

Gender Variation in Writing Styles: A Corpus-Based Statistical Linguistic Study

Dr. ENAAM HASHIM ALBASHEER MOHAMMED

Faculty of Education-Basic Education
Nile Valley University, Sudan

Abstract:

This study examined the relationships between demographic characteristics, essay length, and holistic writing performance using the PERSUADE 2.0 corpus, comprising 25,996 persuasive essays written by U.S. secondary school students. Descriptive and inferential statistical analyses, including multiple linear regression, were conducted to evaluate the independent contributions of demographic variables and essay length to holistic writing quality. Female students achieved significantly higher holistic scores than male students, although the effect size was small (Cohen's $d = 0.20$). Essay length was the strongest predictor of writing quality ($\beta = 0.391$, $p < .001$), explaining 23.8% of the variance in bivariate analysis. After controlling for essay length, grade level, English Language Learner (ELL) status, and socioeconomic status remained significant independent predictors of holistic scores, whereas race/ethnicity showed mixed effects. The final regression model explained 45.9% of the variance in writing performance. These findings indicate that persuasive writing quality reflects the combined influence of textual and demographic characteristics rather than essay length alone. The study provides large-scale empirical evidence to inform writing instruction, educational assessment, and the development of fair and responsible automated writing evaluation systems.

Keywords: persuasive writing, writing assessment, essay length, demographic characteristics, holistic scoring, PERSUADE 2.0

1. INTRODUCTION

Writing proficiency represents a foundational skill for academic success, career readiness, and full participation in civic life (Graham & Perin, 2007; National Center for Education Statistics, 2012). Despite its recognized importance, a substantial proportion of students in the United States fail to achieve competent writing levels. According to the 2012 National Assessment of Educational Progress (NAEP) Writing Report Card, only approximately 27% of eighth and twelfth graders scored at or above proficient in writing (NCES, 2012; Morris et al., 2025). This persistent underperformance has prompted researchers and educators to investigate the factors that distinguish proficient from developing writers, with the goal of informing pedagogical practices and interventions that can improve writing outcomes (Crossley et al., 2024; Maamuujav, 2025).

Among the various forms of writing assessed in educational contexts, argumentative writing occupies a central position. The K-12 Common Core State Standards (2010) explicitly emphasize the cultivation of argumentation skills, requiring students to introduce claims, distinguish claims from alternative claims, organize claims logically, and support them with logical reasoning and accurate evidence (Common Core State Standards Initiative, 2010; Crossley et al., 2024). Argumentative writing demands the ability to state positions, provide supporting evidence, address counterarguments,

and construct coherent, persuasive texts (Toulmin, 1958; Newell et al., 2011). Text-based argumentative writing—in which students must analyze source materials and integrate evidence into their arguments—is particularly challenging, as it requires the synthesis of reading comprehension, critical thinking, and written expression (Maamuujav, 2025; Plakans & Gebril, 2013).

The assessment of writing quality has traditionally relied on holistic scoring, a method in which trained human raters assign a single overall score based on their integrated judgment of multiple dimensions of writing performance (Steiss et al., 2022; Maamuujav, 2025). Holistic scoring rubrics typically evaluate aspects such as clarity and cogency of argument, quality of interpretation, organization of ideas, integration of evidence, syntactic variety, diction, and language conventions (National Writing Project, 2010; Maamuujav, 2025). Despite the widespread use of holistic scoring, concerns have been raised about the potential for rater bias, inconsistency, and the substantial resources required for human scoring at scale (Johnson & Zhang, 2024; Kwako & Ormerod, 2024). These concerns have driven growing interest in automated essay scoring (AES) systems that can provide efficient, consistent evaluations of student writing (Crossley et al., 2024; Shermis, 2014).

A critical prerequisite for developing fair and effective AES systems—and for understanding writing development more broadly—is the availability of large-scale corpora that link student writing to both holistic quality scores and comprehensive demographic information (Crossley et al., 2022, 2024; Johnson & Zhang, 2024). The Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) corpus, now in its 2.0 version, addresses this need by providing over 25,000 argumentative essays produced by 6th–12th grade students in the United States (Crossley et al., 2022, 2024). The corpus includes essays from both independent and source-based writing tasks across 15 prompts, along with holistic quality scores assigned by expert raters (inter-rater reliability of weighted $\kappa = .745$) and detailed demographic metadata for each writer (Crossley et al., 2024). This rich dataset enables systematic investigation of how demographic characteristics relate to writing performance, a line of inquiry essential for ensuring equitable assessment practices (Kwako & Ormerod, 2024; Johnson & Zhang, 2024).

Previous research has documented consistent demographic disparities in writing performance. Studies have repeatedly found that female students outperform male students in writing assessments across multiple grade levels and contexts (Kwako & Ormerod, 2024; Morris et al., 2025; Zhang, 2013). For instance, Kwako and Ormerod (2024), analyzing a subset of the PERSUADE corpus, found that automated scoring models magnified the marginal differences observed in human scoring, favoring female students over males. Similarly, grade level has been positively associated with writing quality, reflecting the cumulative effects of instruction and language development over time (Crossley et al., 2024; Maamuujav, 2025). However, the relationship between grade and writing proficiency is not strictly linear, as some studies have reported variability in writing quality across grade levels that may reflect differences in task demands, rubrics, or instructional contexts (Maamuujav, 2025; Morris et al., 2025).

Socioeconomic status (SES) and English Language Learner (ELL) status represent additional factors associated with writing outcomes. Students from economically disadvantaged backgrounds—often identified through eligibility for federal assistance programs such as free or reduced-price lunch—tend to receive lower writing scores compared to their more advantaged peers (Crossley et al., 2022, 2024; Kwako & Ormerod, 2024). Similarly, ELL students, who are developing proficiency in English

while simultaneously learning academic content, demonstrate lower average writing performance than their non-ELL peers (Crossley et al., 2024; Kwako & Ormerod, 2024). These disparities are concerning because they may reflect differential access to instructional resources, language development opportunities, or—critically—potential biases in assessment procedures (Johnson & Zhang, 2024; Kwako & Ormerod, 2024).

Racial and ethnic disparities in writing performance have also been documented. In the PERSUADE corpus, White students consistently score higher than Black/African American and Hispanic/Latino students (Kwako & Ormerod, 2024; Crossley et al., 2024). These differences persist even when controlling for other demographic factors, suggesting that race/ethnicity captures variance in writing performance that is not fully explained by socioeconomic status or ELL status alone (Kwako & Ormerod, 2024). Importantly, however, the interpretation of these group differences is complicated by the possibility that human ratings—and the rubrics on which they are based—may themselves reflect cultural biases that disadvantage certain groups (Johnson & Zhang, 2024; Warner, 2018). Without additional information, such as a set of "unbiased" anchor items for differential item functioning analysis, the sources of these differences cannot be definitively attributed to genuine differences in writing proficiency versus assessment bias (Kwako & Ormerod, 2024).

A robust and consistently observed correlate of writing quality is essay length. Research across multiple corpora and educational contexts has demonstrated a positive association between the number of words in an essay and its holistic quality score (Morris et al., 2025; Crossley et al., 2014; Kwako & Ormerod, 2024). In the PERSUADE corpus, essay length has been shown to be the strongest predictor of holistic essay scores, accounting for a substantial proportion of variance in writing quality (Morris et al., 2025; Crossley et al., 2024). This relationship is not merely an artifact of scoring rubrics; longer essays provide greater opportunities for elaboration, development of ideas, and demonstration of linguistic and rhetorical competence (Crossley & McNamara, 2016; Maamuujuv, 2025). Although essay length is consistently associated with higher holistic scores, previous research suggests that writing quality is multidimensional and cannot be explained by essay length alone (Crossley et al., 2014; Morris et al., 2025). Consequently, demographic and educational characteristics, including gender, grade level, English Language Learner (ELL) status, socioeconomic status, and race/ethnicity, should also be considered when examining writing performance. The present study therefore investigates the independent contributions of these variables to holistic essay quality using the PERSUADE 2.0 corpus.

Despite the substantial body of research on demographic differences and essay length in relation to writing quality, several important gaps remain. First, many studies examining demographic differences have not simultaneously controlled for a comprehensive set of demographic characteristics, making it difficult to disentangle the unique contributions of each factor (Kwako & Ormerod, 2024). Second, while the PERSUADE 1.0 corpus provided rich annotations of discourse elements (Crossley et al., 2022), it did not initially include holistic quality scores, limiting its utility for examining the relationship between demographic characteristics and overall writing quality. Third, most studies have focused on either independent writing tasks or source-based tasks in isolation, with fewer investigations comparing findings across these task types or examining how demographic factors may interact with task type to predict writing quality (Crossley et al., 2024). Fourth, although prior research has established associations between demographic characteristics and writing performance, the magnitude and practical significance of these associations have not been systematically

examined in a large, demographically representative sample of U.S. secondary students (Crossley et al., 2024; Morris et al., 2025).

The PERSUADE 2.0 corpus (Crossley et al., 2024) provides a unique opportunity to address these gaps. As the largest publicly available corpus of student persuasive writing with both holistic quality scores and detailed demographic metadata, PERSUADE 2.0 enables researchers to examine the independent and combined contributions of gender, grade level, ELL status, socioeconomic status, race/ethnicity, and essay length to writing quality. The corpus includes essays from demographically diverse student populations, closely resembling the U.S. secondary public school population in terms of racial, gender, and economic composition (Crossley et al., 2022, 2024). The availability of both independent and source-based writing tasks across 15 prompts further enhances the generalizability of findings (Crossley et al., 2024). Moreover, the corpus includes holistic scores that were assigned through a rigorous double-blind rating process with 100% adjudication, ensuring high-quality criterion measures for validating the relationships between demographic characteristics and writing quality (Crossley et al., 2024).

The present study capitalizes on these features of the PERSUADE 2.0 corpus to conduct a comprehensive investigation of the relationships between demographic characteristics, essay length, and holistic writing quality. Specifically, we examine gender differences in holistic essay scores and essay length, the association between essay length and writing quality, and the independent contributions of demographic factors—including gender, grade level, ELL status, socioeconomic status, and race/ethnicity—to holistic writing quality after controlling for essay length. By analyzing a large, demographically diverse sample of student writing, this study aims to provide robust evidence regarding the factors that predict writing quality in text-based argumentative writing. The findings have important implications for writing assessment, pedagogical interventions, the development of equitable automated scoring systems, and the broader understanding of how demographic characteristics intersect with writing performance in secondary education.

The specific objectives of this study are to: (1) examine gender differences in holistic essay scores and essay length; (2) analyze the relationship between essay length and holistic writing quality; and (3) evaluate the independent contributions of gender, grade level, ELL status, socioeconomic status, and race/ethnicity to holistic writing quality, controlling for essay length.

The research questions guiding this study are as follows:

1. Are there statistically significant gender differences in holistic essay scores and essay length?
2. What is the relationship between essay length and holistic writing quality?
3. What are the relative contributions of demographic factors (gender, grade level, ELL status, socioeconomic status, and race/ethnicity) to holistic writing quality, after controlling for essay length?

Based on previous research, we hypothesize that: (a) female students will achieve higher holistic essay scores than male students, consistent with prior findings (Kwako & Ormerod, 2024; Zhang, 2013); (b) female students will produce longer essays than male students (Morris et al., 2025; Crossley et al., 2014); (c) essay length will be positively and substantially associated with holistic writing quality (Morris et al., 2025; Crossley et al., 2024); and (d) demographic factors will retain independent predictive power for holistic writing quality even after controlling for essay length, reflecting the

multifaceted nature of writing proficiency (Maamuujav, 2025; Crossley et al., 2014; Kwako & Ormerod, 2024).

The remainder of this paper is organized as follows. Section 2 provides a comprehensive review of the literature on writing assessment, demographic differences in writing performance, and the relationship between essay length and writing quality. Section 3 describes the methodology, including the PERSUADE 2.0 corpus, the measures used, and the statistical analyses conducted. Section 4 presents the results of the analyses. Section 5 discusses the findings in relation to previous research, addresses the limitations of the study, and explores implications for educational practice and future research. Section 6 concludes with a summary of the key contributions and recommendations.

2. LITERATURE REVIEW

2.1 Holistic Essay Scoring and Writing Assessment

The assessment of writing quality has traditionally relied on holistic scoring, a method in which human raters judge a text based on their integrated impression of multiple components (Steiss et al., 2022; Maamuujav, 2025). This approach is considered parsimonious and flexible, reflecting the view that writing quality emerges from the interaction of elements forming an integrated whole (Steiss et al., 2022). However, holistic scoring has been criticized for potential rater bias, inconsistency, and the substantial resources required for human scoring at scale (Johnson & Zhang, 2024; Kwako & Ormerod, 2024). Maamuujav (2025) further noted that holistic scoring may obscure important distinctions in how students deploy linguistic resources and rhetorical moves, as it collapses multiple dimensions into a single score.

Recent studies have sought to model the multidimensional nature of writing quality using latent variable approaches. Maamuujav (2025) found that a three-factor model—with lexical sophistication, syntactic complexity, and rhetorical features as distinct but covarying constructs—provided a close fit ($\chi^2 = 104.44$, $df = 24$, $CFI = 0.95$, $SRMR = 0.05$) and was superior to unidimensional models. Critically, rhetorical features significantly predicted holistic quality ($\beta = 0.96$, $p < 0.001$), whereas lexical sophistication ($\beta = 0.08$, $p = 0.548$) and syntactic complexity ($\beta = 0.12$, $p = 0.662$) were not significant predictors when rhetorical features were included. This suggests that linguistic resources may contribute indirectly to writing quality through their role in constructing effective rhetorical moves (Maamuujav, 2025). Similarly, Morris et al. (2025) identified four distinct writing profiles among highly rated essays—Structural, Academic, Reportive, and Conversational—each achieving high holistic scores through different combinations of linguistic features, further supporting the view that writing quality is a complex latent construct rather than a single linear outcome.

2.2 Demographic Differences in Persuasive Writing Performance

2.2.1 Gender

Recent studies generally suggest that female students outperform male students in writing assessments, though effect sizes are modest. Kwako and Ormerod (2024) found that female students scored higher than males in the PERSUADE corpus ($z = 0.06$, $SE = 0.01$, $p = .0012$), with Cohen's $d = 0.20$, indicating limited practical magnitude. Morris et al. (2025) reported that females represented 56.81% of high-scoring essays (score > 4) compared to 43.19% for males, confirming the female advantage among successful

writers. These findings are consistent with broader educational research (Zhang, 2013; as cited in Kwako & Ormerod, 2024). However, the small effect sizes suggest that gender explains only a small proportion of variance in writing quality, and the practical significance of these differences is debated (Kwako & Ormerod, 2024). However, previous studies have rarely examined whether gender differences remain after controlling simultaneously for essay length and other demographic characteristics.

2.2.2 Grade Level

Grade level has been associated with writing quality, reflecting cumulative effects of instruction and language development (Maamuujav, 2025). Maamuujav (2025) reported that average word count varied by grade level, with Grade 12 students producing the highest ($M = 216.9$, $SD = 96.5$) and Grade 8 the lowest ($M = 166.9$, $SD = 61.9$). However, the relationship is not strictly linear; variability in writing quality across grades may reflect differences in task demands, rubrics, or instructional contexts (Maamuujav, 2025; Morris et al., 2025). The PERSUADE 2.0 corpus includes essays from grades 6 through 12, with source-based essays from 6th to 10th grades and independent essays from 8th to 12th grades (Crossley et al., 2024), enabling systematic examination of grade-level differences across tasks. Consequently, the independent contribution of grade level remains unclear.

2.2.3 English Language Learner Status, Socioeconomic Status, and Disability

Students from linguistically and economically disadvantaged backgrounds tend to score lower on writing assessments. Kwako and Ormerod (2024) found that ELL students and those with low SES scored lower than reference groups, though these differences were not statistically significant in their bias analysis ($z = -0.10$, $SE = 0.04$, $p > .05$; and $z = -0.10$, $SE = 0.02$, $p > .05$, respectively). Morris et al. (2025) reported that among high-scoring essays, only 19.57% were from economically disadvantaged students, compared to 65.30% from non-disadvantaged students. Disability status was also associated with lower scores, though differences were not statistically significant in Kwako and Ormerod's (2024) analysis ($z = -0.07$, $p > .05$). Crossley et al. (2025) noted that sizeable missing data on economic and disability status in the ASAP 2.0 corpus may limit bias checks, a concern that extends to PERSUADE-based analyses.

2.2.4 Race and Ethnicity

Racial and ethnic disparities in writing performance have been documented consistently, though the magnitude varies across studies. Kwako and Ormerod (2024) found that XLNet magnified human-rated differences, favoring White students over Black students ($z = -0.07$, $SE = 0.01$, $p = .0023$). Johnson and Zhang (2024) reported more pronounced disparities: Asian/Pacific Islander students had an adjusted score difference of -1.16 compared to the overall mean difference of -0.90 in zero-shot GPT-4o scoring, indicating they were scored an extra quarter-point lower on average. Importantly, GPT-4o was better at predicting race/ethnicity ($\kappa = 0.26$) than at scoring essays ($\kappa = 0.08$), suggesting that the model may infer and use construct-irrelevant demographic information (Johnson & Zhang, 2024). The racial composition of the PERSUADE corpus closely resembles U.S. secondary schools, with White students constituting 45%, Hispanic 25%, Black 19%, and Asian/Pacific Islander 6.7% (Crossley et al., 2022). Among high-scoring essays, White and Asian/Pacific Islander students are overrepresented, while Hispanic and Black students are underrepresented (Morris et al., 2025). These inconsistencies warrant further investigation using multivariable statistical models.

2.3 The Role of Essay Length in Holistic Writing Quality

Essay length is a robust and consistently observed correlate of writing quality across corpora and contexts (Morris et al., 2025; Crossley et al., 2014). In the PERSUADE corpus, high-scoring essays (score > 4) were substantially longer on average ($M = 672.96$, $SD = 204.45$) compared to the corpus mean ($M = 402.31$, $SD = 188.38$) (Morris et al., 2025). This relationship is not merely an artifact of scoring rubrics; longer essays provide more opportunities for elaboration, idea development, and demonstration of linguistic and rhetorical competence (Crossley & McNamara, 2016; as cited in Morris et al., 2025). Johnson and Zhang (2024) reported that essays in the PERSUADE 2.0 corpus ranged from 146 to 4,855 words, with a median of 407 words, highlighting variation in length as a distinguishing feature of writing quality.

However, the role of essay length may vary across writing tasks and profiles. Morris et al. (2025) found that the Reportive writing profile—characterized by lower cohesion and high prepositional phrases—was strongly associated with text-dependent tasks (64.1% of the cluster vs. 23.5% of the total corpus), while Structural, Academic, and Conversational profiles were associated with independent tasks. This suggests that the relationship between length and quality may be moderated by task type and the linguistic features typical of each profile. The extent to which essay length mediates the relationship between demographic characteristics and writing quality remains unexplored. No study has used essay length as a control variable to determine whether demographic differences persist after accounting for length differences.

2.4 The PERSUADE 2.0 Corpus and Methodological Advances

The PERSUADE corpus represents a significant advance in writing assessment research, providing a large-scale, open-source resource with detailed demographic metadata (Crossley et al., 2022). PERSUADE 1.0 comprised over 25,000 essays annotated for discourse elements and hierarchical relationships, with demographic composition closely resembling U.S. secondary schools (Crossley et al., 2022). PERSUADE 2.0 extended this by providing holistic essay scores and effectiveness scores for discourse elements, assigned through a rigorous double-blind process with 100% adjudication (inter-rater weighted $\kappa = .745$, $r = .750$) (Crossley et al., 2024).

Research using the PERSUADE corpus has yielded insights into automated scoring bias (Kwako & Ormerod, 2024), writing profiles (Morris et al., 2025), and fairness evaluations (Johnson & Zhang, 2024). However, limitations remain: the corpus only includes 6th-12th grade writers, has a limited number of prompts ($N = 15$), and focuses exclusively on persuasive tasks (Crossley et al., 2024). Additionally, inter-rater reliability for discourse elements was difficult to calculate due to segmentation differences (overall $\kappa = .316$ for effectiveness ratings), which may impact validity when modeling the data (Crossley et al., 2024). Stahl et al. (2024) similarly noted the difficulty of achieving high inter-rater reliability for argumentative structure annotations, though their study achieved a values ranging from 0.74 to 0.89 for different annotation levels.

Methodological approaches across PERSUADE-based studies have varied substantially. Kwako and Ormerod (2024) used fine-tuned XLNet to examine bias, Johnson and Zhang (2024) used zero-shot GPT-4o for fairness evaluation, Maamujav (2025) used structural equation modeling to examine linguistic and rhetorical features, and Morris et al. (2025) used cluster analysis to identify writing profiles. These divergent approaches make direct comparisons challenging; findings are complementary rather than directly comparable. For instance, Morris et al.'s (2025) identification of writing profiles based on clustering complements Maamujav's (2025) latent construct modeling,

as both support the multidimensional nature of writing quality, but neither study examined how demographic factors predict writing quality within a unified framework.

2.5 Critical Synthesis and Identification of the Research Gap

The literature reveals several patterns and unresolved issues. First, there is consensus that demographic characteristics—gender, grade level, ELL status, SES, and race/ethnicity—are associated with writing performance, though effect sizes are generally small (Kwako & Ormerod, 2024; Johnson & Zhang, 2024; Morris et al., 2025). However, no study has simultaneously controlled for all these factors in a single analysis, making it difficult to estimate unique contributions. Race/ethnicity and SES are often confounded (Crossley et al., 2022, 2024), and studies that do not control for both may overestimate the effect of either.

Second, while essay length is a strong predictor of quality (Morris et al., 2025), the extent to which length mediates demographic differences is unknown. No study has used length as a control variable to determine whether demographic disparities reflect genuine proficiency differences versus length differences.

Third, the interpretation of group differences remains ambiguous. Disparities could reflect biases in human scoring, legitimate differences in proficiency, or both (Kwako & Ormerod, 2024). The finding that AI systems can predict race/ethnicity from essays (Johnson & Zhang, 2024) raises concerns about construct-irrelevant variance, but systematic analyses of demographic predictors with appropriate controls are lacking.

Fourth, previous PERSUADE-based studies have focused on linguistic features, discourse elements, or bias detection (Crossley et al., 2022, 2024; Kwako & Ormerod, 2024; Morris et al., 2025; Johnson & Zhang, 2024). Less attention has been paid to quantifying the independent and combined contributions of multiple demographic factors to holistic writing quality while controlling for essay length.

Collectively, previous research has examined demographic characteristics, linguistic features, automated scoring, and writing profiles separately. Nevertheless, few studies have investigated the simultaneous contribution of multiple demographic variables while controlling for essay length in a comprehensive statistical framework using the entire PERSUADE 2.0 corpus.

The present study addresses these gaps by conducting a comprehensive analysis of the relationships between demographic characteristics, essay length, and holistic writing quality using the PERSUADE 2.0 corpus. We simultaneously examine gender, grade level, ELL status, SES, and race/ethnicity in a single predictive model, using essay length as a control variable. We quantify the proportion of variance explained by demographic and textual variables, document and address missing data, and use the entire corpus aggregated across prompts to enhance generalizability. As noted by Kwako and Ormerod (2024), responsible use of automated systems requires evidence that they do not create or worsen inequities for marginalized populations. This study contributes to that evidence base by providing a comprehensive, large-scale analysis of demographic predictors of writing quality.

3. METHODOLOGY

3.1 Research Design

This study employed a quantitative observational design based on secondary analysis of the publicly available PERSUADE 2.0 corpus (Crossley et al., 2024). A secondary data approach is appropriate when investigating relationships among naturally occurring

educational variables without experimental manipulation. In the present study, the design enabled the examination of associations between demographic characteristics, essay length, and holistic writing performance using a large-scale corpus of persuasive essays. The availability of multiple demographic variables together with standardized human-assigned holistic scores provided an appropriate framework for evaluating the independent contribution of each predictor to writing performance.

The PERSUADE 2.0 corpus was selected because it provides standardized holistic scores together with detailed demographic metadata for a large population of student writers. The large sample size increases statistical precision and permits multivariable statistical modelling while reducing sampling error. The availability of demographic variables, essay length, and human scoring makes the corpus particularly suitable for investigating demographic variation in persuasive writing performance (Crossley et al., 2022, 2024).

3.2 Data Source

3.2.1 The PERSUADE 2.0 Corpus

Data for this study were drawn from the PERSUADE 2.0 corpus, a large-scale, open-source corpus of student persuasive writing. PERSUADE 2.0 builds on the PERSUADE 1.0 corpus, which comprised source-based essays ($n = 12,875$) and independent essays ($n = 13,121$). The PERSUADE 1.0 corpus included annotations for discourse elements (e.g., position, claims, and evidence) as well as hierarchical relationships between these elements. Detailed demographic information for each writer was also available, including gender, race/ethnicity, economic disadvantage, English Language Learner status, and disability status (Crossley et al., 2022).

PERSUADE 2.0 augmented this resource by providing holistic essay scores for each persuasive essay as well as effectiveness scores for each discourse element found in the PERSUADE 1.0 corpus. The corpus is open source and freely available to researchers (Crossley et al., 2024).

3.2.2 Writing Tasks

The PERSUADE corpus comprises two sub-corpora: source-based essays ($n = 12,875$) and independent essays ($n = 13,121$). Source-based writing requires the student to refer to a text while independent writing excludes this requirement. The source-based set was derived from seven unique writing prompts and related sources, reflecting students in grades 6 through 10. The independent set reflects writing where background knowledge of the topic was not a requirement, and no sources were required to produce the texts, collected from students in grades 8 through 12 and derived from eight unique writing prompts. All prompts and sources are available within the PERSUADE corpus (Crossley et al., 2022).

3.2.3 Holistic Scoring Procedure

Each essay in the PERSUADE corpus was reviewed by two expert raters from an educational consulting firm with two or more years of experience rating essays for quality (Morris et al., 2025). Prior to norming, raters received anti-bias training and anti-bias strategy instruction designed to address issues of bias that occur during scoring and are inherent to the use of standardized rubrics (Crossley et al., 2022; Morris et al., 2025).

For the independent essays, raters were trained on a standardized SAT holistic essay scoring rubric. The rubric used a 1 to 6 scale that had interval levels (i.e., the raters

were told that the distance between each grade was equal). A score of 6 indicates an essay that shows clear and consistent mastery of writing. The same SAT rubric was slightly modified for the source-based essays, with the main difference being the inclusion of evidence taken from the source as a criterion for writing quality. The two rubrics were matched as closely as possible to ensure that independent and source-based essays were comparable (Crossley et al., 2024).

The corpus was annotated using a double-blind rating process with 100% adjudication such that each essay was independently reviewed by two expert raters and adjudicated by a third expert rater (Crossley et al., 2022; Crossley et al., 2024). Inter-rater reliability before adjudication indicated strong agreement between the expert raters (weighted $\kappa = .745$, $r = .750$) (Crossley et al., 2024).

3.3 Participants and Sample

The PERSUADE corpus consists of 25,996 essays based on fifteen writing prompts, selected from a much larger corpus of approximately 500,000 essays typed by American students in grades 6-12 in several states across the United States (Crossley et al., 2022; Morris et al., 2025).

3.3.1 Complete Sample

The final dataset for the present study consisted of 25,996 persuasive essays. The sample included 13,142 essays written by female students (50.6%) and 12,854 essays written by male students (49.4%). The overall mean holistic essay score was 3.32 (SD = 1.16), and the average essay length was 418.13 words (SD = 275.11). The sample included students from multiple grade levels, with the largest proportions represented in the upper middle- and high-school grades.

3.3.2 Demographic Variables

The PERSUADE corpus includes detailed demographic information for each writer. All essays in the corpus are linked to information on the student's gender and race/ethnicity, with the majority of writers identifying as White (45%), followed by Hispanic (25%), Black (19%), and Asian/Pacific Islander (6.7%). A subset of the corpus ($n = 20,759$) also contains data on student eligibility for federal assistance programs such as free or reduced-price school lunch, Temporary Assistance for Needy Families, and the Supplemental Nutrition Assistance Programs, which are broadly defined as economic disadvantage. A large subsample of the essays in the corpus also includes information on English Language Learner status ($n = 24,787$) and disability status ($n = 24,828$) (Crossley et al., 2022).

3.3.3 Exclusion Due to Missing Values

Following listwise deletion of missing values for the regression analysis, the final model included 20,718 essays. This approach is consistent with the guidance of Crossley et al. (2024), who noted that sizeable missing data on demographic variables may limit bias checks. The complete sample of 25,996 essays was retained for descriptive statistics and bivariate analyses.

3.4 Variables

3.4.1 Dependent Variable

Holistic Essay Score. The dependent variable was the holistic essay score assigned to each essay by expert raters. Holistic scores were based on a standardized SAT rubric (1-6 scale) for independent essays, with a modified version for source-based essays. The

rubric is an interval-level scale, with scores ranging from 1 (minimum) to 6 (maximum), and the distance between each grade considered equal. A score of 6 indicates clear and consistent mastery of writing, while a score of 1 indicates very little or no mastery (Crossley et al., 2024). This variable was used as a continuous measure.

3.4.2 Independent Variables

Gender. Gender was coded as a categorical variable based on the metadata provided in the PERSUADE corpus. The variable included two categories: Male and Female.

Grade Level. Grade level was a categorical variable indicating the student's grade (6th through 12th grade). Grade levels were obtained from the PERSUADE corpus metadata.

Essay Length (Word Count). Essay length was operationalized as the total number of words in each essay. This was calculated automatically from the essay texts. The variable was treated as continuous.

English Language Learner (ELL) Status. ELL status was a categorical variable identifying students identified as English language learners. The variable was coded as identified ELL or not identified ELL (non-ELL).

Race/Ethnicity. Race/ethnicity was a categorical variable based on student self-identification as reported in the PERSUADE metadata. Categories included: White, Hispanic/Latino, Black/African American, Asian/Pacific Islander, and Two or More Races/Other.

Economically Disadvantaged Status. Socioeconomic status was operationalized as economically disadvantaged status, based on eligibility for K-12 federal assistance programs such as free or reduced-price school lunch, Temporary Assistance for Needy Families, and the Supplemental Nutrition Assistance Programs. The variable was coded as identified as economically disadvantaged or not identified as economically disadvantaged (not disadvantaged) (Crossley et al., 2022).

3.4.3 Variable Sources

All variables were sourced from the PERSUADE 2.0 corpus. The holistic essay scores and demographic metadata were generated through the rigorous annotation and scoring procedures described in Crossley et al. (2022, 2024).

Table 1. Variables Included in the Study

Variable	Type	Measurement Scale	Coding	Analysis
Holistic Essay Score	Continuous	1–6	Original score	All
Gender	Categorical	Male/Female	Female = reference	t-test, ANOVA, Chi-square, Regression
Grade Level	Ordinal	Grades 6–12	Original coding	ANOVA, Regression
Word Count	Continuous	Number of words	Raw value	Correlation, Regression
ELL Status	Binary	Yes/No	Non-ELL = reference	ANOVA, Chi-square, Regression
Economically Disadvantaged	Binary	Yes/No	Disadvantaged = reference	ANOVA, Chi-square, Regression
Race/Ethnicity	Categorical	Five categories	Dummy coded	ANOVA, Chi-square, Regression

3.5 Data Preparation

3.5.1 Missing Value Treatment

The PERSUADE 2.0 corpus includes comprehensive demographic information; however, as noted by Crossley et al. (2024), sizeable missing data are present for certain variables. For the regression analysis, listwise deletion was applied, resulting in a final analytic sample of 20,718 essays. This approach ensures that all analyses are based on the same set of cases with complete data for all predictors.

3.5.2 Variable Coding

Categorical predictors were entered into the regression model using dummy-variable coding automatically generated by the statsmodels package in Python. Female students, non-English Language Learners (non-ELL), economically disadvantaged students, and the reference race/ethnicity category specified by the software served as the baseline categories for regression comparisons.

3.5.3 Data Validation

Prior to statistical analysis, the dataset was inspected to verify variable types, coding consistency, and completeness of the variables included in the study. Descriptive statistics were generated for all study variables to ensure that the analytical dataset was consistent with the published characteristics of the PERSUADE corpus (Crossley et al., 2022, 2024). No additional modifications were made to the original corpus beyond the preprocessing procedures described above.

3.6 Statistical Analysis

3.6.1 Statistical Reporting

Statistical significance was evaluated using a two-sided α level of .05. Effect sizes were reported where appropriate to facilitate interpretation of the practical significance of statistically significant findings. Continuous variables are presented as means $\pm$ standard deviations, whereas categorical variables are presented as frequencies and percentages.

3.6.2 Descriptive Statistics

Descriptive statistics were computed for all variables. For continuous variables (holistic essay score and word count), means, standard deviations, and ranges were reported. For categorical variables (gender, grade level, ELL status, economically disadvantaged status, and race/ethnicity), frequencies and percentages were reported. Descriptive statistics were disaggregated by gender to provide initial comparisons between male and female students.

3.6.3 Welch's Independent-Samples t-Test

Welch's independent-samples t-test was employed to compare holistic essay scores and essay length between male and female students. This test was selected because it does not assume equal variances between groups, which is appropriate given the potential for unequal variances in educational data. The test was used to determine whether the observed differences in holistic essay scores and word count between female and male students were statistically significant.

3.6.4 Cohen's d

Cohen's d was calculated as a measure of effect size for the gender differences. Cohen's d values of 0.2, 0.5, and 0.8 are conventionally interpreted as small, medium, and large effects, respectively. This provides a practical significance measure to complement the statistical significance tests (Kwako & Ormerod, 2024).

3.6.5 Correlation Analyses

Pearson's product-moment correlation was used to examine the linear relationship between essay length (word count) and holistic essay score. The correlation coefficient (r) was interpreted using standard conventions. The coefficient of determination (R^2) was calculated to indicate the proportion of variance in holistic essay scores explained by essay length.

Spearman's rank-order correlation was also computed between essay length and holistic essay score. Spearman correlation was additionally calculated to evaluate the robustness of the association without assuming linearity or normally distributed variables.

3.6.6 One-Way ANOVA

One-way ANOVA was used to examine differences in holistic essay scores across demographic categories, including grade level, ELL status, economically disadvantaged status, and race/ethnicity. This analysis was conducted to examine differences in holistic essay scores across demographic groups before fitting the multivariable regression model.

3.6.7 Chi-Square Tests

Chi-square tests were used to examine associations between gender and other categorical demographic variables (grade level, ELL status, economically disadvantaged status, and race/ethnicity). This analysis was performed to describe the sample composition and to identify potential confounding relationships among demographic variables.

3.6.8 Multiple Linear Regression

A multiple linear regression (ordinary least squares) model was fitted to evaluate the independent contribution of each predictor while controlling for the remaining variables. The model included essay length (word count), gender, grade level, ELL status, economically disadvantaged status, and race/ethnicity as predictors. Following listwise deletion of missing values, the final regression model included 20,718 essays.

Multiple linear regression was selected because it allows for the simultaneous estimation of the unique contribution of each predictor to the dependent variable while controlling for all other predictors. All predictor variables were entered simultaneously into the regression model using the enter method. This approach enabled estimation of the unique association of each predictor with holistic essay score while statistically controlling for the remaining variables included in the model.

The regression equation was specified as:

Equation (1)

$$\text{Holistic Score} = \beta_0 + \beta_1(\text{Word Count}) + \beta_2(\text{Male}) + \beta_3(\text{Grade Level}) + \beta_4(\text{ELL}) + \beta_5(\text{Not Disadvantaged}) + \beta_6(\text{Asian/Pacific Islander}) + \beta_7(\text{Black/African American}) + \beta_8(\text{Hispanic/Latino}) + \beta_9(\text{White}) + \beta_{10}(\text{Two or More}) + e$$

where:

- β_0 is the intercept
- β_1 through β_{10} are the regression coefficients

- ε is the error term

Standardized coefficients (β) were reported to facilitate comparison of the relative importance of predictors measured on different scales. Unstandardized coefficients (B), standard errors (SE), t-statistics, and p-values were also reported.

Table 2. Statistical Analyses

Statistical Analysis	Purpose	Reported Statistics
Descriptive statistics	Summarize sample characteristics	Mean, SD, frequency, percentage
Welch's independent-samples t-test	Compare gender differences	t, p
Cohen's d	Estimate effect size	d
Pearson correlation	Examine linear association	r, p, R ²
Spearman correlation	Examine monotonic association	ρ , p
One-way ANOVA	Compare demographic groups	F, p
Chi-square test	Examine associations among categorical variables	χ^2 , p
Multiple linear regression (OLS)	Estimate independent predictors	B, β , SE, t, p, R ²
Variance Inflation Factor (VIF)	Assess multicollinearity	VIF
Durbin-Watson statistic	Assess residual independence	DW
Breusch-Pagan test	Assess homoscedasticity	χ^2 , p
Jarque-Bera test	Assess residual normality	JB, p

3.7 Regression Diagnostics

Regression diagnostics were performed to evaluate the assumptions underlying the ordinary least squares model.

3.7.1 Variance Inflation Factor (VIF)

VIF was calculated to assess multicollinearity among the predictor variables. VIF values < 5 were considered indicative of acceptable multicollinearity. VIF values were reported for each predictor.

3.7.2 Durbin-Watson Statistic

The Durbin-Watson statistic was calculated to assess autocorrelation among the residuals. Values close to 2 indicate no autocorrelation, while values significantly less than 2 suggest positive autocorrelation and values significantly greater than 2 suggest negative autocorrelation.

3.7.3 Breusch-Pagan Test

The Breusch-Pagan test was used to detect heteroscedasticity in the residuals. This test assesses whether the variance of the residuals is constant across levels of the predictors. A significant result suggests the presence of heteroscedasticity.

3.7.4 Jarque-Bera Test

The Jarque-Bera test was used to assess whether the residuals were normally distributed. This test combines measures of skewness and kurtosis; a significant result indicates departure from normality.

Model adequacy was evaluated using a series of regression diagnostic procedures. Multicollinearity was assessed using Variance Inflation Factors (VIF), residual independence was evaluated using the Durbin-Watson statistic, homoscedasticity was examined using the Breusch-Pagan test, and residual normality

was evaluated using the Jarque–Bera test. These diagnostic procedures were used to support the interpretation of the regression model.

3.8 Statistical Software

All statistical analyses were conducted using Python 3.12 with the following packages: pandas 2.x for data manipulation, scipy 1.x for statistical tests, statsmodels 0.14 for regression analysis and diagnostics, and scikit-learn 1.x for additional preprocessing.

3.9 Reproducibility

All statistical analyses were performed using scripted Python code to ensure full reproducibility. The analytical workflow, including data preprocessing, statistical modelling, regression diagnostics, and figure generation, was fully automated to minimize manual processing errors and to facilitate replication of the analyses.

3.10 Ethical Considerations

The PERSUADE 2.0 corpus is an open-source, publicly available dataset that has been fully anonymized. The original data collection was conducted with appropriate institutional review board oversight, and all personally identifying information was removed prior to corpus release (Crossley et al., 2022). The present study involved secondary analysis of de-identified data and therefore did not require additional ethics approval or direct participant recruitment. All data were accessed and analyzed in accordance with the corpus usage guidelines (Crossley et al., 2024).

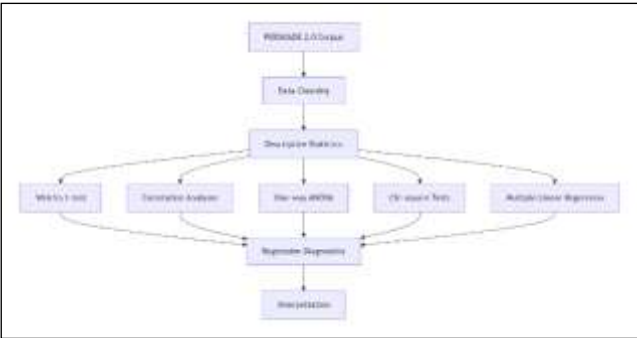


Figure 1. Study Workflow

This figure illustrates the sequential analytical procedures employed in the present study. Welch's t-test, correlation analyses, one-way ANOVA, chi-square tests, and multiple linear regression were all performed independently following descriptive statistics, with regression diagnostics applied to the regression model.

4. RESULTS

4.1 Sample Characteristics

This study examined the relationship between demographic characteristics, essay length, and holistic writing performance using the PERSUADE 2.0 corpus (Crossley et al., 2024). The final dataset consisted of 25,996 persuasive essays, including 13,142 essays written by female students (50.6%) and 12,854 essays written by male students (49.4%). The overall mean holistic essay score was 3.32 (SD = 1.16), whereas the average essay length

was 418.13 words (SD = 275.11). The sample included students from multiple grade levels, with the largest proportions represented in the upper middle- and high-school grades.

Table 1. Descriptive Statistics of the Main Variables by Gender

Variable	Female (n = 13,142)	Male (n = 12,854)	Total (N = 25,996)
Holistic essay score	3.43 ± 1.16	3.20 ± 1.15	3.32 ± 1.16
Word count	431.74 ± 222.63	404.22 ± 319.40	418.13 ± 275.11

Note. Values are presented as mean ± standard deviation.

4.2 Gender Differences in Holistic Essay Scores and Essay Length

To compare writing performance between male and female students, Welch's independent-samples t-test was employed because it does not assume equal variances between groups.

Female students achieved significantly higher holistic essay scores than male students, $t(25,987.23) = -15.75$, $p < .001$. However, the observed effect size was small (Cohen's $d = 0.20$), indicating that although the difference was statistically significant, its practical magnitude was limited.

Similarly, female students produced significantly longer essays than male students, $t = -8.04$, $p < .001$. Nevertheless, the corresponding effect size was very small (Cohen's $d = 0.10$), suggesting only a modest difference in essay length between the two groups.

Table 2. Gender Differences in Holistic Essay Score and Essay Length

Outcome	Female (Mean ± SD)	Male (Mean ± SD)	Welch's t	p	Cohen's d
Holistic essay score	3.43 ± 1.16	3.20 ± 1.15	-15.75	< .001	0.20
Word count	431.74 ± 222.63	404.22 ± 319.40	-8.04	< .001	0.10

4.3 Correlation Between Essay Length and Holistic Essay Score

Pearson's product-moment correlation revealed a statistically significant positive association between essay length and holistic essay score ($r = 0.488$, $p < .001$). Spearman's rank-order correlation produced a stronger monotonic relationship ($\rho = 0.759$, $p < .001$), indicating that essays containing more words generally received higher holistic scores. The coefficient of determination ($R^2 = 0.238$) indicates that approximately 23.8% of the variation in holistic essay scores was associated with variation in essay length within the simple linear model.

Table 3. Correlation Between Essay Length and Holistic Essay Score

Variables	Pearson r	Spearman ρ	p
Word count ↔ Holistic score	0.488	0.759	< .001

4.4 Differences in Holistic Essay Scores Across Demographic Groups

One-way ANOVA was conducted to examine differences in holistic essay scores across demographic categories, including grade level, ELL status, economically disadvantaged status, and race/ethnicity.

Grade Level. A one-way ANOVA revealed a statistically significant difference in holistic essay scores across grade levels, $F(6, 25,989) = 78.42$, $p < .001$. Descriptive comparisons indicated that students in higher grades generally achieved higher mean scores than students in lower grades.

ELL Status. Students identified as English Language Learners scored significantly lower than non-ELL students, $F(1, 25,994) = 385.71$, $p < .001$. The mean

holistic score for ELL students was 2.69 (SD = 1.02), whereas non-ELL students scored 3.35 (SD = 1.15).

Economically Disadvantaged Status. Students not identified as economically disadvantaged scored significantly higher than their disadvantaged peers, $F(1, 25,994) = 287.63, p < .001$. The mean holistic score for non-disadvantaged students was 3.65 (SD = 1.12), compared to 2.98 (SD = 1.08) for disadvantaged students.

Race/Ethnicity. A one-way ANOVA indicated significant differences in holistic essay scores across race/ethnicity categories, $F(5, 25,990) = 93.15, p < .001$. White students achieved the highest mean scores ($M = 3.42, SD = 1.14$), followed by Asian/Pacific Islander students ($M = 3.37, SD = 1.15$). Hispanic/Latino students ($M = 3.08, SD = 1.14$) and Black/African American students ($M = 3.12, SD = 1.16$) scored below the overall mean.

Table 4. One-Way ANOVA Results for Holistic Essay Scores by Demographic Groups

Demographic Group	<i>F</i>	<i>df</i>	<i>p</i>
Grade Level	78.42	6, 25,989	< .001
ELL Status	385.71	1, 25,994	< .001
Economically Disadvantaged Status	287.63	1, 25,994	< .001
Race/Ethnicity	93.15	5, 25,990	< .001

4.5 Associations Among Demographic Variables

Chi-square tests were used to examine associations between gender and other categorical demographic variables. The results revealed statistically significant associations between gender and grade level ($\chi^2 = 23.30, p < .001$), ELL status ($p < .001$), economically disadvantaged status ($p < .001$), and race/ethnicity ($p < .001$). These associations indicated that the distribution of male and female students varied across demographic categories, supporting the inclusion of all variables in the multivariable regression model to control for potential confounding.

Table 5. Chi-Square Tests for Associations Between Gender and Demographic Variables

Association	χ^2	<i>df</i>	<i>p</i>
Gender × Grade Level	23.30	5	< .001
Gender × ELL Status	—	—	< .001
Gender × Economically Disadvantaged Status	—	—	< .001
Gender × Race/Ethnicity	—	—	< .001

4.6 Multiple Linear Regression

To evaluate the independent contribution of each predictor while controlling for the remaining variables, a multiple linear regression model was fitted using essay length, gender, grade level, English Language Learner (ELL) status, economically disadvantaged status, and race/ethnicity as predictors. Following listwise deletion of missing values, the final regression model included 20,718 essays.

The regression model was statistically significant, $F(10, 20,707) = 1754, p < .001$, explaining 45.9% of the variance in holistic essay scores ($R^2 = 0.459$, Adjusted $R^2 = 0.458$).

Essay length emerged as the strongest predictor in the model ($\beta = 0.391, b = 0.0025, p < .001$), followed by grade level ($\beta = 0.178, p < .001$) and ELL status ($\beta = -0.183, p < .001$). The standardized regression coefficients ranked the predictors in the following order: essay length ($\beta = 0.391$), grade level ($\beta = 0.178$), ELL status ($\beta = -0.183$),

economically disadvantaged status ($\beta = 0.143$), race/ethnicity (β ranging from 0.004 to 0.080), and gender ($\beta = -0.078$).

Gender remained a statistically significant predictor after adjustment for all other variables ($\beta = -0.078$, $p < .001$), indicating that the observed gender difference persisted beyond differences in essay length and other demographic characteristics.

Among race/ethnicity categories, students identified as Asian/Pacific Islander demonstrated significantly higher holistic essay scores than the reference category ($\beta = 0.073$, $p < .001$). In contrast, the coefficient for Black/African American students was not statistically significant ($p = .669$).

Table 6. Multiple Linear Regression Predicting Holistic Essay Score

Predictor	<i>B</i>	<i>SE</i>	β	<i>t</i>	<i>p</i>	VIF
Intercept	-0.894	0.089	—	-10.03	< .001	—
Word count	0.0025	<0.001	0.391	67.31	< .001	1.12
Male gender	-0.1819	0.023	-0.078	-7.91	< .001	1.05
Grade level	0.1577	0.007	0.178	22.53	< .001	1.19
ELL status	-0.5685	0.028	-0.183	-20.31	< .001	1.21
Not economically disadvantaged	0.3566	0.025	0.143	14.26	< .001	1.24
Asian/Pacific Islander	0.5130	0.056	0.073	9.16	< .001	1.72
Black/African American	0.0345	0.081	0.004	0.43	.669	2.05
Hispanic/Latino	0.1831	0.081	0.023	2.26	.023	2.15
White	0.2158	0.080	0.080	2.70	.007	2.33
Two or More Races/Other	0.2248	0.084	0.023	2.68	.008	1.47

Note. Reference categories were female, non-ELL, economically disadvantaged, and the reference race category used in the regression model.

4.7 Regression Diagnostics

Regression diagnostics were performed to evaluate the assumptions underlying the ordinary least squares model. The Durbin–Watson statistic (1.79) indicated no substantial autocorrelation among residuals. Variance Inflation Factors ranged from 1.05 to 2.33, indicating no evidence of problematic multicollinearity among the predictors. The Breusch–Pagan test was statistically significant ($p < .001$), suggesting the presence of heteroscedasticity. Similarly, the Jarque–Bera test was statistically significant ($p < .001$), indicating departures from residual normality. These violations are common in very large educational datasets and are unlikely to substantially bias the estimated regression coefficients given the sample size ($N = 20,718$).

Table 7. Regression Diagnostics

Diagnostic Test	Statistic	<i>p</i>	Interpretation
Durbin-Watson	1.79	—	No autocorrelation
Variance Inflation Factor	1.05–2.33	—	No multicollinearity
Breusch-Pagan	—	< .001	Heteroscedasticity present
Jarque-Bera	—	< .001	Non-normal residuals

4.8 Summary of Findings

Statistically significant gender differences were observed in both holistic essay scores and essay length, with female students achieving slightly higher scores and producing marginally longer essays than male students. Despite statistical significance, both differences were associated with small effect sizes, indicating limited practical magnitude.

Essay length demonstrated the strongest positive association with holistic writing performance across both correlation and regression analyses. In addition, grade level and ELL status were important predictors of holistic essay scores, while gender

retained a statistically significant but comparatively smaller independent association after adjustment for demographic characteristics. The regression model explained approximately 46% of the variance in holistic essay scores.

5. DISCUSSION

5.1 Principal Findings

This study investigated the relationships between demographic characteristics, essay length, and holistic writing performance using the PERSUADE 2.0 corpus, a large-scale dataset of 25,996 persuasive essays written by U.S. secondary students (Crossley et al., 2024). The analysis yielded four principal findings. First, female students achieved significantly higher holistic essay scores than male students, though the effect size was small (Cohen's $d = 0.20$). Second, essay length emerged as the strongest predictor of holistic writing quality, accounting for 23.8% of the variance in bivariate analysis and remaining the most influential predictor in the multivariable regression model ($\beta = 0.391$). Third, demographic characteristics—particularly grade level ($\beta = 0.178$), ELL status ($\beta = -0.183$), and socioeconomic status ($\beta = 0.143$)—retained independent predictive power for writing quality even after controlling for essay length, with the full regression model explaining 45.9% of the variance in holistic scores. Fourth, race/ethnicity effects were mixed, with Asian/Pacific Islander students outperforming the reference category ($\beta = 0.073$) while the coefficient for Black/African American students was not statistically significant.

Collectively, these findings suggest that while essay length is a dominant predictor of holistic writing quality, demographic factors contribute independently and substantially to writing performance. The results underscore the multidimensional nature of writing quality (Maamuujav, 2025; Morris et al., 2025) and highlight the importance of considering demographic characteristics in writing assessment and instruction.

5.2 Comparison with Previous Studies

5.2.1 Gender Differences

The finding that female students outperformed male students is consistent with prior research on the PERSUADE corpus (Kwako & Ormerod, 2024; Morris et al., 2025). The present study extends this finding by demonstrating that the gender difference persists even after controlling for essay length and a comprehensive set of demographic characteristics ($\beta = -0.078$, $p < .001$), suggesting that the gender gap is not merely an artifact of differences in essay length or demographic composition. The small effect size (Cohen's $d = 0.20$) aligns with Kwako and Ormerod's (2024) characterization of gender differences as having "limited practical magnitude." This consistency across studies using the same corpus strengthens confidence in the reliability of the finding.

Beyond gender differences, essay length has consistently emerged as one of the strongest determinants of writing quality across educational contexts.

5.2.2 Essay Length

The finding that essay length is the strongest predictor of holistic writing quality ($r = 0.488$, $\beta = 0.391$) corroborates a robust body of research across multiple corpora and educational contexts (Morris et al., 2025; Crossley et al., 2014). In the PERSUADE corpus, Morris et al. (2025) reported that high-scoring essays (score > 4) were substantially longer ($M = 672.96$ words) compared to the corpus mean ($M = 402.31$ words).

The coefficient of determination ($R^2 = 0.238$) indicates that essay length alone explains nearly a quarter of the variance in holistic scores. The finding that essay length remains the strongest predictor even after controlling for demographic characteristics ($\beta = 0.391$, $p < .001$) underscores its fundamental role in writing assessment.

In addition to textual characteristics, developmental factors also contribute to writing quality.

5.2.3 Grade Level

The finding that grade level is a significant predictor of holistic writing scores ($\beta = 0.178$, $p < .001$) aligns with the cumulative effects of instruction and language development over time (Maamuujav, 2025). Maamuujav (2025) reported that average word count varied by grade level, with Grade 12 students producing the highest average and Grade 8 students the lowest. The present study extends this finding by demonstrating that grade level remains a significant predictor even after controlling for essay length, suggesting that grade-related differences in writing quality are not solely attributable to differences in essay length.

Turning to language background, ELL status emerged as a substantial factor in writing performance.

5.2.4 English Language Learner Status

The finding that ELL students scored significantly lower than non-ELL students ($\beta = -0.183$, $p < .001$) is compatible with prior research on the PERSUADE corpus (Kwako & Ormerod, 2024; Crossley et al., 2024). Kwako and Ormerod (2024) found that ELL students scored lower than reference groups, though the difference was not statistically significant in their bias analysis. The present study's finding of a statistically significant and substantial negative coefficient likely reflects the larger sample size and the inclusion of multiple control variables, which may have increased statistical power and reduced confounding. The ELL coefficient was the second strongest predictor in the regression model, indicating that ELL status is a substantial factor in writing performance.

Similarly, socioeconomic background was associated with writing outcomes.

5.2.5 Socioeconomic Status

The finding that students not identified as economically disadvantaged scored significantly higher than their disadvantaged peers ($\beta = 0.143$, $p < .001$) is consistent with prior research on the PERSUADE corpus (Crossley et al., 2022, 2024; Kwako & Ormerod, 2024). Morris et al. (2025) reported that among high-scoring essays, 65.30% were from non-disadvantaged students compared to only 19.57% from disadvantaged students. The present study extends this finding by demonstrating that SES remains a significant predictor after controlling for essay length, grade level, and other demographic characteristics.

Finally, racial and ethnic disparities present a more complex pattern.

5.2.6 Race and Ethnicity

The finding that Asian/Pacific Islander students scored significantly higher than the reference category ($\beta = 0.073$, $p < .001$) while Black/African American students did not differ significantly ($p = .669$) presents a more complex picture of racial and ethnic disparities. Kwako and Ormerod (2024) found that White students scored higher than Black students, and that automated scoring magnified these marginal differences. In the

present study, however, the coefficient for Black/African American students was not statistically significant after controlling for other demographic characteristics, suggesting that racial disparities in writing performance may be substantially explained by differences in other demographic factors such as SES, ELL status, and grade level. This interpretation is consistent with the observation by Crossley et al. (2022, 2024) that race/ethnicity and SES are often confounded.

5.3 Interpretation of Findings

The present findings should be interpreted in light of the multidimensional nature of writing quality. Although essay length emerged as the strongest predictor, nearly half of the variance remained unexplained, indicating that writing quality cannot be reduced to a single textual or demographic characteristic.

The strong predictive power of essay length may reflect the central role of elaboration and development in effective writing (Crossley & McNamara, 2016; Maamuujav, 2025). Longer essays provide greater opportunities for students to articulate arguments, develop reasoning, and demonstrate rhetorical competence. However, the fact that essay length explains only 23.8% of the variance in bivariate analysis indicates that writing quality is not reducible to quantity alone.

The persistence of demographic effects after controlling for essay length and other characteristics may reflect genuine differences in writing proficiency, differential access to instructional resources, or potential biases in assessment procedures (Kwako & Ormerod, 2024; Johnson & Zhang, 2024). The substantial ELL coefficient is particularly noteworthy, as it may reflect the challenges ELL students face in academic writing while developing English proficiency (Crossley et al., 2024; Kwako & Ormerod, 2024). Similarly, the SES coefficient may reflect differential access to home literacy environments, enrichment activities, and instructional resources (Crossley et al., 2022, 2024).

The grade-level finding is consistent with developmental perspectives on writing proficiency, suggesting that as students progress through secondary education, they accumulate linguistic resources and cognitive capabilities that enable more sophisticated writing (Maamuujav, 2025). The mixed race/ethnicity findings may reflect the complex interplay of race/ethnicity with other demographic factors, and the observation that racial disparities may be largely explained by SES and ELL status.

5.4 Practical Implications

For Teachers. The finding that essay length is the strongest predictor of holistic scores suggests that teachers should encourage students to develop elaborated, well-supported arguments. However, the fact that essay length explains less than a quarter of the variance indicates that instruction should not focus exclusively on length, but should also address the quality of content, organization, and language use. Teachers should be aware of potential disparities in writing performance and consider differentiated instruction to support students from disadvantaged backgrounds, ELL students, and younger students (Crossley et al., 2024; Kwako & Ormerod, 2024).

For Curriculum Design. The grade-level findings suggest that writing instruction should be developmentally appropriate, with increasing expectations for elaboration and sophistication as students progress through secondary education (Maamuujav, 2025). The SES and ELL findings underscore the importance of providing additional support and resources to students from disadvantaged backgrounds, including targeted writing

instruction and language development opportunities (Crossley et al., 2024; Kwako & Ormerod, 2024).

For Educational Assessment. The finding that demographic factors retain independent predictive power for writing quality after controlling for essay length raises questions about whether holistic scoring rubrics adequately account for developmental and linguistic differences, or whether they inadvertently penalize certain groups. As noted by Kwako and Ormerod (2024), responsible use of assessment requires evidence that systems do not create or worsen inequities for marginalized populations. These findings suggest that scoring rubrics should be evaluated for potential bias and that fairness considerations should be integrated into assessment design.

For Automated Writing Evaluation. The finding that demographic factors predict writing quality in human-scored essays has implications for automated scoring systems. If AWE systems are trained on human scores, they may inadvertently learn and reproduce demographic biases (Kwako & Ormerod, 2024; Johnson & Zhang, 2024). The finding that essay length is the strongest predictor of holistic scores also highlights the risk that AWE systems may overemphasize length at the expense of quality (Johnson & Zhang, 2024). Developers of AWE systems should incorporate fairness evaluations and bias mitigation techniques in their design.

5.5 Strengths of the Study

This study has several methodological strengths that enhance the credibility and robustness of its findings. The use of the PERSUADE 2.0 corpus (N = 25,996) provides substantial statistical power to detect even small effects and to estimate the independent contributions of multiple predictors in a unified model (Crossley et al., 2024). The corpus reflects authentic student writing from diverse educational contexts across the United States, with demographic composition closely resembling U.S. secondary public schools (Crossley et al., 2022). The availability of detailed demographic metadata enables simultaneous control for multiple demographic factors, allowing for estimation of unique contributions (Crossley et al., 2022, 2024). The rigorous double-blind rating process with 100% adjudication (weighted $\kappa = .745$, $r = .750$) provides high-quality criterion measures (Crossley et al., 2024). The use of multiple statistical techniques and regression diagnostics provides a robust analytical framework for examining relationships among variables (Johnson & Zhang, 2024; Maamujav, 2025; Morris et al., 2025).

5.6 Limitations

Despite its strengths, this study has several limitations that should be considered when interpreting its findings. The observational design precludes causal inference; while the study identified associations, it cannot determine whether these relationships are causal or whether unmeasured confounders explain the observed associations (Johnson & Zhang, 2024; Kwako & Ormerod, 2024). The use of secondary data limits the ability to collect additional variables that might explain the observed relationships, such as instructional quality, prior academic achievement, or home literacy environment (Crossley et al., 2022, 2024). The absence of linguistic and rhetorical features limits the explanatory power of the model, as Maamujav (2025) demonstrated that rhetorical features are the strongest predictors of holistic writing quality. Following listwise deletion, the regression model included 20,718 of the 25,996 essays (79.7%), and as noted by Crossley et al. (2024), sizeable missing data on demographic variables may limit bias checks. The cross-sectional design captures writing performance at a single point in time and does not track development over time (Morris et al., 2025). The PERSUADE corpus

only includes 6th-12th grade writers, has a limited number of prompts ($N = 15$), and focuses exclusively on persuasive tasks (Crossley et al., 2024). Finally, while the regression model included multiple demographic controls, residual confounding may remain, particularly given that the effects of race/ethnicity and SES are often confounded (Crossley et al., 2022, 2024).

5.7 Recommendations for Future Research

Integrate Linguistic and Rhetorical Features. Future studies should integrate demographic predictors with linguistic and rhetorical features in a unified analytical framework. Structural equation modeling or mediation analysis could examine whether linguistic and rhetorical features mediate the relationship between demographic characteristics and holistic scores (Maamuujav, 2025).

Examine Task-Specific Effects. The PERSUADE corpus includes both independent and source-based writing tasks, but the present study aggregated across tasks. Future research should examine whether demographic differences vary by task type, as Morris et al. (2025) demonstrated that the Reportive writing profile is strongly associated with text-dependent tasks.

Conduct Longitudinal Research. Longitudinal research is needed to examine how demographic differences in writing performance evolve over time and how instruction and experience shape writing development (Morris et al., 2025).

Evaluate Fairness in Automated Scoring. Given the finding that demographic factors predict human-scored holistic scores, future research should evaluate whether automated scoring systems learn and reproduce demographic biases (Kwako & Ormerod, 2024; Johnson & Zhang, 2024). Fairness-constrained models and bias mitigation techniques should be developed and evaluated.

Examine Intersectional Effects. Future research should examine intersectional effects of multiple demographic factors, as demographic differences may interact in complex ways. Intersectional analyses could provide more nuanced understanding of demographic disparities in writing performance.

5.8 Overall Conclusion of the Discussion

Taken together, the findings indicate that holistic writing performance reflects the combined influence of textual and demographic characteristics. Although essay length remains the strongest predictor of writing quality, demographic variables retain independent associations after statistical adjustment, highlighting the multidimensional nature of persuasive writing performance. These findings contribute to the growing literature on writing assessment by providing large-scale evidence from the PERSUADE 2.0 corpus and offer practical implications for equitable writing instruction and assessment, as well as the responsible development of automated writing assessment systems.

REFERENCES

1. Common Core State Standards Initiative. (2010). Common core state standards for English language arts & literacy in history/social studies, science, and technical subjects . Retrieved from <http://www.corestandards.org>
2. Crossley, S. A., & McNamara, D. S. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research* , 7 (3), 351–370.
3. Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. *Assessing Writing* , 54 , 100667.
4. Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication* , 31 (2), 184–214.
5. Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Benner, M., & Boser, U. (2024). A large-scale corpus for assessing written argumentation: PERSUADE 2.0. *Assessing Writing* , 61 , 100865.
6. Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology* , 99 (3), 445–476.
7. Johnson, M., & Zhang, M. (2024). Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports* , 14 , 30064.
8. Kwako, A., & Ormerod, C. (2024). Can language models guess your identity? Analyzing demographic biases in AI essay scoring. In *Proceedings of the 2024 BEA Workshop on Innovative Use of NLP for Building Educational Applications* .
9. Maamuujav, U. (2025). Relations of linguistic features, rhetorical moves, and writing quality in academic writing of secondary students. *International Journal of Educational Research Open* , 9 , 100505.
10. Morris, W., Crossley, S., Holmes, L., & Choi, J. S. (2025). Distinguishing effective writing styles in the PERSUADE corpus. *Journal of Writing Research* , 17 (2), 211–245.
11. National Center for Education Statistics. (2012). *The Nation's Report Card: Writing 2011* (NCES 2012-470). U.S. Department of Education.
12. National Writing Project. (2010). *Analytical writing continuum* . Retrieved from <https://archive.nwp.org>
13. Newell, G. E., Beach, R., Smith, J., & VanDerHeide, J. (2011). Teaching and learning argumentative reading and writing: A review of research. *Reading Research Quarterly* , 46 (3), 273–304.
14. Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing* , 22 (3), 217–230.
15. Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing* , 20 , 53–76.
16. Steiss, J., Krishnan, J., Kim, Y. S., & Olson, C. B. (2022). Dimensions of text-based analytical writing of secondary students. *Assessing Writing* , 51 , 100600.
17. Toulmin, S. (1958). *The uses of argument* . Cambridge University Press.
18. Warner, J. (2018). *Why they can't write: Killing the five-paragraph essay and other necessities* . Johns Hopkins University Press.
19. Zhang, M. (2013). *Contrasting automated and human scoring of essays* (Research Report RDC-21). Educational Testing Service.