Comparison of Originality Intensity (OI) Across Demographic Subgroups

EXECUTIVE SUMMARY

The goal of this report is to determine whether Originality Intensity (OI), an automated computational index of how far a single text diverges from common AI output, assigns systematically different scores to students based on identity and other characteristics inherent to the individual: race/ethnicity, gender, English language learner (ELL) status, economic disadvantage, and disability status. We examined this using an independently collected dataset described in a peer-reviewed article in *Assessing Writing*: the held-out test partition of the Automated Student Assessment Prize 2.0 (ASAP 2.0) corpus (Crossley, Baffour, Burleigh, & King, 2025), comprising 7,421 argumentative essays written by U.S. students in grades 6 through 10. The dataset is open and all analyses can be independently verified. Neither the essays nor the demographic information originate with us, and the dataset was purpose-built to compare groups varying in individual characteristics like identity and disability.

The majority of research on bias in AI-based evaluation of writing, including large language models (LLMs) used as judges, automated essay scoring, and AI-text detectors, finds that racial and ethnic minority students, English language learners, students with disabilities, and speakers of non-dominant dialects receive **lower** scores or less favorable judgments than their peers. The present results show the **reverse**. Differences between groups on OI were small to moderate, and in every case the historically disadvantaged group scored **higher**. Black/African American (d = 0.33), Hispanic/Latino (d = 0.27), Asian/Pacific Islander (d = 0.20), and multiracial students (d = 0.10) scored above White students; ELL students scored above non-ELL students (d = 0.39); students with an identified disability scored above students without one (d = 0.48); and economically disadvantaged students scored slightly above students who were not (d = 0.12). Female students scored slightly higher than male students (d = 0.08).

While more research is needed, the most likely explanation of these results follows from what OI measures. OI captures divergence from common AI output, and students from racial and ethnic minority backgrounds, students learning English, and students with disabilities are likely to bring distinctive backgrounds, experiences, and perspectives to their writing, in both the ideas they develop and the style in which they express them, that depart from the most common ideas and style from AI-generated text. OI appears to be capturing that divergence rather than penalizing it, which is the opposite of what has been documented for LLM judges and other automated evaluations of writing.

Background

When a large language model (LLM) is asked to judge a piece of writing, how does the identity of the writer impact the judgment? LLMs are now routinely used as judges of text, essays, ranking job applicants, and evaluating other models' outputs. A growing body of evidence shows that LLMs' judgments vary with the writer's identity in ways the content does not justify. Hofmann, Kalluri, Jurafsky, and King (2024), writing in *Nature*, showed that language models hold negative, covert raciolinguistic stereotypes about speakers of African American English, leading models to assign them to less prestigious jobs and to recommend conviction and the death penalty more often; alignment to human preferences concealed rather than removed this bias. In a randomized audit of roughly 361,000 résumés scored by five commercial LLMs, An, Huang, Lin, and Tai (2025) found that otherwise identical candidates received lower scores when signaled as Black and male. In the evaluation of student writing specifically, Warr, Oster, and Isaac (2025) found that ChatGPT's scores for identical passages shifted with implied racial cues even when explicit racial labels did not. Other researchers found that AI assigned lower scores to Black/African American than to White students on held-out essays from the same corpus analyzed in this paper (Burleigh, Holmes, Boser, Smith, & Baffour, 2025).

Audits of LLM-as-a-judge systems document numerous systematic biases (Ye et al., 2025), including a preference for text resembling the judge's own generations over human writing of equal quality (Panickssery, Bowman, & Feng, 2024), which by construction disadvantages writers whose language departs from the model's default. Laurito et al. (2025), writing in the *Proceedings of the National Academy of Sciences* (*PNAS*), adapted classic employment-discrimination designs to show that LLMs acting as decision-makers systematically chose items (products, academic

papers, and films) whose descriptions were written by LLMs over humans, a preference human evaluators showed far more weakly. They termed this preference of artificial intelligence (AI) systems for AI-generated content AI–AI bias, and warned that LLMs deployed as gatekeepers will implicitly favor AI agents and AI-assisted people. This bias has measurable consequences.

In a large, controlled résumé correspondence experiment, Xu, Li, and Jiang (2025) found that LLMs preferred résumés they had generated themselves over human-written résumés of comparable quality in 68% to 92% of comparisons across commercial and open-source models. In simulated hiring pipelines, candidates who used the same model as the screening LLM were 23% to 60% more likely to be shortlisted than equally qualified candidates who submitted their own writing. In a published commentary on Laurito et al., Yu (2025), also in *PNAS*, argues that such preferences may stem less from recognizing AI authorship than from the stylistic regularities LLMs favor, and warns of gatekeeping effects in which human writers who do not use, or do not write like, these systems are marginalized as LLMs take on evaluative roles across domains.

The same pattern appears when automated systems judge whether writing was produced by a human or AI – that is, in AI detectors. More than half of essays written by non-native English speakers were misclassified as AI-generated by widely used detectors, while essays written by native English speaking, U.S. eighth graders were classified almost perfectly (Liang, Yuksekgonul, Mao, Wu, & Zou, 2023); Black teenagers were roughly twice as likely as White or Latino teenagers to report schoolwork falsely flagged as AI-generated (Madden, Calvin, Hasse, & Lenhart, 2024); and texts by likely-autistic writers were flagged more often than texts by others (Chambers & Kelley, 2026). Across these literatures, the consistent finding is that automated judgments disfavor racial and ethnic minority writers, English language learners, and writers with disabilities.

We investigate whether Originality Intensity (OI), an automated computational index of how far a single text diverges from common AI output, reproduces this pattern, assigning lower scores to the groups that other automated judgments disfavor, using ASAP 2.0, a corpus built explicitly to permit bias analyses of automated scoring (Crossley et al., 2025).

Race/Ethnicity

The analysis includes 7,377 essays from the five racial/ethnic groups large enough to analyze (see Appendix A2). Mean Originality Intensity (OI) differed across groups, $F(4, 7372) = 35.33$, $p = 2.8 \times 10^{-29}$, $\eta^2 = .019$, and the ordering runs opposite to the one reported for large language model (LLM) judges and artificial intelligence (AI) text detectors. White students had the lowest mean OI ($M = 90.9$), followed by multiracial ($M = 97.5$), Asian/Pacific Islander ($M = 104.1$), Hispanic/Latino ($M = 109.3$), and Black/African American students ($M = 113.5$); see Figure 1.

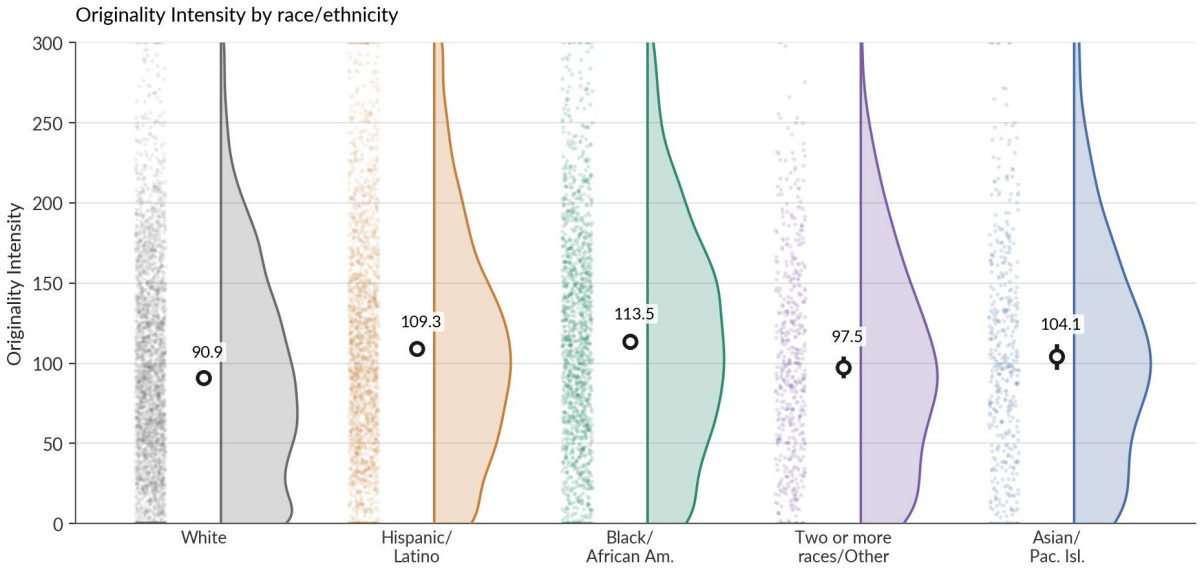


Figure 1. Originality Intensity (OI) by race/ethnicity. Each group shows the kernel density of the raw scores (right), every essay jittered (left), and the group mean with its 95% confidence interval (center, open circle).

Table 1 compares each group with White students, using Cohen's *d* and the *p*-value from a *t*-test. Positive values indicate that the named group scored higher than White students. Because four comparisons are made, *p*-values are also reported after Holm's (1979) sequential correction (see Appendix A4 for estimation details).

Table 1. Originality Intensity (OI) by Race/Ethnicity, with each group compared with White students. n = number of essays; M = mean; SD = standard deviation; d = Cohen's d; CI = confidence interval.

Group	*n*	*M* (*SD*)	*d* [95% CI]	*p*-value	Holm *p*
White	2,852	90.9 (66.4)	—	—	—
Hispanic/Latino	2,186	109.3 (71.0)	0.27 [0.21, 0.32]	4.5×10^{-21}	1.4×10^{-20}
Black/African American	1,382	113.5 (70.1)	0.33 [0.27, 0.40]	3.4×10^{-24}	1.4×10^{-23}
Two or more races/Other	531	97.5 (64.8)	0.10 [0.01, 0.19]	.034	.034
Asian/Pacific Islander	426	104.1 (69.2)	0.20 [0.10, 0.30]	1.4×10^{-4}	2.9×10^{-4}

Gender, English language learner status, economic status, and disability status

The four remaining characteristics show the same direction of findings as the Race/Ethnicity results (Figure 2). Female students scored slightly higher than male students (M = 104.6 versus 99.3) (d = 0.08). Economically disadvantaged students scored modestly higher than students who were not (M = 105.2 versus 96.7, d = 0.12). The two largest differences in the report involve groups that prior research identifies as most exposed to unfavorable automated judgments. English language learner (ELL) students scored higher than non-ELL students (M = 124.9 versus 98.2, d = 0.39), and students with an identified disability scored higher than students without one (M = 129.6 versus 97.2, d = 0.48).

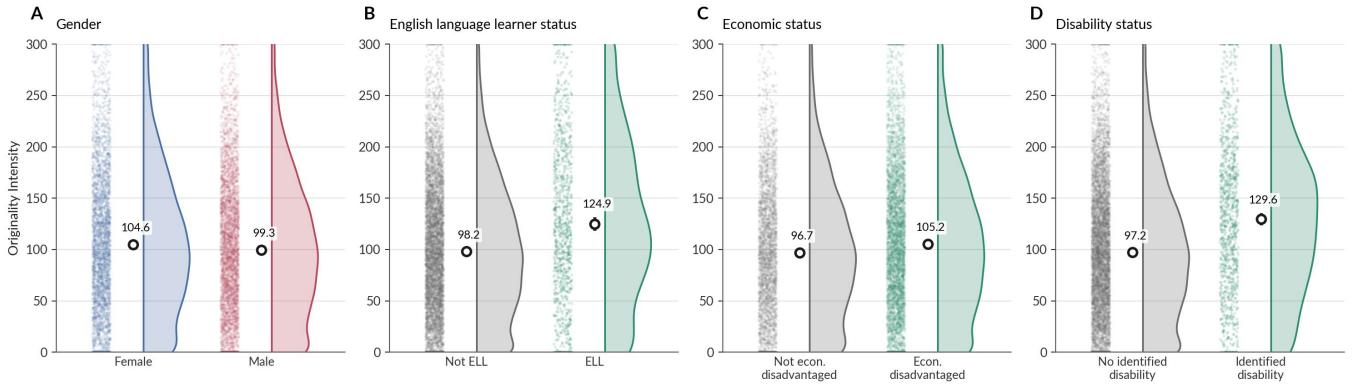


Figure 2. Originality Intensity by gender (A), English language learner status (B), economic status (C), and disability status (D), plotted with the conventions of Figure 1. In every panel, the group that prior research finds is disfavored by automated judgments scores higher on average.

Table 2 reports the corresponding effect sizes. Positive values indicate that the first-named group (female, ELL, economically disadvantaged, identified disability) scored higher. Welch-corrected tests give the same conclusions throughout.

Table 2. Originality Intensity (OI) by gender, English language learner (ELL) status, economic status, and disability status. n = number of essays; M = mean; SD = standard deviation; d = Cohen's d; CI = confidence interval; df = degrees of freedom.

Comparison	*n*	*M* (*SD*)	*d* [95% CI]	*p*-value
Gender (df = 7409)				
Female	3,475	104.6 (70.4)	0.08 [0.03, 0.12]	9.2×10^{-4}
Male	3,936	99.3 (67.9)		
English language learner status (df = 7189)				
ELL	998	124.9 (80.4)	0.39 [0.32, 0.46]	6.7×10^{-30}
Not ELL	6,193	98.2 (66.6)		
Economic status (df = 7409)				
Economically disadvantaged	4,481	105.2 (70.8)	0.12 [0.08, 0.17]	2.7×10^{-7}
Not economically disadvantaged	2,930	96.7 (66.2)		
Disability status (df = 7409)				
Identified disability	1,054	129.6 (75.8)	0.48 [0.41, 0.54]	8.8×10^{-46}
No identified disability	6,357	97.2 (66.9)		

Summary

Originality Intensity (OI) does not reproduce the pattern of bias documented for large language model (LLM) judges and other automated evaluations of writing. No group that prior research identifies as disfavored received lower OI scores than its comparison group. Instead, differences were small to moderate and consistently reversed: Black/African American, Hispanic/Latino, Asian/Pacific Islander, and multiracial students scored higher than White students. English language learner (ELL) students, students with an identified disability, and economically disadvantaged students scored higher than their peers.

It is important to note that all essays were rated on overall quality by expert human raters, who were unaware of any of the authors' individual characteristics during the rating process. An analysis of essay quality revealed majority groups tended to score higher than minority groups (Crossley et al., 2025). To ensure differences in essay quality could not confound OI comparisons between groups, we conducted multiple analyses that statistically control for essay quality across groups and all OI comparisons showed the same patterns and effect sizes compared to analyses without controlling for essay quality (see Appendix B). Put simply, differences in essay quality does not account for or contribute to the reversal outcomes we found, where minority groups showed higher OI scores than majority groups.

While more research is needed, these results are most plausibly explained by what OI is designed to capture. Because OI measures divergence from common artificial intelligence (AI) output, writing that draws on experiences, perspectives, and ways of expressing ideas that are less represented in the most common patterns of AI-generated text will score higher. Students from racial and ethnic minority backgrounds, students learning English, and students with disabilities are likely to bring such distinctive backgrounds to their ideas and style, and OI appears to be rewarding that divergence rather than penalizing it.

Appendix A: Methodology

A1. Dataset

The essays come from the Automated Student Assessment Prize 2.0 (ASAP 2.0) corpus, introduced by Crossley, Baffour, Burleigh, and King (2025) in *Assessing Writing*. ASAP 2.0 is the successor to the original ASAP corpus, which the William and Flora Hewlett Foundation funded in 2011 to test whether machine learning models could score student essays as reliably as trained human raters, and which became the standard training corpus for automated essay scoring. The original corpus had well-known limitations: its eight prompts spanned four different writing tasks and were scored on different rubrics with scales ranging from 0–3 to 0–30, six of the eight prompts were handwritten and later transcribed, the source texts for its source-based prompts were not released, only its training set was made public, and it contained no demographic information about the writers, making it impossible to test automated scoring models for bias. ASAP 2.0 was built to address each of these problems. Its stated purpose is to support the development of more sophisticated and unbiased automated scoring systems, and in particular to allow researchers to assess bias in scoring models by grade level, topic, race/ethnicity, English language learner (ELL) status, and gender.

ASAP 2.0 contains 24,278 persuasive essays written by U.S. students in grades 6, 8, 9, and 10 as part of state-level standardized writing tests. All essays were typed on computers, and all were source-based: students read one or more source texts and were required to integrate information from them into an argument. Of the essays, 12,871 were drawn from the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) 2.0 corpus (itself selected from a larger collection of roughly 500,000 student essays to resemble the U.S. secondary public school population in race, gender, and economic composition), and 11,829 had never been released. Essays were chosen to maximize the demographic information available for each writer. The published descriptions identify the data as coming from state standardized assessments in the United States but do not name the participating states or the years of test administration. The corpus was collected, curated, and scored by The Learning Agency and colleagues at Vanderbilt University, with support for the PERSUADE collection from the Bill & Melinda Gates Foundation, the Chan Zuckerberg Initiative, and Schmidt Futures.

Every essay was scored holistically on a 1–6 rubric adapted from the Scholastic Aptitude Test (SAT) for source-based writing, which rewards an effective point of view, critical thinking, appropriate use of evidence from the sources, organization, and command of language. Scoring was done by 34 expert raters employed by an outside consulting firm, using a double-blind process with 100% adjudication; inter-rater agreement before adjudication was weighted $\kappa = .745$ for the PERSUADE essays and $\kappa = .659$ for the new essays. All raters completed implicit-bias training modules from the Kirwan Institute at The Ohio State University, wrote reflections on them, and participated in group training on bias in essay scoring before being trained on each prompt. Raters also flagged personally identifying information, which was removed.

In 2024 the corpus was the basis of a public Kaggle competition in which roughly 70% of essays were released for model training and 30% were withheld as a test set, stratified by score, prompt, and demographic information. Competitors saw only the essay text and its score, never the demographic data. More than 2,500 teams developed over 60,000 scoring models. The full corpus, including the withheld test set, source texts, and rubric, has since been released under a Creative Commons license. The present analysis uses that withheld test partition, $N = 7{,}421$ essays (5,164 in the private test set and 2,257 in the public validation set), written in grades 6, 8, and 10 in response to five source-based prompts. Because these essays were never released during the competition and Originality Intensity (OI) was not developed on them, they provide an independent sample for this analysis.

A2. Demographic variables and sample size

Variable definitions follow Crossley et al. (2025). Economic disadvantage is eligibility for free or reduced-price lunch; disability status is having an individualized education plan or 504 plan; English language learner (ELL) status is as reported by the data sources. Records missing a given variable were excluded from analyses of that variable only: race/ethnicity was missing for 13 essays, gender, economic status, and disability status for 10 each, and ELL status for 230.

The corpus records six racial/ethnic categories, but only 31 essays in the test partition were written by American Indian/Alaskan Native students. A group of this size is too small to support a reliable comparison. Native students were therefore excluded from the racial/ethnic analyses, which are based on the remaining 7,377 essays. Their essays are retained in the analyses of gender, ELL status, economic status, and disability status.

A3. Originality Intensity

Originality Intensity (OI) was computed on each essay's full text exactly as in prior reports, with no reference to the prompt, the source text, or other essays. Scores range from 0 to 300 (mean $M = 101.8$, standard deviation $SD = 69.1$, median $Mdn = 95.8$). Every essay in the partition exceeds 150 words, so no length filter was needed and no exclusions were applied.

A4. Statistical procedures

Group means are reported with 95% confidence intervals from the t distribution. Cohen's d uses the pooled within-group standard deviation, and its confidence interval is obtained by inverting the noncentral t distribution. Significance tests are two-tailed independent-samples t tests assuming equal variances; Welch-corrected tests give the same conclusions throughout. The racial/ethnic omnibus test is a one-way analysis of variance (ANOVA), and the four comparisons with White students are corrected with Holm's (1979) sequential procedure.

Appendix B: Originality Intensity (OI) Comparisons in Samples Matched on Human-Rated Essay Quality Scores

B1. Purpose and matching procedure

This appendix accompanies the report *Examining Originality Intensity (OI) for Bias Linked to Student Identity and Individual Characteristics*, which compared OI across race/ethnicity, gender, English language learner (ELL) status, economic status, and disability status in the held-out test partition of the Automated Student Assessment Prize 2.0 (ASAP 2.0) corpus (Crossley, Baffour, Burleigh, & King, 2025). The groups compared in that report differ in their average human-rated essay scores. For example, Asian/Pacific Islander students had the highest mean holistic score and students with an identified disability the lowest. To determine whether the differences in OI reported there depend on these differences in rated essay quality, we repeated every comparison in samples in which the two groups being compared were matched on mean human-rated score.

Each comparison was matched separately, so each has its own matched sample; for race/ethnicity, each group was matched with White students in its own pair, mirroring the comparisons with White students in the main report. As in the main report, the 31 American Indian/Alaskan Native essays were excluded because the group is too small for a reliable comparison. Within each pair, essays were removed in equal numbers from both groups: one of the highest-scoring essays from the group with the higher mean score, then one of the lowest-scoring essays from the group with the lower mean score, alternating until the two group means differed by no more than 0.01 points on the 1–6 scale. Because scores take only whole values, the essay removed at a given score level was chosen at random. Removing essays from both ends in equal numbers brings the means together while retaining as many essays as possible from each group. *Table B1* shows the effect of matching. The number of essays removed ranged from 1 per group for the multiracial–White comparison to 375 per group for the disability comparison. Because the same number was removed from each group, the proportional loss was larger for the smaller group; the largest were for students with an identified disability (36% of 1,054), Asian/Pacific Islander students (23% of 426), and English language learner (ELL) students (16% of 998).

Table B1. Essays removed and mean human-rated essay scores before and after matching, by comparison. Scores are on the 1–6 holistic rubric. n = number of essays; M = mean.

Group	*n* original	*n* removed	*n* matched	Score *M* original	Score *M* matched
Hispanic/Latino vs White					
Hispanic/Latino	2,186	25	2,161	2.88	2.90
White	2,852	25	2,827	2.93	2.91
Black/African American vs White					
Black/African American	1,382	63	1,319	2.79	2.87
White	2,852	63	2,789	2.93	2.88
Two or more races/Other vs White					
Two or more races/Other	531	1	530	2.92	2.92
White	2,852	1	2,851	2.93	2.93
Asian/Pacific Islander vs White					
Asian/Pacific Islander	426	99	327	3.51	3.01
White	2,852	98	2,754	2.93	3.00
Female vs male					
Female	3,475	103	3,372	2.99	2.92
Male	3,936	102	3,834	2.86	2.91
ELL vs not ELL					
ELL	998	164	834	2.63	2.89
Not ELL	6,193	164	6,029	2.96	2.90
Economically disadvantaged vs not					
Economically disadvantaged	4,481	270	4,211	2.79	2.91
Not economically disadvantaged	2,930	271	2,659	3.12	2.92
Identified disability vs not					
Identified disability	1,054	375	679	2.37	2.87
No identified disability	6,357	375	5,982	3.01	2.88

B2. Race/ethnicity

After matching on mean human-rated score, every racial/ethnic group still scored higher on Originality Intensity (OI) than White students (Figure B1, Table B2). Black/African American (d = 0.30, compared with 0.33 before matching), Hispanic/Latino (d = 0.26, compared with 0.27), and Asian/Pacific Islander students (d = 0.28, compared with 0.20) scored above White students, and the difference for multiracial students remained small (d = 0.10, compared with 0.10). All four comparisons remained statistically significant after Holm's (1979) correction. Because each group was matched with White students in a separate pair, the omnibus test across all five groups was not repeated.

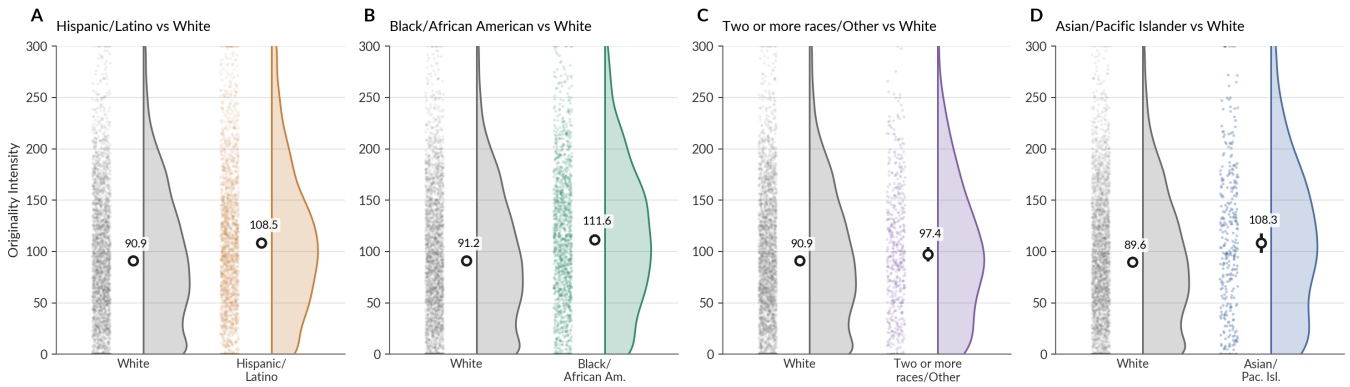


Figure B1. Originality Intensity for each racial/ethnic group and White students in samples matched on mean human-rated essay score (A, Hispanic/Latino; B, Black/African American; C, two or more races/other; D, Asian/Pacific Islander), with each group's raw scores shown as a kernel density (right) and jittered points (left), and the group mean with its 95% confidence interval (center, open circle). The White sample differs slightly across panels because each pair was matched separately.

Table B2. Originality Intensity by race/ethnicity in score-matched pairs, with each group compared with White students in its own matched pair. n = number of essays; M = mean; SD = standard deviation; d = Cohen's d; CI = confidence interval.

Group	*n*	Score *M*	OI *M* (*SD*)	*d* [95% CI]	*p*-value	Holm *p*
Hispanic/Latino vs White (df = 4986)						
Hispanic/Latino	2,161	2.90	108.5 (70.5)	0.26 [0.20, 0.31]	2.0×10^{-19}	7.8×10^{-19}
White	2,827	2.91	90.9 (66.5)			
Black/African American vs White (df = 4106)						
Black/African American	1,319	2.87	111.6 (69.1)	0.30 [0.24, 0.37]	2.2×10^{-19}	7.8×10^{-19}
White	2,789	2.88	91.2 (66.5)			
Two or more races/Other vs White (df = 3379)						
Two or more races/Other	530	2.92	97.4 (64.8)	0.10 [0.01, 0.19]	.036	.036
White	2,851	2.93	90.9 (66.4)			
Asian/Pacific Islander vs White (df = 3079)						
Asian/Pacific Islander	327	3.01	108.3 (74.1)	0.28 [0.17, 0.40]	1.6×10^{-6}	3.2×10^{-6}
White	2,754	3.00	89.6 (65.3)			

B3. Gender, English language learner status, economic status, and disability status

The four remaining comparisons also retained their direction after matching (Figure B2, Table B3). Female students scored slightly above male students ($d = 0.11$, compared with 0.08 before matching), and the small difference favoring economically disadvantaged students became smaller still ($d = 0.06$, compared with 0.12). The two largest differences were reduced but remained clearly present: English language learner (ELL) students scored above non-ELL students ($d = 0.27$, compared with 0.39), and students with an identified disability scored above students without one ($d = 0.33$, compared with 0.48). All four differences remained statistically significant, and Welch-corrected tests gave the same conclusions throughout.

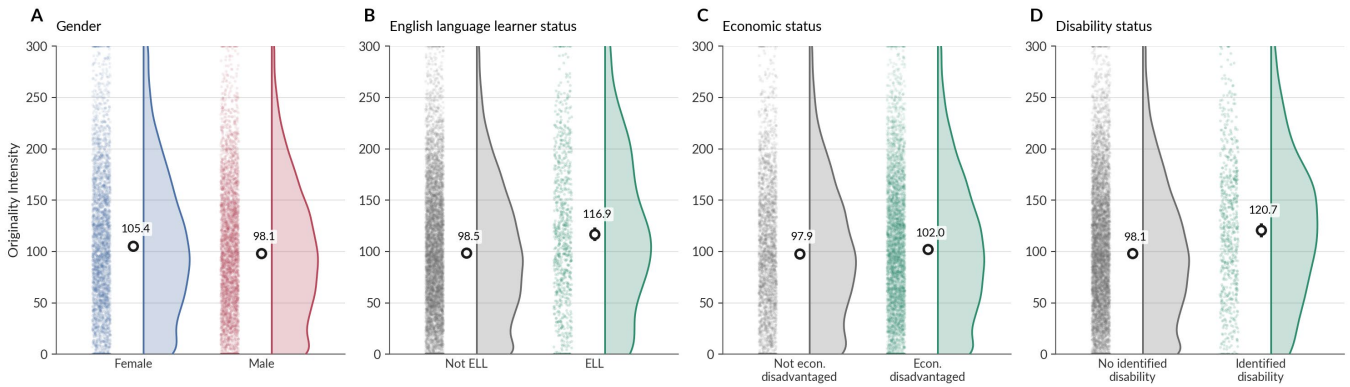


Figure B2. Originality Intensity by gender (A), English language learner status (B), economic status (C), and disability status (D) in samples matched on mean human-rated essay score, with each group's raw scores shown as a kernel density (right) and jittered points (left), and the group mean with its 95% confidence interval (center, open circle).

Table B3. Originality Intensity by gender, English language learner (ELL) status, economic status, and disability status in score-matched samples. n = number of essays; M = mean; SD = standard deviation; d = Cohen's d; CI = confidence interval; df = degrees of freedom.

Comparison	*n*	Score *M*	OI *M* (*SD*)	*d* [95% CI]	*p*-value
Gender (df = 7204)					
Female	3,372	2.92	105.4 (70.8)	0.11 [0.06, 0.15]	7.5×10^{-6}
Male	3,834	2.91	98.1 (67.2)		
English language learner status (df = 6861)					
ELL	834	2.89	116.9 (75.7)	0.27 [0.20, 0.34]	2.2×10^{-13}
Not ELL	6,029	2.90	98.5 (66.9)		
Economic status (df = 6868)					
Economically disadvantaged	4,211	2.91	102.0 (68.7)	0.06 [0.01, 0.11]	.015
Not economically disadvantaged	2,659	2.92	97.9 (67.2)		
Disability status (df = 6659)					
Identified disability	679	2.87	120.7 (69.9)	0.33 [0.25, 0.41]	2.2×10^{-16}
No identified disability	5,982	2.88	98.1 (67.5)		

B4. Statistical control for human-rated score in the full samples

Matching discards essays and equates only group means. As a complementary analysis that keeps every essay and each group's full score distribution, we repeated all comparisons in the full samples of the main report using analysis of covariance (ANCOVA), with the human-rated holistic score as a covariate. Each model regressed Originality Intensity (OI) on group membership and score. Group differences are reported as adjusted means evaluated at the overall mean score and as the standardized contrast $\psi*$ of Shieh (2023), the adjusted group difference divided by the residual standard deviation of the model, with an exact 95% confidence interval from the noncentral *t* distribution. Because the residual standard deviation is smaller than the raw standard deviation, $\psi*$ is not on exactly the same scale as Cohen's *d* (Olejnik & Algina, 2000); here the difference is small, because score explained about 5% of the variance in OI. As in the main report, the American Indian/Alaskan Native group was excluded from the race/ethnicity model, and the four comparisons with White students were corrected with Holm's (1979) procedure. We also tested whether the relationship between score and OI differed between groups (the homogeneity-of-slopes assumption) by adding a group-by-score interaction to each model.

Across all models, higher human-rated scores were associated with lower OI, about 14 to 16 points per rubric point. Controlling for score, mean OI still differed across racial/ethnic groups, $F(4, 7371) = 35.14$, $p = 4.1 \times 10^{-29}$, partial $\eta^2 = .019$, and the relationship between score and OI did not differ across racial/ethnic groups, $F(4, 7367) = 1.60$, $p = .171$. Every group scored higher than White students after adjustment (Table B4): Asian/Pacific Islander ($\psi* = 0.34$), Black/African American ($\psi* = 0.30$), Hispanic/Latino ($\psi* = 0.26$), and multiracial students ($\psi* = 0.10$), all

statistically significant after Holm correction. As in the matched analysis, the Asian/Pacific Islander difference was larger after adjustment than before (adjusted $M = 113.5$ compared with an observed M of 104.1), because that group's higher rated scores would otherwise predict lower OI.

Table B4. Originality Intensity by race/ethnicity controlling for human-rated essay score (ANCOVA), with each group compared with White students. Adjusted means are evaluated at the overall mean score. n = number of essays; M = mean; CI = confidence interval; ψ∗ = standardized contrast (Shieh, 2023).

Group	*n*	Observed *M*	Adjusted *M* [95% CI]	ψ∗ [95% CI]	*p*-value	Holm *p*
White	2,852	90.9	91.1 [88.6, 93.5]	—	—	—
Hispanic/Latino	2,186	109.3	108.6 [105.8, 111.4]	0.26 [0.21, 0.32]	2.7×10^{-20}	9.4×10^{-20}
Black/African American	1,382	113.5	111.4 [107.8, 114.9]	0.30 [0.24, 0.37]	2.3×10^{-20}	9.4×10^{-20}
Two or more races/Other	531	97.5	97.5 [91.8, 103.2]	0.10 [0.00, 0.19]	.041	.041
Asian/Pacific Islander	426	104.1	113.5 [107.1, 119.9]	0.34 [0.23, 0.44]	1.3×10^{-10}	2.6×10^{-10}

The other four characteristics showed the same pattern after adjustment (Table B5). Students with an identified disability ($\psi* = 0.35$), ELL students ($\psi* = 0.32$), and female students ($\psi* = 0.11$) scored higher than their comparison groups, and economically disadvantaged students scored only marginally higher ($\psi* = 0.05$, $p = .041$). The homogeneity-of-slopes assumption held for disability status ($p = .298$) but not for gender, ELL status, or economic status (all $p < .001$). In those three comparisons, the group difference was largest among essays with low rated scores and shrank toward zero among essays with high rated scores. The ELL advantage was 29.9 points at a score of 2, 17.4 points at a score of 3, and 4.9 points at a score of 4 (not significant); the female advantage was 12.9 points at a score of 2, 7.0 points at a score of 3, and 1.0 points at a score of 4 (not significant); and the economic-status difference was 10.5 points at a score of 2 and not significant at scores of 3 or 4. For these three characteristics, the adjusted differences in Table B5 are averages over this varying pattern. At no score level did any group difference reverse direction significantly.

Table B5. Originality Intensity by gender, English language learner (ELL) status, economic status, and disability status controlling for human-rated essay score (ANCOVA). Adjusted means are evaluated at the overall mean score. n = number of essays; M = mean; CI = confidence interval; ψ∗ = standardized contrast (Shieh, 2023); df = residual degrees of freedom.

Comparison	*n*	Observed *M*	Adjusted *M* [95% CI]	ψ∗ [95% CI]	*p*-value
Gender (df = 7408)					
Female	3,475	104.6	105.7 [103.5, 108.0]	0.11 [0.06, 0.16]	2.4×10^{-6}
Male	3,936	99.3	98.3 [96.2, 100.4]		
English language learner status (df = 7188)					
ELL	998	124.9	120.4 [116.2, 124.6]	0.32 [0.25, 0.39]	1.0×10^{-20}
Not ELL	6,193	98.2	98.9 [97.3, 100.6]		
Economic status (df = 7408)					
Economically disadvantaged	4,481	105.2	103.1 [101.1, 105.1]	0.05 [0.00, 0.10]	.041
Not economically disadvantaged	2,930	96.7	99.8 [97.4, 102.3]		
Disability status (df = 7408)					
Identified disability	1,054	129.6	121.9 [117.7, 126.0]	0.35 [0.28, 0.42]	1.5×10^{-24}
No identified disability	6,357	97.2	98.5 [96.8, 100.1]		

B5. Interpretation

Neither matching on mean human-rated essay quality nor controlling for it statistically in the full samples removed or reversed any of the group differences in Originality Intensity (OI), and the two approaches gave closely similar estimates. Every group that scored higher in the main report still scored higher in the score-matched samples, and every comparison remained statistically significant. The racial/ethnic differences were essentially unchanged, and the Asian/Pacific Islander difference grew once that group's highest-rated essays were set aside, which indicates that

higher OI is not simply a feature of weaker essays. The ELL and disability differences were reduced by roughly one-third, showing that part of those differences was associated with lower rated essay quality in those groups, but differences of about a quarter to a third of a standard deviation remained. The economic-status difference, already small, was halved.

There are limitations to note. First, matching equates group means on the human-rated score but not the full score distributions; the ANCOVA retains the full distributions but assumes a linear relationship between score and OI, and for gender, ELL status, and economic status that relationship differed between groups. Second, trimming removes the highest-scoring essays from one group and the lowest-scoring essays from the other, so the matched samples are less representative of each group than the full samples; for example, the matched Asian/Pacific Islander sample excludes that group's highest-rated essays, and the matched disability sample excludes more than a third of that group. Third, in both analyses, the human-rated score is itself a judgment that may contain group differences unrelated to writing quality.

References

An, J., Huang, D., Lin, C., & Tai, M. (2025). Measuring gender and racial biases in large language models: Intersectional evidence from automated resume evaluation. *PNAS Nexus, 4*(3), pgaf089. https://doi.org/10.1093/pnasnexus/pgaf089

Burleigh, L., Holmes, L., Boser, U., Smith, K., & Baffour, P. (2025, March 31). *Identifying limitations and bias in ChatGPT essay scores: Insights from benchmark data*. The Learning Agency. https://the-learning-agency.com/the-cutting-ed/article/identifying-limitations-and-bias-in-chatgpt-essay-scores-insights-from-benchmark-data/

Chambers, S., & Kelley, M. C. (2026). *The misclassification of autistic writing as AI-generated* (arXiv:2607.14729). arXiv. https://arxiv.org/abs/2607.14729

Crossley, S. A., Baffour, P., Burleigh, L., & King, J. (2025). A large-scale corpus for assessing source-based writing quality: ASAP 2.0. *Assessing Writing, 65*, 100954. https://doi.org/10.1016/j.asw.2025.100954

Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). AI generates covertly racist decisions about people based on their dialect. *Nature, 633*, 147–154. https://doi.org/10.1038/s41586-024-07856-5

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Laurito, W., Davis, B., Grietzer, P., Gavenčiak, T., Böhm, A., & Kulveit, J. (2025). AI–AI bias: Large language models favor communications generated by large language models. *Proceedings of the National Academy of Sciences, 122*(31), e2415697122. https://doi.org/10.1073/pnas.2415697122

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns, 4*(7), 100779. https://doi.org/10.1016/j.patter.2023.100779

Madden, M., Calvin, A., Hasse, A., & Lenhart, A. (2024). *The dawn of the AI era: Teens, parents, and the adoption of generative AI at home and school*. Common Sense Media.

Olejnik, S., & Algina, J. (2000). Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology, 25*, 241–286. https://doi.org/10.1006/ceps.2000.1040

Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems, 37*, 68772–68802.

Shieh, G. (2023). Assessing standardized contrast effects in ANCOVA: Confidence intervals, precision evaluations, and sample size requirements. *PLOS ONE, 18*(2), e0282161. https://doi.org/10.1371/journal.pone.0282161

Warr, M., Oster, N. J., & Isaac, R. (2025). Implicit bias in large language models: Experimental proof and implications for education. *Journal of Research on Technology in Education, 57*(6). https://doi.org/10.1080/15391523.2024.2395295

Xu, J., Li, G., & Jiang, J. Y. (2025). AI self-preferencing in algorithmic hiring: Empirical evidence and insights. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 8*(3), 2757–2758. https://doi.org/10.1609/aies.v8i3.36755 (Extended version: arXiv:2509.00462)

Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N. V., & Zhang, X. (2025). Justice or prejudice? Quantifying biases in LLM-as-a-judge. In *Proceedings of the International Conference on Learning Representations (ICLR 2025)*. https://arxiv.org/abs/2410.02736

Yu, Z. (2025). Addressing the broader implications of AI–AI bias in decision-making systems. *Proceedings of the National Academy of Sciences, 122*(40), e2521127122. https://doi.org/10.1073/pnas.2521127122