-human comparison data. Specifically, IEA prioritizes the available humans scores and identifies “best” available human score for each student response. For example, when multiple human scores exist for a single response, scores assigned by expert human scorers are prioritized over scores assigned by a single scorer from the pool of scorers. Scores from two randomly selected humans from a pool of scorers were used for all human-human comparison data.

Note that we also reported both within-difference and between-difference if needed when comparing these metrics. Within-difference is defined as $(A-B)-(C-D)$ while between-difference is defined as $Mean_{AandB} - Mean_{CandD}$. A and B are the values for two subgroups under the same group while C and D are the values for two subgroups under another.

Controlled Variables

Training datasets comprised 6,000 responses (3,000 training, 3,000 validation) maintaining consistent demographic composition. Test datasets contained 600 responses per demographic group.

4 Results

4.1 Demographic Representation Effects

Table 1 displays descriptive statistics of the values of QWK_{IH} , ΔQWK and SMD_{IH} across prompts, training methods and traits at each density level of demographic group representation in the engine training dataset. For example, Row 1 indicates the metric performance for female students when the model is trained on data containing 10% female and 90% male students, whereas Row 2 shows performance for male students when the model is trained on 10% male and 90% female data.

The table also reports the average flag rates for each metric. Flag rates represent the percentage of combinations that do not meeting pre-defined thresholds (i.e., $QWK < 0.70$, $\Delta QWK < -0.10$, or $|SMD| > 0.10$) across five prompts, two training

methods, subgroup categories (i.e., two gender groups, two ELL groups, three race groups, two SWD groups, or two traits). For example, there are $5 \times 2 \times 2 \times 2 = 40$ total combinations for gender comparisons.

The result indicates that including different densities for demographic subgroups in model training has minimal influence on model performance. Specifically, QWK values varied by less than 0.04 across density levels, with largest variability in ELL (0.04 within-difference between 30% and 70% densities as highlighted in yellow; $[(0.84-0.76)-(0.82-0.78) = 0.04]$). With respect to flag rates, the values were almost universally zero, with one slight deviation (yellow highlights).

Moreover, ΔQWK deviations remained small (absolute value ≤ 0.05), with flag rates nearly universally zero except for ELL, where 17% of models flagged at all densities (green highlights) except 90%, which produced zero flags. SMD remained near zero across densities (largest absolute difference: 0.05). SMD flag rates were minimal except for ELL and gender subgroups at specific densities (blue highlights). In total, a small number of models trained with $\leq 70\%$ subgroup representation showed higher flag rates, but these were minor and inconsistent. Models trained with higher subgroup inclusion exhibited no flagged conditions.

4.2 Training Method Comparisons

Table 2 shows descriptive statistics of the values of QWK_{IH} , ΔQWK and SMD_{IH} and their associated average flag rates across prompts, traits and density levels of demographic representation for each engine training method (Feature and Transformer). The results indicate that transformer models performed slightly better than feature-based models.

Transformer models exhibited slightly higher QWK values, with the largest difference occurring for ELL students (0.03 within-approach difference and 0.02 between-approach difference as highlighted in yellow). Flag rates for QWK were universally zero except for 4% for Hispanic using feature-based models and 3% for SWD using transformer models.

ΔQWK deviations remained small (absolute

value ≤ 0.05). 29% of ELL feature-based models produced $\Delta QWK < -0.10$ (green highlights), while transformer-based ELL models produced zero such flags.

Standardized mean differences were minimal across both training methods ($|SMD| \leq 0.05$), with no substantial within-differences except for ELL students (blue highlights). Training methods by demographic group interactions emerged for ELL and SWD, with flag rates of 7% for both ELL groups under feature-based and transformer approaches, respectively, and 9% for non-SWD students using transformer modeling (blue highlights).

5 Discussion

The purpose of this study is to evaluate whether AS engines maintain stable performance when training data composition and training methods vary. Results demonstrate that IEA’s scoring performance is stable across demographic representation densities and training methods. Subgroup-related differences were small in magnitude, inconsistent in direction, and fully explainable by data-representation factors or expected random fluctuation.

RQ 1: To what extent does variation in subgroup representation within training data affect the performance of AS engines?

Results from the training-density analyses showed that subgroup concentration exerted minimal influence on model results in most conditions. Differences in QWK and SMD values across densities were small (≤ 0.05) and inconsistent in direction. Deviations appeared mainly under certain extreme subgroup under-representation and disappeared once balanced data were used in most cases.

We acknowledge that there may be confounding factors related to ELL status and racial background. Controlling for these potential confounders may represent an important direction for future research.

RQ 2: How does the choice of training method (feature-based vs. transformer-based) affect the stability of subgroup results?

The feature-based and transformer models yielded broadly similar results. The transformer configuration exhibited slightly lower variability

across subgroups and eliminated a small number of ELL-related flag conditions found in the feature-based models. The magnitude of these differences was modest, suggesting that both approaches perform comparably, with transformer models offering a small consistency advantage.

6 Conclusion

This study shows that AS engines can maintain stable performance across demographic subgroups under varied training conditions. We found that the densities of different demographic representations in the model training dataset minimally affect scoring performances except under extreme under-representation conditions for most subgroup variables. Moreover, transformer and feature-based models perform comparably, with transformer models showing slight consistency advantages. In summary, AS performance aligns with human rater variability. The results support adopting transformer-based models as the operational standard to maximize scoring stability.

7 References

He, P., Gao, J., & Chen, W. (2023). DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing (No. arXiv:2111.09543). arXiv. <https://doi.org/10.48550/arXiv.2111.09543>

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

Table 1: Statistical Indicators by Density Across Demographic Groups

Variable	Density	Group	QWK _{LH}				Δ QWK				SMD _{LH}			
			Min	Ave	Max	Flag	Min	Ave	Max	Flag	Min	Ave	Max	Flag
Gender	10	Female	0.76	0.83	0.88	0.00	-0.05	-0.01	0.06	0.00	-0.03	0.03	0.13	0.11
		Male	0.76	0.85	0.88	0.00	-0.05	-0.02	0.03	0.00	-0.04	0.00	0.05	0.00
	30	Female	0.76	0.83	0.88	0.00	-0.04	-0.01	0.05	0.00	-0.04	0.02	0.07	0.00
		Male	0.71	0.84	0.88	0.00	-0.08	-0.02	0.02	0.00	-0.04	0.01	0.06	0.00
	50	Female	0.76	0.83	0.89	0.00	-0.05	-0.01	0.06	0.00	-0.04	0.04	0.10	0.05
		Male	0.72	0.83	0.88	0.00	-0.07	-0.02	0.02	0.00	-0.02	0.03	0.12	0.11
	70	Female	0.76	0.83	0.89	0.00	-0.04	-0.01	0.05	0.00	-0.04	0.02	0.06	0.00
		Male	0.71	0.84	0.88	0.00	-0.08	-0.02	0.02	0.00	-0.03	0.01	0.06	0.00
	90	Female	0.77	0.84	0.89	0.00	-0.05	-0.01	0.05	0.00	-0.06	0.02	0.09	0.00
		Male	0.72	0.84	0.89	0.00	-0.07	-0.02	0.02	0.00	-0.03	0.02	0.06	0.00
ELL	10	Non-ELL	0.76	0.83	0.88	0.00	-0.07	-0.04	0.00	0.00	-0.12	0.05	0.23	0.19
		ELL	0.71	0.77	0.84	0.00	-0.12	-0.04	0.06	0.17	-0.02	0.02	0.11	0.17
	30	Non-ELL	0.76	0.84	0.88	0.00	-0.07	-0.03	0.00	0.00	-0.05	0.03	0.08	0.00
		ELL	0.70	0.76	0.81	0.00	-0.13	-0.05	0.03	0.17	-0.07	0.00	0.04	0.00
	50	Non-ELL	0.73	0.83	0.87	0.00	-0.05	-0.03	0.01	0.00	-0.05	0.04	0.17	0.06
		ELL	0.73	0.78	0.80	0.00	-0.10	-0.04	0.02	0.17	-0.07	0.00	0.03	0.00
	70	Non-ELL	0.70	0.82	0.88	0.05	-0.05	-0.02	0.00	0.00	-0.01	0.03	0.08	0.00
		ELL	0.73	0.78	0.81	0.00	-0.11	-0.03	0.03	0.17	-0.02	0.02	0.05	0.00
	90	Non-ELL	0.75	0.83	0.88	0.00	-0.05	-0.02	0.00	0.00	-0.03	0.02	0.08	0.00
		ELL	0.75	0.79	0.83	0.00	-0.08	-0.02	0.05	0.00	-0.02	0.00	0.02	0.00
Race	15	Black	0.72	0.82	0.88	0.00	-0.06	-0.03	0.02	0.00	-0.07	0.01	0.06	0.00
		Hispanic	0.71	0.83	0.88	0.00	-0.07	-0.03	0.02	0.00	-0.02	0.02	0.04	0.00
		White	0.71	0.83	0.88	0.00	-0.05	-0.02	0.02	0.00	-0.07	0.01	0.08	0.00
	30	Black	0.73	0.83	0.89	0.00	-0.05	-0.02	0.03	0.00	-0.07	0.01	0.08	0.00
		Hispanic	0.67	0.82	0.87	0.06	-0.09	-0.03	0.01	0.00	-0.03	0.01	0.13	0.06
		White	0.72	0.83	0.88	0.00	-0.05	-0.02	0.03	0.00	-0.05	0.02	0.11	0.06

	55	Black	0.73	0.82	0.89	0.00	-0.05	-0.02	0.02	0.00	-0.06	0.01	0.06	0.00
		Hispanic	0.71	0.83	0.88	0.00	-0.07	-0.03	0.02	0.00	-0.04	0.02	0.06	0.00
		White	0.70	0.83	0.89	0.00	-0.05	-0.02	0.01	0.00	-0.06	0.01	0.12	0.06
SWD	10	Non-SWD	0.73	0.83	0.88	0.00	-0.06	-0.02	0.05	0.00	0.01	0.04	0.11	0.06
		SWD	0.72	0.85	0.88	0.00	-0.08	-0.02	0.02	0.00	-0.05	-0.01	0.04	0.00
	30	Non-SWD	0.72	0.83	0.89	0.00	-0.05	-0.02	0.03	0.00	-0.01	0.04	0.13	0.06
		SWD	0.74	0.83	0.88	0.00	-0.05	-0.02	0.03	0.00	-0.08	-0.01	0.04	0.00
	50	Non-SWD	0.75	0.83	0.88	0.00	-0.05	-0.01	0.06	0.00	-0.04	0.02	0.09	0.00
		SWD	0.74	0.84	0.88	0.00	-0.05	-0.02	0.03	0.00	-0.05	-0.01	0.03	0.00
	70	Non-SWD	0.70	0.83	0.88	0.00	-0.05	-0.01	0.03	0.00	-0.02	0.03	0.13	0.06
		SWD	0.70	0.84	0.88	0.07	-0.06	-0.02	0.02	0.00	-0.05	0.00	0.06	0.00
	90	Non-SWD	0.73	0.83	0.88	0.00	-0.07	-0.02	0.03	0.00	-0.06	0.02	0.11	0.06
		SWD	0.72	0.83	0.88	0.00	-0.07	-0.03	0.01	0.00	-0.05	0.01	0.05	0.00

Note. I_H indicates comparisons between IEA and human (i.e., best human as defined in this study). H_H indicates comparisons between human and human (i.e., two randomly selected humans from a pool of raters). QWK_{I_H} indicates the QWK values computed between IEA and human scores. ΔQWK indicates the difference between QWK_{I_H} and QWK_{H_H} . SMD_{I_H} indicates the SMD values computed between IEA and human scores.

Table 2: Statistical Indicators by Training Approach Across Demographic Groups

Variable	Training	Group	QWK _{I_H}				Δ QWK				SMD _{I_H}			
			Min	Ave	Max	Flag	Min	Ave	Max	Flag	Min	Ave	Max	Flag
Gender	Feature	Female	0.76	0.83	0.89	0.00	-0.05	-0.02	0.05	0.00	-0.06	0.02	0.11	0.02
		Male	0.71	0.83	0.88	0.00	-0.08	-0.03	0.00	0.00	-0.04	0.01	0.12	0.03
	Transformer	Female	0.76	0.83	0.88	0.00	-0.05	0.00	0.06	0.00	-0.04	0.03	0.13	0.04
		Male	0.76	0.85	0.89	0.00	-0.06	-0.01	0.03	0.00	-0.03	0.02	0.11	0.02
ELL	Feature	Non-ELL	0.73	0.83	0.88	0.00	-0.07	-0.04	-0.01	0.00	-0.12	0.01	0.07	0.00
		ELL	0.70	0.76	0.82	0.00	-0.13	-0.05	0.04	0.29	0.00	0.03	0.11	0.07
	Transformer	Non-ELL	0.70	0.83	0.88	0.02	-0.07	-0.02	0.01	0.00	-0.03	0.05	0.23	0.07
		ELL	0.74	0.79	0.84	0.00	-0.10	-0.03	0.06	0.00	-0.07	-0.01	0.04	0.00
Race	Feature	Black	0.72	0.82	0.89	0.00	-0.06	-0.03	0.01	0.00	-0.06	0.02	0.07	0.00
		Hispanic	0.67	0.81	0.88	0.04	-0.09	-0.04	-0.01	0.00	-0.02	0.02	0.13	0.04
		White	0.73	0.83	0.89	0.00	-0.05	-0.03	0.00	0.00	-0.07	0.00	0.12	0.04
	Transformer	Black	0.73	0.83	0.88	0.00	-0.05	-0.01	0.03	0.00	-0.07	0.00	0.08	0.00
		Hispanic	0.75	0.84	0.88	0.00	-0.04	-0.02	0.02	0.00	-0.04	0.01	0.06	0.00
		White	0.70	0.83	0.88	0.00	-0.05	-0.01	0.03	0.00	-0.05	0.03	0.11	0.04
SWD	Feature	Non-SWD	0.70	0.83	0.89	0.00	-0.07	-0.02	0.02	0.00	-0.02	0.02	0.07	0.00
		SWD	0.72	0.83	0.88	0.00	-0.08	-0.03	0.03	0.00	-0.05	0.01	0.06	0.00
	Transformer	Non-SWD	0.72	0.83	0.88	0.00	-0.05	-0.01	0.06	0.00	-0.06	0.03	0.13	0.09
		SWD	0.70	0.84	0.88	0.03	-0.07	-0.01	0.03	0.00	-0.08	-0.01	0.04	0.00

Note. I_H indicates comparisons between IEA and human (i.e., best human as defined in this study). H_H indicates comparisons between human and human (i.e., two randomly selected humans from a pool of raters). QWK_{I_H} indicates the QWK values computed between IEA and human scores. Δ QWK indicates the difference between QWK_{I_H} and QWK_{H_H}. SMD_{I_H} indicates the SMD values computed between IEA and human scores.