# Identifying Biases in Large Language Model Assessment of Linguistically Diverse Texts

**Lionel Meng**
**Shamya Karumbaiah, Ph.D.**
**Vivek Saravanan**
**Daniel Bolt, Ph.D.**
University of Wisconsin - Madison / 1025 W Johnson St, Madison, WI 53706
**Correspondence: lhmeng@wisc.edu, shamya.karumbaiah@wisc.edu**

## Abstract

The development of Large Language Models (LLMs) to assess student text responses is rapidly progressing but evaluating whether LLMs equitably assess multilingual learner responses is an important precursor to adoption. Our study provides an example procedure for identifying and quantifying bias in LLM assessment of student essay responses.

## 1 Introduction

The application of Large Language Models (LLMs) for assessing student essays affords numerous avenues of research within learning analytics. Particularly for high stakes assessment contexts where annotated data is often sparse or difficult to acquire, the use of LLMs becomes particularly attractive. However, for LLMs to be ethically applied to educational assessment, they must be evaluated for equity across diverse student subpopulations. One subpopulation of particular concern is multilingual students. In high stakes testing contexts, acquiring sufficient annotated data for multilingual students is often unrealistic for reasons such as test security, student privacy, diversity in linguistic practices, and low population size. Furthermore, traditional methods of algorithmic bias assessment that rely on broad demographic categories such as age or gender are prone to mis-characterize the complex heterogeneous backgrounds of such students, potentially making them ineffective. Direct empirical comparisons across subpopulations can also be complicated by difficulties in separating bias from impact (Ackerman, 1992). Understanding causes of differential item functioning is a notorious challenge in the use of empirical data for evaluating bias (Zumbo, 2007).

## 2 Aim

Our study aims to illustrate a procedure by which LLM performance can be assessed for equity by systematically manipulating texts with construct-irrelevant linguistic variations and characterizing resultant score change. We refer to these variations as perturbations and the resulting scores as perturbed scores.

## 3 Sample(s)

Texts are from the Hewlett Foundation: Automated Essay Scoring competition data (Hammer et al., 2012). The sample consists of 5875 actual essay responses written by students in grades 7 through 10 in response to prompts that did not have accompanying reading passages (essay numbers 1, 2, 7, 8). Essays range from 150 to 550 words in length.[1]

## 4 Methods

Analyses of texts begins by feeding original, monolingual texts to the target LLM, GPT-4o, for scoring.

### 4.1 LLM Prompting

We employed GPT-4o as the LLM for essay grading. The model was prompted with "Grade the essay below with a score between 0 and 100 based on content, ignoring language errors. Your response must be exactly one number between 0 and 100". Scores were normalized to range from 0 to 1.

### 4.2 Text Quality and Baseline Reference Values

Each monolingual text was scored twice by the LLM. The first of each of these scores was chosen to be a reference value for subsequent analyses, and is henceforth referred to as the "original score." We refer to the second score as the "replicate score."

Texts were then grouped into quartiles based on original scores. Resultant quartile sizes, from first to fourth, were as follows: 1767, 2600, 619,

[1]This dataset is openly accessible at [https://www.kaggle.com/competitions/asap-aes/overview].



*Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress*, pages 204–210

889. For each quartile, the first three statistical moments of the difference between original and perturbed scores were calculated. The inclusion of higher order moments reflects the notion that equity in measurement transcends expected score differences, and includes equivalence in precision as well as potential for outlier scores, etc.

Corresponding moments were also calculated for the difference between original and replicate scores to serve as a baseline. Specifically, an expected signed difference between the original score and replicate score (first order), an expected squared deviation between the original score and replicate score (second order), and an expected signed cubed deviation between the original and replicate score were calculated (third order).

### 4.3 Construct-Irrelevant Linguistic Variations

Construct-irrelevant linguistic variations here are defined as linguistic features of the text that are not directly related to the content proficiency intended to be measured. For a student essay response on a science test question, for example, it might refer to spelling errors that the student makes. The idea is that the underlying student response may be scientifically accurate despite the linguistic variation.

While it is true that large proportions of linguistic variations may impede LLM scoring of the text, not unlike barriers to comprehension that may occur with a human scorer, the label of "construct-irrelevant" is used to highlight that these are not the intended target construct of measurement. Indeed, where the LLM scoring becomes difficult due to perturbations, this difficulty itself becomes a form of inequity (Prabhakaran et al., 2019).

The four linguistic variations analyzed in this study were: 1) spelling errors, 2) noun transfers (i.e., borrowing nouns across languages; e.g., “Tierra” instead of “Earth”), 3) cognates (i.e., borrowing words with similar meaning, spelling, and pronunciation; e.g., “océano” instead of “ocean”), and 4) Spanglish (a hybrid use of both languages; e.g., “en la Earth” instead of “on the Earth”).

We build an algorithm in which eligible words or phrases at which the above linguistic variations could occur is first determined for each text. Then, the linguistic variations above are randomly introduced to each of the texts at these words or phrases, resulting in transformed, or perturbed versions of the text with the same underlying response meaning. Under this scheme, the magnitude of the perturbation can be controlled. For this study, we introduce perturbations of the following magnitudes: 20, 40, 60, and 80 percent.

### 4.4 Assessment of Inequity

The perturbed texts are scored by GPT-4o, such that each text has not only an original score, but a perturbed score as well. Using these values, we apply procedures conceptually derived from Lord’s (1980) notions of equity.

#### 4.4.1 First Order Inequity

We refer to the signed difference between the original score and perturbed score as "error" for each text. By calculating the expected error (original - perturbed) across texts for each quartile, we can determine a quartile-specific bias value attributable to the linguistic perturbations. Expected perturbation error values greater than the expected replicate error values suggest first order inequity.

#### 4.4.2 Second Order Inequity

By calculating the expectation of squared deviations between the original and perturbed score across essays within each quartile, we can get quartile-specific variances of deviations. We take the square root of these values to get standard deviations, and compare to the corresponding standard deviation for replicate scores as a reference. Standard deviation values that surpass the reference values suggest second-order inequity.

#### 4.4.3 Third Order Inequity

We also calculate the expectation of signed cubed deviations between the original and perturbed score across essays for each quartile. Values greater than quartile-specific third order baseline reference values defined using replicate scores suggest third order inequity.

## 5 Results

A Wilcoxon test (see Table 2) was conducted to compare original scores and perturbed scores, confirming that differences in scores seen due to linguistic perturbations of varying magnitudes were statistically significant. Our sample size of 5875 texts naturally predisposes the test to be significant, even with small average deviations. While practical significance of such LLM audits are context-specific and best determined on a case-by-case basis, for this study readers are referred to Figure 1, Figure 2, and Figure 3 where the effect of linguistic perturbations are quantified on the scale of the text

| perturbation & magnitude | Wilcoxon test statistic | p value |
|---|---|---|
| spanglish 20 | 3533900 | p < 0.05 |
| spanglish 40 | 3581410 | p < 0.05 |
| spanglish 60 | 3481962 | p < 0.05 |
| spanglish 80 | 3727746 | p < 0.05 |
| cognates 20 | 4146682 | p < 0.05 |
| cognates 40 | 4147836 | p < 0.05 |
| cognates 60 | 3969052 | p < 0.05 |
| cognates 80 | 3847161 | p < 0.05 |
| noun transfer 20 | 3944342 | p < 0.05 |
| noun transfer 40 | 3794908 | p < 0.05 |
| noun transfer 60 | 3524270 | p < 0.05 |
| noun transfer 80 | 3498297 | p < 0.05 |
| spelling 20 | 2669144 | p < 0.05 |
| spelling 40 | 1732722 | p < 0.05 |
| spelling 60 | 1436490 | p < 0.05 |
| spelling 80 | 1132988 | p < 0.05 |

Table 1: Wilcoxon test results comparing original and perturbed scores by linguistic variation and magnitude.

scores (0 to 1), and as such may serve as effect size measures.

Figure 1 displays quartile-specific results for bias in each linguistic perturbation, faceted by magnitude of perturbation. A general trend whereby increasing magnitudes of perturbation result in greater mean error can be observed. Additionally, for all linguistic variations, mean error values trend positive as successive quartile results are compared for all magnitudes of perturbation. Mean error values for all perturbations exceed the first order baseline reference level in the positive direction, although error values are still negative in the first quartile for all perturbations aside from spelling errors.

Figure 2 displays quartile-specific results for the expected value of squared deviations between original scores and perturbed scores, converted to standard deviations for each linguistic perturbation, faceted by magnitude of perturbation. Quartile-specific baseline reference values are represented as horizontal lines. With the exception of noun transfer at 60% magnitude for the first quartile, all standard deviation values were greater than baseline reference values within their respective quartiles. A moderate trend can be observed such that for texts in the first quartile, standard deviations tend to be high across linguistic variations.

Figure 3 displays quartile-specific results for the expected value of cubed deviations between original scores and perturbed scores for each linguistic perturbation, faceted by magnitude of perturbation. Skewness values exceed baseline reference values in the positive direction for all linguistic perturbations, although skewness values are still negative in the first quartile for all perturbations aside from spelling errors. Additionally, cognate skewness values only barely surpass baseline reference values in the second quartile.

## 6 Discussion

For texts in the second to fourth quartile of text quality, sensitivity of GPT-4o scores to linguistic perturbations of varying magnitudes in all analyses suggests inequitable assessment of student knowledge in its application. First order results indicate the presence of bias in LLM scoring, second order results further indicate differences in precision, and third order results indicate the an increased likelihood for extreme cases of discrepant results in the positive direction. Additional interpretation for results of texts in the first quartile are presented below.

First order analysis results for the first quartile in Figure 1 show mean error values surpassing the baseline reference value in the positive direction while remaining negative. In interpreting these results, however, it should be appreciated that the negative value associated with the reference condition likely represents a "regression to the mean" phe-

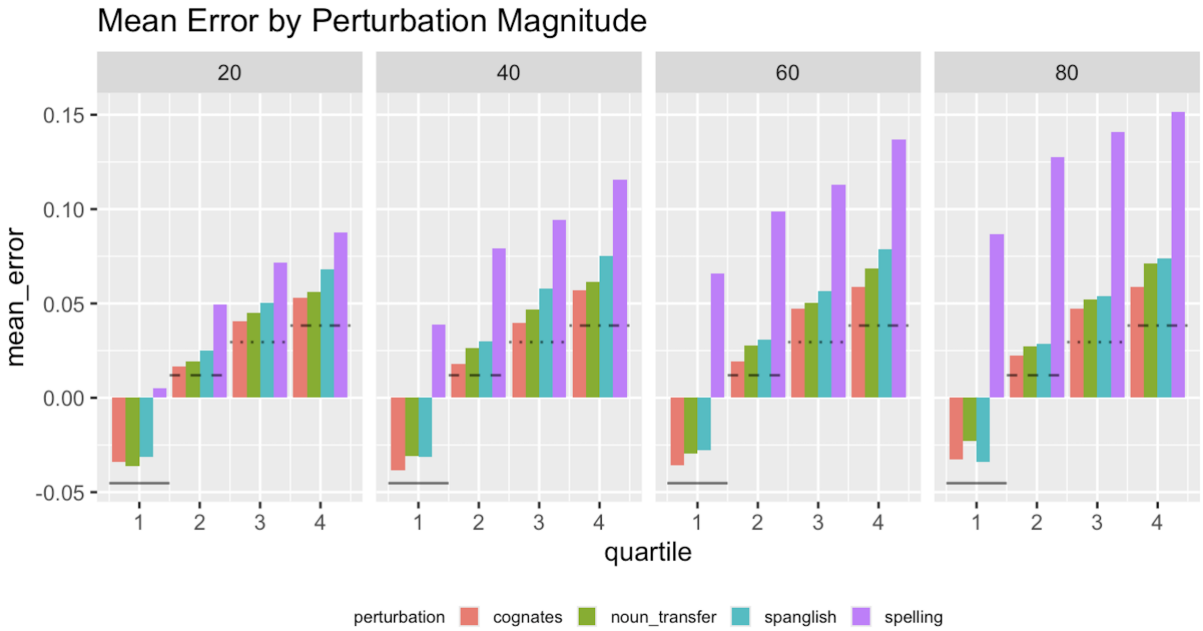


Figure 1: Error (original score - perturbed score) averaged across texts for each linguistic variation by original score quartile. Graphs are faceted by perturbation magnitude. Horizontal reference lines for signed difference between original and replicate score are included for each quartile.

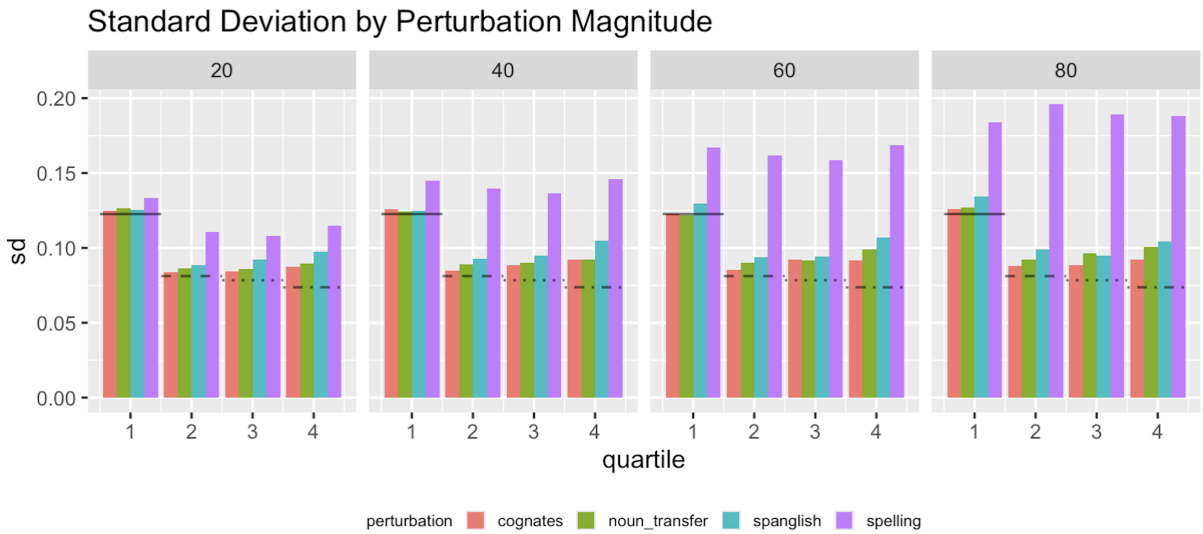


Figure 2: Standard deviation of differences averaged across texts for each linguistic variation by original score quartile. Graphs are faceted by perturbation magnitude. Horizontal reference lines for square root of expected squared deviations between original and replicate scores are included for each quartile.

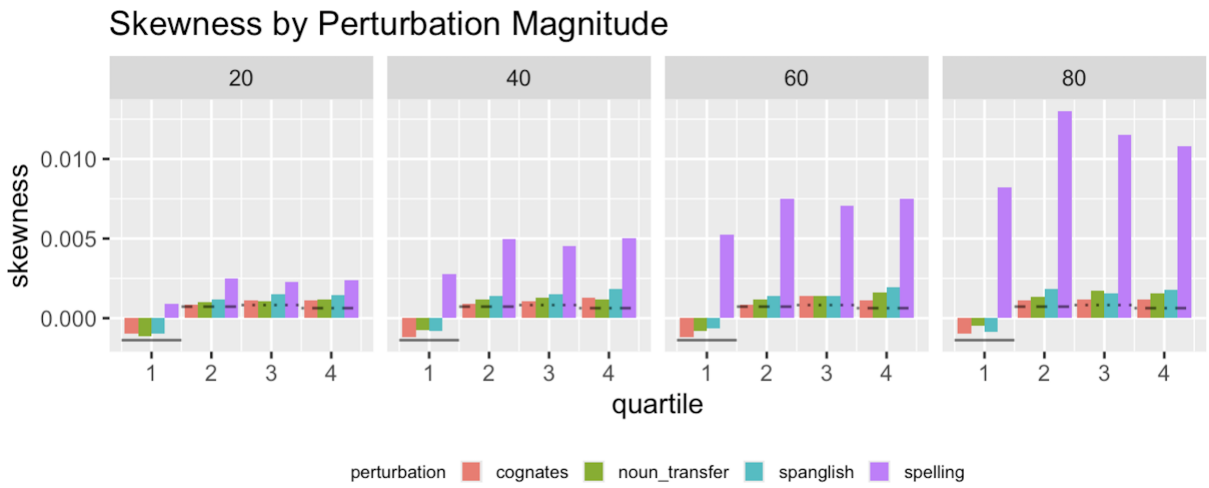


Figure 3: Skewness averaged across texts for each linguistic variation by original score quartile. Graphs are faceted by perturbation magnitude. Horizontal reference lines for expected cubed deviations between original and replicate scores are included for each quartile.

Cognate Perturbation 40% Magnitude Sample Texts with Large Error

| Text ID | Original Score | Perturbed Score | Error |
|---|---|---|---|
| 19047 | 0.15 | 0.60 | -0.45 |
| 240 | 0.25 | 0.65 | -0.40 |
| 18088 | 0.10 | 0.50 | -0.40 |
| 21140 | 0.10 | 0.50 | -0.40 |
| 1245 | 0.20 | 0.60 | -0.40 |
| 19190 | 0.45 | 0.15 | 0.30 |
| 19479 | 0.50 | 0.20 | 0.30 |
| 276 | 0.45 | 0.10 | 0.35 |
| 19320 | 0.45 | 0.10 | 0.35 |
| 18578 | 0.45 | 0.00 | 0.45 |

Table 2: Sample texts with large magnitudes of error for cognate perturbations at 40% magnitude in the 1st quartile. Scores are on a scale from 0 to 1. Error is calculated as original score - perturbed score. Results show potential for error scores in either direction.

nomenon, an expected statistical result, as scoring error is on average negative in the lowest quartile. Thus the larger values observed under perturbation, although still often negative, can nevertheless be viewed as a first order equity violation (albeit generally small), in that less than the expected regression correction is observed under perturbation. Additionally, it is important to keep in mind that mean error scores are muted due to cancellation from the signed nature of the quantity; paired with large variance of scores (see Figure 2), this leaves a non-trivial likelihood of inequities in scoring for particular students. In other words, for texts produced by students of developing proficiency, the target LLM would be expected to grade multilingual student text responses with lower levels of precision than monolingual student text responses. As an example, Table 2 shows the 5 most negative error value texts and 5 most positive error value texts for cognate perturbations at 40% magnitude in the first quartile. Comparably large magnitudes of error in either direction illustrate how noisy assessment of perturbed texts as shown by second order analysis results from Figure 2 can manifest. Given that academic decisions for students occur in consideration of individual scores, not group-aggregated values such as mean scores, this is potential reason for concern. Third order analysis results (see Figure 3) for the first quartile are similar to first order results in that expected cubed deviations are on average negative and regression to the mean corrections are more weakly observed under perturbation, indicating mild third order equity violations. We conclude that the effect of linguistic perturbation results in violations of equity in all three orders, with particularly strong results for spelling errors.

## 7 Limitation and Future Directions

Our study intended to highlight a methodology for examining the effects of perturbations on LLM scoring. A primary limitation of our results relates to the constantly changing nature of LLMs. It is likely that the validity of audit results for any given LLM will have limited longevity. As such, stakeholders are advised to audit their target LLMs as close to the time of application as possible.

As the authenticity of the algorithmically-introduced linguistic variations can be questioned, results from this audit procedure should be interpreted cautiously. More developed ways to introduce these perturbations can be implemented in the future to improve the validity of the audit procedure.

One challenging aspect of this procedure lies in determining comparable magnitudes of different perturbations. In this study, we opted to use the number of eligible words in the text for a given linguistic variation. For spelling error, this includes all words in the text. However, for noun transfer, only nouns in the text would be included. Thus, a 20% magnitude spelling error perturbation involves more words being perturbed than in a 20% magnitude noun transfer perturbation. This is likely why in our results, spelling error perturbations show the largest bars across all quartiles for all analyses. Depending on the text feature and context, different methods for normalization may be preferable.

While this study focused on a limited selection of linguistic variations, the audit procedure can be applied for analysis of various other text features. This gives stakeholders flexibility to choose those features that are most appropriate to their context and use case. Analysis of additional features of essays (ie. length) could also allow for investigation of potential moderating effects on linguistic perturbations.

When establishing baseline reliability, only two trials of test-retest analysis were conducted. While this decision was made for illustration of concept, for more robust audits of LLMs, more replications should be included. On a related note, perturbed texts were only passed to the LLM for scoring once each in this study - more replications could be considered for improved auditing of LLM scoring. Increasing replications of both original and perturbed scores would additionally afford the opportunity to analyze the effect of text features on scoring at the individual essay level.

Another natural future direction of this study is to incorporate higher moments for analyses, which have the potential to illustrate further nuances of potential inequities in LLM scoring. There is no theoretical limit to moments that can be analyzed.

## 8 Conclusion

Our study has provided an example procedure for evaluating LLM scoring of texts for equity, incorporating algorithmically introduced linguistic perturbations and higher order moment analyses in characterizing impacts on stakeholders. We believe this procedure to be useful in the following ways:

First, in such contexts as educational testing with

multilingual student populations where annotated data is sparse, such a procedure has the potential to augment our ability to evaluate whether LLMs are ethically appropriate for application.

Second, due to the experimental nature of this process whereby the effects of the perturbation can be isolated, sources of LLM bias can be directly studied. Furthermore, LLM scoring does not suffer from carryover effects the way human raters might, allowing true replications of scores to be obtained for study. By further investigating how distributions of target features may vary across groups (e.g. multilingual vs. monolingual students), stakeholders can leverage audit results to infer how LLM scores may manifest as bias at the subpopulation level.

Third, this procedure is accessible in that it can be conducted by stakeholders in various contexts (not just education) for evaluation of their target LLM, and with respect to various features beyond those targeted in this study.

## References


Terry A. Ackerman. 1992. A Didactic Explanation of Item Bias, Item Impact, and Item Validity From a Multidimensional Perspective. *Journal of Educational Measurement*, 29(1):67–91.

Ben Hammer, Jaison Morgan, lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The Hewlett Foundation: Automated Essay Scoring.

Frederic M. Lord. 1980. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Erlbaum, Hillsdale, NJ.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*.

Bruno D Zumbo. 2007. Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, and Where It Is Going. *Language Assessment Quarterly*, 4(2):223–233.