

On Promoting Individual-Level Fairness in Automated Scoring

Michael Fauss and Matthew S. Johnson and Ikkyu Choi

ETS Research Institute, Princeton, NJ, USA

{mfauss, msjohnson, ichoi001}@ets.org

Abstract

The problem of promoting individual fairness in automated scoring is investigated. First, a recently proposed definition of individual-level fairness in terms of score distribution distances is revisited and adapted to the problem at hand. Based on this definition, three different automated scoring models are trained under individual fairness constraints. The accuracy, fairness and resulting population score distributions of these models are then evaluated on a subset of the PERSUADE 2.0 corpus, and compared to an unconstrained baseline model. Our findings show that explicitly enforcing individual fairness during model training significantly reduces fairness violations on the test set, but comes at the cost of a slightly reduced accuracy and a tendency to regress scores towards the mean.

1 Introduction

Concerns about fairness have long shaped educational measurement. Traditionally, the field has operationalized fairness at the group level, using tools such as differential item functioning (DIF), measurement invariance, and statistical criteria such as independence, separation, and sufficiency (Barocas et al., 2023; Johnson et al., 2022; Johnson and McCaffrey, 2023; Suk and Han, 2024). These methods seek to ensure that test scores do not systematically disadvantage members of protected social groups.

The adoption of automated scoring systems powered by artificial intelligence (AI) and machine learning has brought new urgency to fairness concerns. Modern scoring algorithms can capture nuanced response patterns, but can also introduce subtler forms of bias that elude group-based diagnostics. Moreover, as more demographic and contextual variables are considered along with their intersections (Suk and Han, 2025), the population segments into smaller and smaller subgroups, and ensuring fairness across all possible intersections eventually becomes unachievable.

Individual fairness, as proposed by Dwork et al. (2012), reframes the problem: rather than aggregating outcomes within broad groups, it requires that similar responses be treated similarly. For automated scoring, this means that similar responses, regardless of demographics or background, should receive similar scores. This echoes familiar psychometric notions such as local independence and response consistency, but formalizes them within an algorithmic decision-making context.

However, this appealingly intuitive principle raises challenging questions. How should “similarity” between responses be defined in a way that reflects rubric-relevant features while ignoring extraneous ones? What criteria can be used to detect, quantify, or guarantee individual fairness? While group-level fairness has a mature methodological toolkit, individual-level fairness in educational measurement remains nascent.

Johnson et al. (2026) proposed to bridge this gap by defining the similarity of two responses as the similarity of their respective *score distributions*. That is, two responses are similar if the probabilities of them receiving a certain score are similar. This definition will be made formal and discussed in more detail shortly.

While Johnson et al. (2026) focused on defining and assessing individual fairness in automated scoring, the paper at hand addresses the follow-up question of how to design automated scoring systems that are fair on an individual level, or at least promote individual-level fairness.

The paper is organized as follows: Individual fairness and its proposed specialization are introduced in Section 2. The optimization problem underlying the training of automated scorers that promote individual fairness is stated in Section 3. An experiment based on a subset of the PERSUADE 2.0 corpus (Crossley et al., 2024) is detailed and discussed in Section 4. Section 5 concludes the paper and discusses avenues for future work.

2 Individual Fairness

As mentioned in the introduction, the idea underlying individual fairness is that similar inputs should lead to similar outputs. In the context of automated scoring, this means that two responses that admit similar context-relevant features are assigned similar scores.

In order to make this notion formal, let $w_1, w_2 \in \mathcal{W}$ be two responses to a given item. For example, w_1 and w_2 could be two essays written in response to the same prompt, with $\mathcal{W}$ being the set of all strings shorter than a given maximum length. A scorer $S: \mathcal{W} \rightarrow \mathcal{K} = \{1, \dots, K\}$ is assumed to map the responses to an integer-valued score between 1 and K . A scorer that is individually fair in the sense of [Dwork et al. \(2012\)](#) satisfies

$$\delta(S(w_1), S(w_2)) \leq d(w_1, w_2), \quad (1)$$

where δ is a distance between scores and d is distance between responses. The question of how to choose suitable distances is addressed next.

2.1 Distance Between Scores

Defining a useful and meaningful score distance is a more nuanced problem than one might expect. In particular, a definition based on a straightforward score difference, say, $|S(w_1) - S(w_2)|$, is rather limiting in practice. For example, both human and machine scores are typically “noisy” in the sense that different raters can assign different scores to the same response. A closely related issue is that responses may not match the given rubric exactly, so that different scores can be equally justified.

[Johnson et al. \(2026\)](#) proposed to define the difference between scores as a difference between *score distributions*, that is,

$$\delta(S(w_1), S(w_2)) = D(P_{S|w_1}, P_{S|w_2}), \quad (2)$$

where D is a statistical distance and we use the notation $S|w$ to indicate that S is a random variable.

This definition overcomes many of the problems mentioned above. Moreover, the case of a single, deterministic score can be embedded in this framework by allowing degenerate distributions.

2.2 Distance Between Responses

One of the key ideas in [\(Johnson et al., 2026\)](#) is that, given a trusted scorer, any difference in the construct-relevant features of two responses will be reflected in the respective score distributions. Put

another way, differences between responses that have no impact on the score distribution are likely construct-irrelevant. Therefore, assuming that the scoring mechanism can be trusted, the score distribution distance is also a suitable response distance:

$$d(w_1, w_2) = D(P_{S|w_1}, P_{S|w_2}). \quad (3)$$

While this approach might appear circular, it has several advantages. First, explicitly identifying and isolating construct-relevant features can be highly challenging for complex or constructed responses. Defining response similarity as score distribution similarity forgoes this need. Instead, it relies on the trusted scorers, typically trained human experts, to implicitly identify the relevant features and score the responses accordingly. Second, the proposed approach provides a transparent and meaningful interpretation of response similarity. This is in contrast to, for example, definitions based on norms or angles of possibly high-dimensional feature vectors that are virtually impossible to interpret for humans. Finally, the proposed definition is universal in the sense that it can be applied across item and response types—the same definition of response similarity can be used for written, spoken or selected responses.

The main drawback of defining response similarity as score distribution similarity is that it requires a reference scorer that is trusted to be accurate and fair. In practice, this commonly means that a training set of responses is scored by multiple human raters, resulting in an empirical score distribution. Letting the number of raters go to infinity, this empirical score distribution is assumed to converge to the true score distribution. On the other end of the spectrum, if only a single rater is available to score the training data, one can either fall back to deterministic scores or obtain approximate score distributions by smoothing the degenerate ones.

There can also be scenarios where the aim is not to learn from human raters, but to use a trusted yet complex, possibly slow and expensive automated scorer to train a simpler one. For example, an ensemble of LLMs could be used to generate scores for the responses in the training set, which are then used to fit a simpler model.

3 Optimization Problem

Irrespective of the nature of the reference scorer, the underlying optimization problem remains the same: Given a set of training responses with trusted

score distributions, design an automated system that replicates these distributions as closely as possible while satisfying individual fairness constraints.

Let $\mathcal{W}_N = \{w_1, \dots, w_N\}$ be a set of training responses. We assume that for each response $w_n \in \mathcal{W}_N$ a score distribution, $\mathbf{p}_n$, was generated by a trusted scorer. The aim is to use this training data to design an automated scorer that is individually fair in the sense of Dwork et al. (2012). Following the approach proposed in that paper, the design process involves two steps. First, for each training response, w_n , a corresponding optimal score distribution, $\mathbf{q}_n^*$, is obtained by solving a constrained optimization problem. Then, the obtained score distributions are used to fit a model that extrapolates these distributions to unseen responses—compare (Dwork et al., 2012, Remark 2.1).

3.1 Solution on the Training Set

For the sake of simplicity and because of its widespread use in practice, we use the mean squared error (MSE) as a cost function. That is, our goal is to minimize

$$\mathbb{E}[(S|W - M|W)^2], \quad (4)$$

where W is a randomly chosen response, S denotes the reference scorer and M denotes the scorer we seek to train. Note that the expected value in (4) is taken over the random response, W , as well as over the score distributions generated by S and M . In practice, the expectation over W is unknown, but can be approximated by the sample average over the training data, that is,

$$\frac{1}{N} \sum_{n=1}^N \mathbb{E}[(S|w_n - M|w_n)^2]. \quad (5)$$

Now, let $\mathbf{p}_n \in [0, 1]^K$ denote the probability mass function (PMF) of $S|w_n$, that is, $p_{n,k} = P[S = k | W = w_n]$. Analogously, let $\mathbf{q}_n \in [0, 1]^K$ denote the PMF of $M|w_n$. Using this notation, we can write the inner expectation in (5) as

$$\sum_{k=1}^K \sum_{j=1}^K (k-j)^2 p_{n,k} q_{n,j} = \mathbf{p}_n^\top \mathbf{C} \mathbf{q}_n, \quad (6)$$

where the elements of the matrix $\mathbf{C} \in \mathbb{R}^{K \times K}$ are the quadratic costs associated with two scores, that is, $C_{k,j} = (k-j)^2$. Note that in (6) we assume the scores S and M to be independent conditioned on a given response, w_n . We believe that this assumption is typically satisfied in practice. If there are

reasons to assume that it is not, the optimization problem can be formulated in terms of the joint distributions of $(S, M)|w_n$, with the added constraint that the marginal distribution of $S|w_n$ equals $\mathbf{p}_n$. However, we consider this extension beyond the scope of this paper.

Using the notation in (6), the optimization problem can be written as

$$\min_{\mathbf{q}_1, \dots, \mathbf{q}_N} \frac{1}{N} \sum_{n=1}^N \mathbf{p}_n^\top \mathbf{C} \mathbf{q}_n \quad \text{s.t.} \quad (7)$$

$$D(\mathbf{q}_n, \mathbf{q}_m) \leq D(\mathbf{p}_n, \mathbf{p}_m) \quad \forall n, m \leq N, \quad (8)$$

with the constraint in (8) enforcing individual fairness on the training data set.

3.2 Extrapolation of the Solution

Solving (7) results in a set of conditional score distributions, $\mathbf{q}_1^*, \dots, \mathbf{q}_N^*$, one for each training sample. However, in order to score unseen responses, a general scoring rule needs to be “extrapolated” from the score distributions of the training samples.

In practice, this step is a traditional supervised classification problem, with the difference that the model is not trained to predict *scores*, but *score distributions*. This scenario is often referred to as “learning from soft labels” and has been studied extensively in the literature—see, for example, (de Vries and Thierens, 2025).

4 Experiment

In this section, we perform an experiment to explore the feasibility and potential impact of training an automated scorer under individual fairness constraints. For the sake of reproducibility, we use the freely available PERSUADE 2.0 corpus (Crossley et al., 2024) of argumentative essays. More specifically, we used the essays written in response to the “Distance Learning” prompt.

Our aim is to simulate a scenario in which a reliable reference scorer is given, and a simpler machine scorer is trained on its output. Since the focus of the experiments is on the impact of the individual fairness constraints, we tried to keep the technical difficulties and intricacies to a minimum. Specifically, we assume that all scorers in this experiment take essay embedding vectors as an input, and use ordinal logistic regression (Agresti, 2013; Pedregosa-Izquierdo, 2015) to map them to score distributions. This scorer design is simple, yet not too different from current-generation automated

scoring systems that often use complicated features in combination with relatively simple classifiers or regression models (Attali and Burstein, 2004). Moreover, ordinal logistic regression models can be trained on soft labels in a straightforward manner (Quost, 2014). Python code to replicate our results is available at (Fauss et al., 2026).

4.1 Setup

In order to obtain a reference scorer, we used the Amazon Titan Text Embeddings V2 model (AWS, 2023) to generate embeddings of size 1024 for all $N = 1498$ “distance learning” essays in the training set. We used the holistic essay scores included in the PERSUADE 2.0 corpus, which are on a scale from 1 to $K = 6$, to fit an ordinal logistic regression model. This model was then used to generate reference score distributions, $\mathbf{p}_1, \dots, \mathbf{p}_N$, for all essays in the training set.

Given the reference score distributions, the optimization problem in (7) was solved for three different statistical distance metrics, namely,

1. Jeffreys divergence (Jeffreys, 1948),

$$D_J(\mathbf{p}, \mathbf{q}) = D_{KL}(\mathbf{p} \parallel \mathbf{q}) + D_{KL}(\mathbf{q} \parallel \mathbf{p}); \quad (9)$$

2. Jensen–Shannon divergence (Lin, 1991),

$$D_{JS}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} (D_{KL}(\mathbf{p} \parallel \mathbf{m}) + D_{KL}(\mathbf{q} \parallel \mathbf{m})), \quad (10)$$

where $\mathbf{m} = \frac{1}{2}(\mathbf{p} + \mathbf{q})$; and

3. total variation distance (Gibbs and Su, 2002),

$$D_{TV}(\mathbf{p}, \mathbf{q}) = \frac{1}{2} \|\mathbf{p} - \mathbf{q}\|_1. \quad (11)$$

In 1. and 2., D_{KL} denotes the standard Kullback–Leibler (KL) divergence. The use of Jeffreys divergence was proposed by Johnson et al. (2026), based on the fact that it is a natural symmetrization of the KL divergence. However, it has the well-known drawback of being infinite if $\mathbf{p}$ and $\mathbf{q}$ do not share the same support. The Jensen–Shannon divergence overcomes this problem by using a mixture distribution for reference. Finally, use of the total variation distance was proposed in (Dwork et al., 2012) and has the advantage that it is linear in both arguments, thus reducing the numerical complexity of the optimization problem in (7) compared to the exponential/logarithmic distances in 1. and 2. Note that, while we limit our study to these three, there

are various other candidate distances that might warrant closer inspection in the future.

The outcome of the three optimization problems corresponding to the statistical distances are three sets of optimal “individually fair” score distributions, $\mathbf{q}_{J,n}^*$, $\mathbf{q}_{JS,n}^*$ and $\mathbf{q}_{TV,n}^*$, where $n = 1, \dots, N$. These score distributions are in turn used to fit three new ordinal logistic regression models, representing three simpler automated scoring systems that are supposed to approximate the reference scorer. In order to simulate the complexity difference, we used embeddings of size 512 to fit the three individually fair models.

Finally, we fitted a 512-dimensional ordinal regression model without any individual fairness constraints. The corresponding score distributions are denoted by $\mathbf{q}_{UC,n}^*$, $n = 1, \dots, N$. This model serves as a baseline against which the impact of the individual fairness constraints is compared.

4.2 Results

In this section, we explore both the accuracy and the fairness of the trained models. To this end, all five models detailed above (reference model + three constrained models + unconstrained model) were used to generate score distributions for the 659 “distance learning” essays in the test set of the PERSUADE 2.0 corpus. Throughout our experiments, the score distributions generated by the reference model are assumed to constitute the “ground truth,” in the sense that, ideally, the smaller models should recover these distributions exactly.

4.2.1 Accuracy

In order to gauge how well the four simpler models approximate the reference model, we evaluated the mean squared error (MSE) and Cohen’s quadratically weighted kappa (QWK) on the test set. The results are summarized in Table 1. The unconstrained model achieves the highest accuracy according to both metrics. This is expected, as additional constraints necessarily come at the expense of “raw accuracy.” Interestingly, the accuracy ranking of the fairness constrained scorers depends on the metric. In terms of MSE, the Jeffreys divergence performs best and the total variation distance worst; in terms of QWK, the ranking is reversed.

4.2.2 Fairness

We evaluate the fairness of the different models by inspection of the degree to which they satisfy the individual fairness constraints on the test data set.

Constraint Type	MSE	QWK
Unconstrained	0.555	0.793
Jeffreys	0.612	0.738
Jensen–Shannon	0.617	0.749
Total Variation	0.621	0.756

Table 1: Accuracy of machine scorers trained with different types of individual fairness constraints.

We define the pairwise constraint violation as

$$\Delta_{\bullet}(n, m) = D(\mathbf{q}_{\bullet,n}^*, \mathbf{q}_{\bullet,m}^*) - D_{\bullet}(\mathbf{p}_n, \mathbf{p}_m), \quad (12)$$

where n and m index the responses, $\mathbf{p}_n$ denotes the reference score distribution for response n , and $\mathbf{q}_{\bullet,n}^*$ denotes the corresponding score distribution assigned by the trained model of interest. Here, $\bullet$ is a placeholder for the different models, namely, the three models fitted under individual fairness constraints and the unconstrained model.

Since our test set contains $N = 659$ responses, and all distances/divergences used in the paper are symmetric, we obtain $\frac{N(N-1)}{2} = 216\,811$ unique response pairs. Histograms of the corresponding constraint violations, as defined in (12), are shown in Figure 1. In all six plots, positive values indicate a constraint violation, negative values indicate feasibility. In Figure 1, the columns correspond to different distances used in the definition of individual fairness, while the rows distinguish between constrained and unconstrained models. More specifically, the results in the first row were obtained from models trained explicitly under the corresponding individual fairness constraints, while the results in the second row were generated using the unconstrained baseline model.

By inspection, the pairwise constraint violations of the fairness-constrained models all follow approximately the same pattern: the distribution has a single mode at zero, and two approximately symmetric, fast decaying tails. This shape might be surprising at first glance since it implies that approximately half of the unique response pairs violate the individual fairness constraint. However, we would argue that this is expected and in line with the conceptual approach. Qualitatively speaking, in order for the individual fairness constraints to have an impact on the trained model, many of them should bind, that is, the optimal solution should satisfy them with equality. However, this equality is only exact on the training data set. On the test

data set, the feature distributions will slightly differ. This difference manifests as noise in the score distributions. This effect can be observed in the first row of Figure 1, where the constraints are no longer satisfied with exact equality. However, the deviations are concentrated around zero and there is no obvious bias in either direction. This indicates that the observed residuals are indeed caused by random noise, and not by systematic model mismatch.

Systematic model mismatch can, however, be clearly observed in the outputs of the unconstrained model in the second row. In all three cases, the histogram is significantly skewed to the right, meaning that constraint violations are more likely to occur than not. Moreover, the right tails are significantly heavier than their counterparts in the first row, that is, violations of individual fairness are not only more likely, but also more severe. Note that the spike at the end of the right tail in the plot on the lower left contains many cases in which the constraint violation was infinite. This is a consequence of using Jeffreys divergence, which can become infinite for distributions with non-identical support.

The picture painted by the constraint violation histograms is supported by their summary statistics. In Table 2, the share of constraint violations and their average values are given for all models. In the rightmost column, the number in parenthesis is the conditional average over the right tail of the histogram, that is, over all cases that violated the fairness constraint. It can be seen that the constrained models violate fairness constraints in less than 50 % of the cases, while the share is between 76 % and 86 % for the unconstrained model. Analogously, the expected constraint violation is slightly negative for the constrained models, and clearly positive for the unconstrained one. Looking at the conditional averages, it can be seen that the severity of the violation increases as well. The average violation approximately doubles for both bounded distances, and “explodes” for the unbounded Jeffreys divergence.

4.2.3 Impact on Population Score Distribution

In addition to their impact on the overall accuracy, it is instructive to study the effect of individual fairness constraints on the population score distribution. In particular, fairness promoting constraints can have the effect of “compressing” the population score distribution, by regressing the scores towards the mean. In turn, extreme scores are assigned less frequently—compare (Fauss and Choi, 2025).

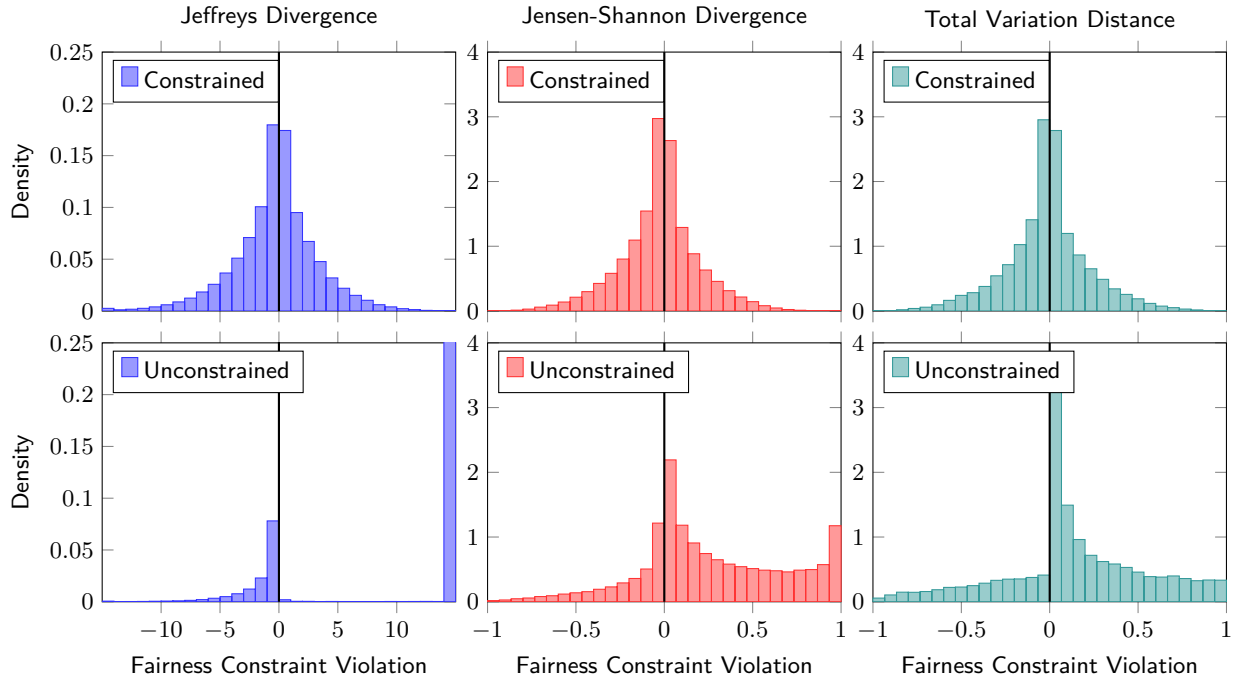


Figure 1: Histograms of pairwise fairness violations of different models.

Constraint Type	Share of Constraint Violations		Average (Positive) Constraint Violation	
	Constrained	Unconstrained	Constrained	Unconstrained
Jeffreys	47.8 %	86.3 %	−0.246 (2.351)	∞ (∞)
Jensen–Shannon	44.8 %	76.2 %	−0.017 (0.104)	0.185 (0.290)
Total Variation	47.0 %	76.0 %	−0.013 (0.162)	0.131 (0.292)

Table 2: Share and average of constraint violations for different models.

This compression can indeed be observed in our experiments. In Figure 2, the difference between the mean score of the respective fair model and the mean score of the reference model is plotted against the latter. Note that for positive values the mean score of the fair model is higher than that of the reference model, that is, it tends to overestimate the score, while for negative values it is lower, that is, it tends to underestimate the score. By inspection of Figure 2, all three fair models tend to overestimate the scores of low-scoring essays, and underestimate the scores of high-scoring essays. This effect is most pronounced for the Jeffreys divergence. In fact, the maximum probability of receiving a score of 1 was 7.37×10^{-4} over the test set. That is, the model effectively clips the score scale at 2. The compression effect is less extreme for the Jensen–Shannon divergence and the total variation distance, but can still be seen clearly in the data.

Finally, the variances of the score distributions are plotted against the respective mean scores in Figure 3. The graphs of all four models show the same pattern, with troughs around the integers and peaks in between. Although, in principle, this pattern is expected as non-integer mean scores imply that the model is uncertain how the response should be scored, the smoothness of the graphs is remarkable, and likely a consequence of the underlying ordinal logistic regression.

It can be seen in Figure 3 that the models with the largest compression also show the largest reduction in variance. In particular, the model based on the Jeffreys divergence admits the lowest variance uniformly across all mean scores. Among the fairness constrained models, the variance of the one based on the total variation distance resembles that of the reference model most closely. Interestingly, for mean scores below approximately 2.5, both the total variation and the Jensen–Shannon based model show higher variances than the reference model. We conjecture that this is a side effect of these models’ tendency to overestimate the quality of low-scoring responses, so that the resulting score distributions are less concentrated on the lowest score than that of the reference model.

5 Discussion and Outlook

The main takeaways from this paper are:

1. Score distribution distances are a potentially useful tool to operationalize individual fairness in the sense of [Dwork et al. \(2012\)](#).

2. Simply training a model on human scores or scores generated by a trusted reference scorer does not guarantee individual fairness. That is, the trained model might give different scores to responses that human raters or the reference system would have scored identically.

3. It is possible to promote, albeit not guarantee, individual fairness by explicitly enforcing it during training.

4. The cost of promoting individual fairness is, first, a slight reduction of overall accuracy, and second, a more concentrated population score distribution. Put differently, the fair models tend to give more similar scores at the expense of their discriminatory power. In particular, they might avoid assigning extreme scores, even when they are justified.

5. The distance metric used in the individual fairness constraints matters. In the example considered here, all three distance metrics showed particular characteristics and resulted in different population score distributions.

6. Based on our limited experiments, the total variation distance appears to be the most useful in practice: it leads to a linearly constrained optimization problem, achieved the highest QWK, and showed the least overall score compression.

Open questions and avenues for future work include the following:

1. The results presented in this paper are based on a single prompt and a single model (ordinal logistic regression). A more comprehensive study is required to investigate if and under which conditions and assumptions the findings generalize beyond this setting.

2. The number of constraints in (8) grows quadratically with the number of response. For training sets on the order of 10^3 , as in our experiment, this did not cause major problems. However, for larger training sets, compute and/or memory constraints are likely to become a problem.

3. Can the problem of score compression caused by individual fairness constraints be alleviated? For example, a constraint that matches the population score distributions of the trained and reference model on the training set could be added in (7). Alternatively, the concentration of the population score distribution could be reduced by adding terms that penalize low-entropy or low-variance solutions.

4. Instead of evaluating the performance of different distance metrics numerically, it could be useful to approach the problem of picking a suitable distance axiomatically, that is, by establishing crite-

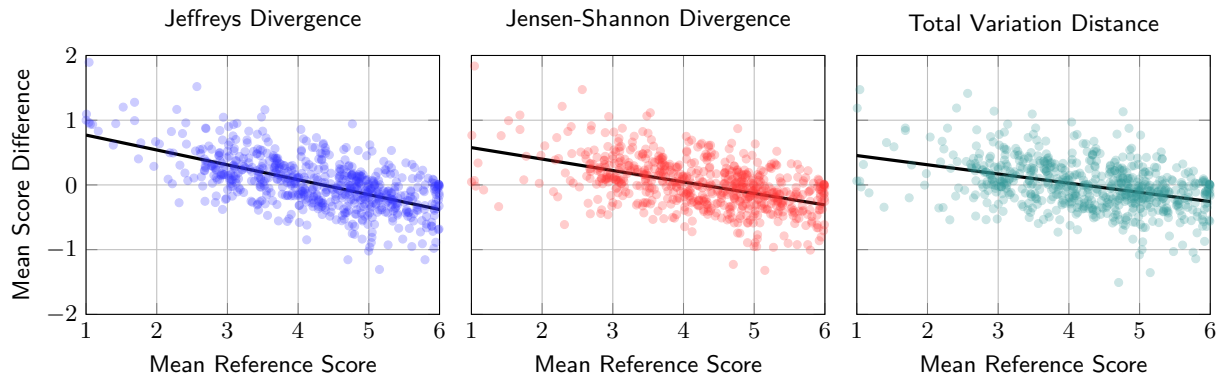


Figure 2: Scatter plot of mean score difference between fitted and reference model against mean score of reference model.

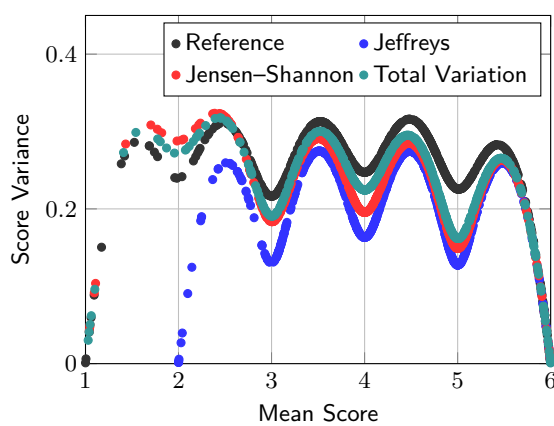


Figure 3: Scatter plot of score variance against mean score of respective model.

ria that it should meet in the context of automated scoring. This could narrow the set of candidate distances significantly.

5. The interplay between group and individual fairness should be explored in more detail. Does promoting individual fairness also promote group fairness and vice versa? Or can the two be in conflict? To the best of our knowledge, these questions have not yet been investigated systematically.

References

- A. Agresti. 2013. *Categorical Data Analysis*, third edition. Wiley, Hoboken, NJ, USA.
- Y. Attali and J. Burstein. 2004. *Automated essay scoring with e-rater v.2.0. ETS Research Report Series*, 2004.
- AWS. 2023. *Amazon titan embeddings*. Accessed: 6/18/2026.
- S. Barocas, M. Hardt, and A. Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, Cambridge, MA, USA.
- S. A. Crossley, P. Baffour, Y. Tian, A. Franklin, M. Benner, and U. Boser. 2024. *A large-scale corpus for assessing written argumentation: PERSUADE 2.0. Assessing Writing*, 61.
- S. de Vries and D. Thierens. 2025. *Learning with confidence: training better classifiers from soft labels. Machine Learning*, 114(11):238.
- C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. *Fairness through awareness*. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, pages 214–226, New York, NY, USA. ACM.
- M. Fauss and I. Choi. 2025. *A fairness-promoting detection objective with applications in AI-assisted test security*. In *Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con): Coordinated Session Papers*, pages 107–114, Pittsburgh, PA, USA. NCME.
- M. Fauss, M. S. Johnson, and I. Choi. 2026. *Code repository*. https://github.com/mifauss/Individual_Fairness.
- A. L. Gibbs and F. E. Su. 2002. *On choosing and bounding probability metrics. International Statistical Review*, 70(3):419–435.
- H. Jeffreys. 1948. *Theory of Probability*, second edition. Oxford University Press, Oxford, UK.
- M. S. Johnson, M. Fauss, and I. Choi. 2026. *An approach for assessing individual-level fairness in automated scoring*. In *Annual Meeting of the National Council on Measurement in Education (NCME)*, Los Angeles, CA, USA. To appear.
- M. S. Johnson, X. Liu, and D. F. McCaffrey. 2022. *Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. Journal of Educational Measurement*, 59(3):338–361.
- M. S. Johnson and D. F. McCaffrey. 2023. *Evaluating fairness of automated scoring in educational measurement*. In *Advancing Natural Language Processing in Educational Assessment*, pages 142–164. Routledge.

- J. Lin. 1991. [Divergence measures based on the Shannon entropy](#). *IEEE Transactions on Information Theory*, 37(1):145–151.
- F. Pedregosa-Izquierdo. 2015. [Feature extraction and supervised learning on fMRI : from practice to theory](#). Theses, Université Pierre et Marie Curie - Paris VI.
- B. Quost. 2014. [Logistic regression of soft labeled instances via the evidential EM algorithm](#). In *Belief Functions: Theory and Applications*, pages 77–86, Cham. Springer International Publishing.
- Y. Suk and K. T. Han. 2024. [A psychometric framework for evaluating fairness in algorithmic decision making: Differential algorithmic functioning](#). *Journal of Educational and Behavioral Statistics*, 49(2):151–172.
- Y. Suk and K. T. Han. 2025. [Evaluating intersectional fairness in algorithmic decision making using intersectional differential algorithmic functioning](#). *Journal of Educational and Behavioral Statistics*, 50(5):833–862.