Received: 15 May 2024 | Accepted: 21 April 2025

DOI: 10.1111/bjet.13596

ORIGINAL ARTICLE

British Journal of Educational Technology BERA

NLP-enabled automated assessment of scientific explanations: Towards eliminating linguistic discrimination

ChanMin Kim[1] | **Rebecca J. Passonneau**[2] | **Eunseo Lee**[3] | **Mahsa Sheikhi Karizaki**[2] | **Dana Gnesdilow**[4] | **Sadhana Puntambekar**[4,5]

[1]Learning, Design, and Technology, The Pennsylvania State University, University Park, Pennsylvania, USA

[2]Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania, USA

[3]Educational Psychology, The Pennsylvania State University, University Park, Pennsylvania, USA

[4]Interactive Learning and Design Lab, The University of Wisconsin, Madison, Wisconsin, USA

[5]Educational Psychology, The University of Wisconsin, Madison, Wisconsin, USA

Correspondence
ChanMin Kim, Learning, Design, and Technology, The Pennsylvania State University, 314D Keller Building, University Park, PA 16802, USA.
Email: cmk604@psu.edu

Funding information
National Science Foundation, Grant/Award Number: 2010351 and 2010483

Abstract

As use of artificial intelligence (AI) has increased, concerns about AI bias and discrimination have been growing. This paper discusses an application called PyrEval in which natural language processing (NLP) was used to automate assessment and provide feedback on middle school science writing without linguistic discrimination. Linguistic discrimination in this study was operationalized as unfair assessment of scientific essays based on writing features that are not considered normative such as subject-verb disagreement. Such unfair assessment is especially problematic when the purpose of assessment is not assessing English writing but rather assessing the content of scientific explanations. PyrEval was implemented in middle school science classrooms. Students explained their roller coaster design by stating relationships among such science concepts as potential energy, kinetic energy and law of conservation of energy. Initial and revised versions of scientific essays written by 307 eighth-grade students were analyzed. Our manual and NLP assessment comparison analysis showed that PyrEval did not penalize student essays that contained non-normative writing features. Repeated measures ANOVAs and GLMM analysis results revealed that essay quality significantly improved from initial to revised essays

after receiving the NLP feedback, regardless of non-normative writing features. Findings and implications are discussed.

KEYWORDS
artificial intelligence, natural language processing, automated writing assessment, codesign, middle school classrooms, science writing, linguistic discrimination

Practitioner notes

What is already known about this topic

- Advancement in AI has created a variety of opportunities in education, including automated assessment, but AI is not bias-free.
- Automated writing assessment designed to improve students' scientific explanations has been studied.
- While limited, some studies reported biased performance of automated writing assessment tools, but without looking into actual linguistic features about which the tools may have discriminated.

What this paper adds

- This study conducted an actual examination of non-normative linguistic features in essays written by middle school students to uncover how our NLP tool called PyrEval worked to assess them.
- PyrEval did not penalize essays containing non-normative linguistic features.
- Regardless of non-normative linguistic features, students' essay quality scores significantly improved from initial to revised essays after receiving feedback from PyrEval. Essay quality improvement was observed regardless of students' prior knowledge, school district and teacher variables.

Implications for practice and/or policy

- This paper inspires practitioners to attend to linguistic discrimination (re)produced by AI.
- This paper offers possibilities of using PyrEval as a reflection tool, to which human assessors compare their assessment and discover implicit bias against non-normative linguistic features.
- PyrEval is available for use on github.com/psunlpgroup/PyrEvalv2.

INTRODUCTION

Writing science explanations is a core science practice (Duschl & Osborne, 2002; Sandoval & Millwood, 2005). Engaging students in writing explanations in science is a topic of interest to educators and researchers, as is the timely assessment of written explanations (eg, Krajcik & McNeill, 2015; McNeill & Berland, 2017). Especially with regard to assessments in science classrooms, there have been unanswered inquiries about inequity and unfairness to students who are from non-normative cultures (eg, Lee, 2005; Lyon et al., 2012). For

example, Barton and Tan (2009) noted that "the exclusive nature of school science culture with its own ways of doing, speaking, and being [is] sometimes in conflict with the ways of being of students from nonnormative cultures" (p. 51). While much effort, such as teacher education for culturally and linguistically responsive pedagogy (eg, Rutt & Mumba, 2022), has been made to foster inclusion in science classrooms, linguistic discrimination in writing assessments has been reported in numerous studies (eg, Faulkner-Bond & Sireci, 2015; Jank, 2017; Johnson & VanBrackle, 2012; Mahboob & Szenes, 2010).

Linguistic discrimination in this study is operationalized as unfair assessment of written explanations based on features that are not considered normative. For example, there are two incidents of subject-verb disagreement [italicized] in the following sentence: When the roller coaster cart *go* down to the hill, the kinetic energy *begin* to increase. Such non-normative features are surface-level features that do not hinder communication about the content of explanations, but empirical studies have shown that readers assign lower scores to writing with non-normative grammar than to writing with normative grammar (Appelman & Schmierbach, 2018; Johnson & VanBrackle, 2012). For example, non-normative grammatical features have emerged in English used by many members of African American communities; such non-normative grammar is often perceived as inadequate despite decades of efforts in promoting critical language awareness (eg, Alim, 2010; Blodgett et al., 2020; Rosa & Burdick, 2017). Human assessors of state-mandated writing exams failed high or intermediate quality essays containing non-normative features of African American English (AAE) more than the same quality essays containing non-normative features of non-AAE (Johnson & VanBrackle, 2012). In another study, regardless of types of non-normative features, human assessors rated informativeness, credibility and quality of news articles lower when containing a large number of non-normative grammatical features (Appelman & Schmierbach, 2018). Science ability of people with lower normative English proficiency, including that of English learners in the US, is often devalued (Amano et al., 2021; Lee, 2005; Lynch et al., 2021; Peters, 2023). Unfair assessment of student scientific explanations that contain non-normative features is problematic when the purpose of assessment is not assessing English writing quality, but rather assessing scientific explanation ability. The premise of the present study is that feedback regarding scientific explanations should be grounded in fair assessment focused on science, because as Amano et al. (2021) argues, "Less fluent language skills do not equate to poorer quality of science" (p. 1121).

With drastic advancement in AI, one can imagine that machines can support fair assessment of writing. The literature reports optimistic expectations that AI could remove human bias against students' race, gender and other backgrounds because AI treats every student equally and consistently (Qin et al., 2020). While this imagined trust is not totally unreasonable, AI is not bias-free (Bearman & Ajjawi, 2023). Automated assessment of student writing is biased when data used to train the algorithms are not representative of the students to be served and also when assessment criteria are developed by mainstream groups (Dixon-Román et al., 2020). AI is not bias-free even when AI assessment is more consistent across varying text than that of human scorers (Wilson et al., 2024; Zhai et al., 2020). For example, if an algorithm is developed based on a norm that multiple non-normative features make scientific writing un-understandable (Peters, 2023), AI would assign a low score to writing that contains more than one non-normative feature. In this case, the AI system would not be a valid assessment tool in science classrooms in which the focus is on what students write about science rather than how they express it. In a recent study, bias was observed in automated assessment of scientific explanations written by English learners (Wilson et al., 2024). Notwithstanding the existence of numerous studies using AI to automate writing assessments, there is a dearth of studies focused on linguistic discrimination practiced in AI-enabled automated writing assessment.

RESEARCH QUESTIONS

Our *ultimate* goal is to eliminate bias in automated assessment of scientific explanations. In this paper, we discuss use of an application called PyrEval (Karizaki et al., 2024; Singh et al., 2022) in which natural language processing (NLP) was used to automate content assessment and provide formative feedback on middle school science writing without linguistic discrimination. In these essays, students explained their roller coaster design using such science concepts as potential energy, kinetic energy and the law of conservation of energy. Student essays were analyzed to examine if automated assessment, provided as per science concepts and their relations regardless of non-normative writing features, helped to improve essay quality. We hypothesized that, given fair content assessment and feedback centered on science concepts and relations, changes in essay quality would not be different between students whose essays contained non-normative writing features and students whose essays contained no non-normative features. The following research questions guided our study.

1. Does PyrEval assess essays regardless of non-normative writing features or does it penalize essays that contain non-normative features?
2. Is there any difference in essay quality change between students whose essays contained non-normative writing features and students whose essays did not contain non-normative writing features?
3. Does the type and number of non-normative writing features in essays predict essay quality change?

CONCEPTUAL FRAMEWORK OF THE STUDY

This study is grounded in literature that (a) reports AI-enabled automated writing assessment and its linguistic discrimination (eg, Blodgett et al., 2020; Dixon-Román et al., 2020; Goldshtein et al., 2024; Litman et al., 2021; Wilson et al., 2024), (b) argues against linguistic discrimination rooted in inequitable ideologies of languagelessness (eg, Alim, 2010; Amano et al., 2021; Lynch et al., 2021; Rosa, 2016, 2019; Rosa & Burdick, 2017), (c) calls for culturally and linguistically congruent writing assessments (eg, Huang, 2009; Johnson & VanBrackle, 2012; Lee, 2005) and (d) attempts to establish the potential of AI in addressing inequity (eg, Abdilla et al., 2020; Lin et al., 2021; Sumner, 2018).

AI-enabled automated writing assessment and linguistic discrimination

Automated writing assessment research has been rapidly increasing in recent years with the rise of AI. A wide spectrum of approaches to automated writing assessment has been developed and implemented in educational contexts for various purposes such as incorporating writing feature measures including scores for writing organization (Boulanger & Kumar, 2020), use of latent semantic analysis-based scoring for open-ended short answer responses to creativity testing items (LaVoie et al., 2020), human-centric scoring for formative feedback on ethical reasoning (Lee et al., 2023), incorporating Grammarly to detect errors (Almusharraf & Alotaibi, 2023), and use of semantic similarity tools to evaluate English learners' fact-based writing (Wang, 2022). A variety of automated assessment methods have been used also in science writing. For example, Gerard et al. (2019) used c-raterML to automatically score scientific explanations of sixth graders and guide their collaborative

revisions. Gerard and Linn (2022) designed the Annotator to support integrated revisions. Personalized methods of automated writing assessment were also developed to support students with low prior knowledge (Tansomboon et al., 2017). Transparency in automated writing assessment and feedback has been also studied to improve students' trust in and use of automated assessment (Conijn et al., 2023; Edelblut, 2020; Lee et al., 2019; Tansomboon et al., 2017). Integration of generative AI such as ChatGPT in automated assessment has been growing (eg, Escalante et al., 2023).

While numerous studies advanced automated writing assessment, there is a growing concern about linguistic discrimination of AI tools. For example, based on their review of 146 articles related to bias in NLP, Blodgett et al. (2020) argued for critical needs in NLP research and development that center lived experiences of linguistically minoritized groups. Dixon-Román et al. (2020) documented the potential failure of AI automated writing assessment to acknowledge literacies of minoritized groups. Recently, more researchers voiced their realization that automated writing assessment should be equitable without discrimination, for example, “based on dialect, language background, race, ethnicity, gender, or other demographic variables discernable via writing” (Goldshtein et al., 2024, p. 422).

Linguistic discrimination of AI against minoritized groups pertains not only to their writing but also to their speech. Recent studies consistently reported poor performance of speech recognition AI on African American English, English learners, and other nonmainstream English use (Brandt & Hazel, 2024; Cunningham et al., 2024; Jeon et al., 2024; Martin & Wright, 2023; Ngueajio & Washington, 2022). Fundamental causes for such discriminatory development are grounded in a lack of sociolinguistic understanding and consideration of minoritized linguistic features (Martin & Wright, 2023). Along these lines, linguistic dominance perceiving "superiority of *standardized* [emphasis added] US English" is criticized for AI creating linguistic oppression to non-normative English users (Payne et al., 2024, p. 553).

While limited, a few empirical studies specifically investigated the bias of automated writing assessment rather than student learning through assessment. For example, Litman et al. (2021) examined whether their automated writing assessment models did a disservice to African Americans, males or economically disadvantaged students, and they found small but significant bias. Yang et al. (2024) tested nine existing automated writing assessment methods and found that carefully engineered traditional machine learning models had less bias against specific genders, English learners, or economically disadvantaged students than neural network models. Wilson et al. (2024) developed automated writing assessment models with a high level of agreement with human assessors but also found bias against English learners (ie, more lenient on writing of non-English learners) as observed in human assessors' bias against English learners.

In recent years, some effort has begun to mitigate linguistic discrimination in automated writing assessment. Correnti et al. (2022) noted their use of representative data in training their automated writing assessment system to prevent bias against certain groups of students, but no related data analysis or discussion was presented. Litman et al. (2021) implemented a variety of bias mitigation strategies and also reported trade-offs such as reduced reliability when bias was reduced. While these studies exemplify effort in minimizing linguistic discrimination, little research has examined actual linguistic features against/for which AI tools may have discriminated. Demographic information was mostly used in these studies. The present study inspected non-normative linguistic features in student science writing to investigate whether our NLP tool disadvantaged students whose essays contained such features.

Linguistic discrimination from inequitable ideologies of *languagelessness*

Unfair assessment of scientific essays containing non-normative writing features is problematic pedagogically as noted above, but it is also problematic in that it perpetuates languagelessness that devalues linguistic diversity across cultures. Languagelessness is a construct that refers to the process (and outcome) of assigning degraded value to the totality of a person, beyond assessing language proficiencies/capacities, because of (a) perceived lacks in their language proficiency, or (b) their stigmatized linguistic practices (Rosa, 2016). Languagelessness is shown in empirical evidence in which the ability of writers and scientists, and credibility of their knowledge, are demoted when non-normative linguistic features are noticed (Amano et al., 2021; Appelman & Schmierbach, 2018; Johnson & VanBrackle, 2012). This is problematic in science, because "ignoring linguistic diversity in science … can perpetuate hegemonic patterns of knowledge production" (Lynch et al., 2021, p. 270). In science classrooms, languagelessness could lead to improper scaffolding and even dehumanize the learning process. Culturally and linguistically congruent writing assessments are needed (eg, Huang, 2009; Johnson & VanBrackle, 2012; Lee, 2005) even more with rapid adoption of AI. As discussed above, linguistic discrimination of AI without sociolinguistic understanding of minoritized linguistic features creates linguistic oppression (Martin & Wright, 2023; Payne et al., 2024).

Nonetheless, considering the potential of AI in addressing inequity (eg, Abdilla et al., 2020; Lin et al., 2021; Sumner, 2018), automated assessment systems can play a role in resisting languagelessness and promoting equity by not being judgmental about non-normative linguistic features. While there is no AI that is culturally and linguistically congruent yet, our work suggests methods that are potentially useful in this regard.

METHOD

Setting and context

Our NLP-enabled automated content assessment tool, called PyrEval, was implemented in seven eighth grade science classrooms in two midwestern school districts in the United States. Students learned about height, mass, energy and the law of conservation of energy while conducting virtual roller coaster experiments. They then wrote essays about their roller coaster design with explanations of the science behind their design in their digital science notebook (see Figure 1).

PyrEval assessed student essays by detecting the presence or absence of six main ideas (see Table 1). According to PyrEval's assessment of each main idea, automated feedback was provided. The feedback listed (a) the main ideas that were detected in their essays with a checkmark and (b) the main ideas that were not detected with a question mark (see Figure 2). Students revised their essays after receiving feedback.

NLP-enabled automated assessment tool: PyrEval

PyrEval was originally developed by one of the researchers on this research project and her colleagues to assess short passages summarizing source content (Gao et al., 2019; Passonneau et al., 2018). PyrEval was grounded in the wise-crowd content assessment model in which models of important propositions are developed based on content summaries written by expert writers or proficient students (Passonneau et al., 2018). Once vector

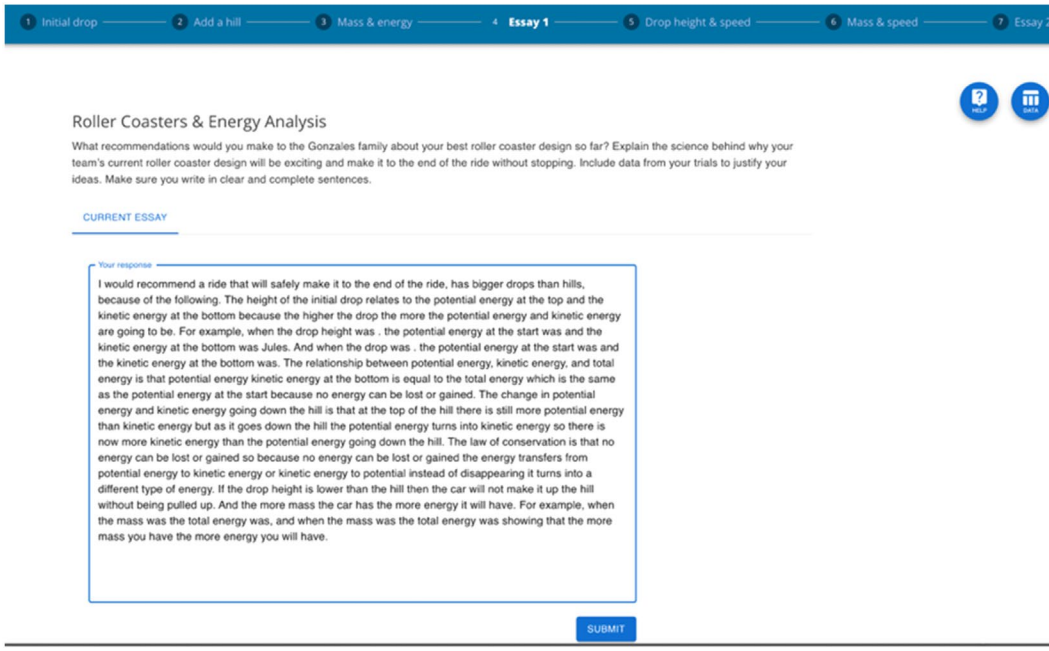

FIGURE 1 Essay submission interface screenshot.

TABLE 1 List of main ideas.

Content units (CUs)	Main ideas
CU0	1. There is an inverse relationship between potential energy and kinetic energy
CU1	2. The initial drop height should be higher than the hill height
CU2	3. The law of conservation of energy states that energy cannot be created or destroyed, only transformed
CU3	4. Greater mass equals greater (any kind of) energy
CU4	5. Total energy on the roller coaster ride remains the same if there is no friction
CU5	6. The greater height the greater the potential energy

representations of propositions are created, PyrEval creates a model of assigning a higher importance weight to ideas that a greater number of wise crowd members included. For example, in the present study, the six main ideas (Table 1) were identified as ideas with the highest weight (5) in the wise crowd model and they were also highly aligned with important science ideas students were expected to learn during the unit; as shown in Figure 3, Sentence 6 was assessed as a segment in which main idea 3 (the law of conservation of energy) was explained and labeled as Content Unit 2 expressed by five of the wise crowd members (ie, labeled as (1) through (5)).

This NLP method works with a small set of reference responses. This method also differs from many other automated writing assessments that focus on complete propositions. This method identifies similarity of meaning (ie, ideas; content units), rather than structure. How much of the important content students captured in their essays is assessed rather than how the content is expressed. This method is "independent of wording" (Gao et al., 2019, p. 404), which enables PyrEval to assess students' essays based on the *coverage of ideas* in

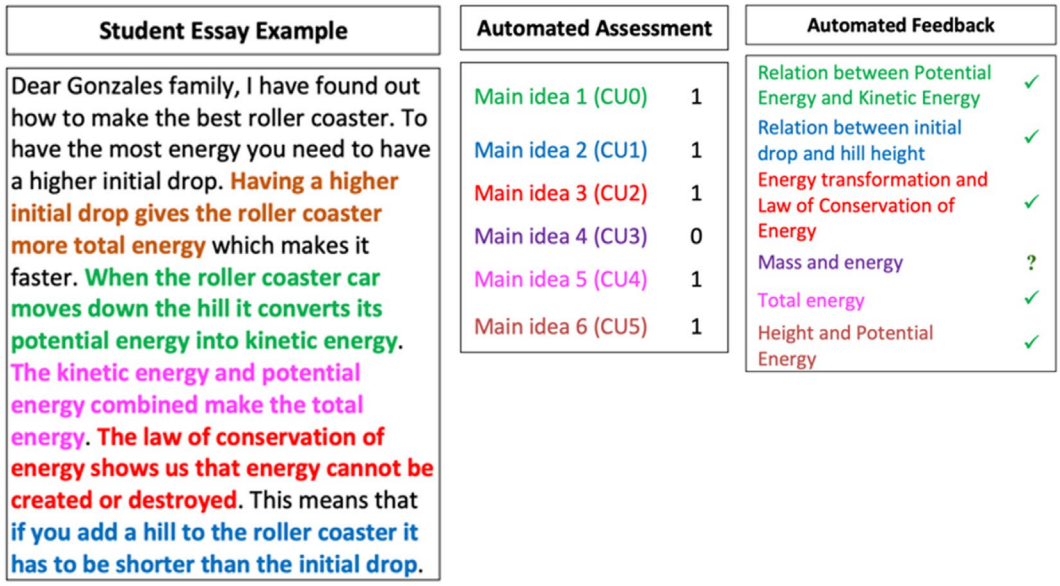

FIGURE 2 Sample essay, assessment and feedback. Font colors are used to illustrate the main ideas on which automated assessment and feedback focused. Feedback that students received did not include colored text (except for colors in check and question marks). Automated feedback also included PyrEval's accuracy level.

```
Sentence: 6 | Segmentation: 0

        Segment ID: 0 | Content Unit: 2 [Weight: 5]
        Segment: ..................... The Law of conservation and energy states that energy can be
                 ..................... created or destroyed but it can convert to other types of energy.
        Content Unit: ..............  (1) According to the law of conservation of energy, you can not
                      ..............      create energy, so we can not create more kinetic energy for
                      ..............      the car.
                      ..............  (2) This brings in total energy, because the law of conservation
                      ..............      states that energy cant be created nor destroyed.
                      ..............  (3) The Law of Conservation of Energy says that energy cant be
                      ..............      created nor destroyed,
                      ..............  (4) since the Law of Conservation of Energy states that energy can
                      ..............      not be created nor destroyed.
                      ..............  (5) The Law of conservation of Energy states that energy can be
                      ..............      transformed into other kinds of energy but can not be created
                      ..............      or destroyed.
```

FIGURE 3 Part of a log output that PyrEval generated showing the sentence assessed as one of the main ideas/content units.

their essays by comparing their propositions to the weighted content in the model. That is, PyrEval in the present study was designed to assess essays as per science concepts and their relations, regardless of non-normative writing features. For example, Figure 3 shows the wise crowd sample sentences (labeled as (1)–(5)) for Content Unit 2, related to the law of conservation of energy, that PyrEval used to identify whether the essay included information about the law of conservation of energy. The sentence (labeled as Segment) that PyrEval identified as Content Unit 2 includes non-normative English ("can convert to"). In other words, PyrEval assessed this segment as containing one of the main ideas regardless of how the main idea was expressed.

In adapting the NLP method for the present study, (a) historical essay data from one of the two participating school districts, (b) essay data from implementing the first version of

PyrEval and (c) interdisciplinary co-design were used to create wise crowd content models. Historical essay data were from middle school classrooms in which design-based science learning modules that involved scientific essay writing about roller coaster design were implemented. A portion of this historical data was modified to create reference essays as PyrEval was adapted for middle school student writing. We then used essay data from implementing the first version of PyrEval in two school districts. Interdisciplinary team members engaged with multiple rounds of modification in reference essays and wise crowd content models. Not only NLP experts and education researchers but also science teachers from both school districts participated in co-design of PyrEval and classroom implementations. Our PyrEval development is detailed elsewhere (Karizaki, et al., 2024; Singh et al., 2022).

Data sources

Scientific explanation essays

Initial and revised versions of scientific essays that 307 eighth-grade students wrote after their first three virtual experiments (Essay 1 in Figure 1) were used in this study. The initial version was revised after they received feedback from PyrEval.

Essay quality scores

Essay quality was measured by the sum of the scores PyrEval assigned according to the inclusion of the six main ideas/content units (CUs). When a CU was detected, 1 was assigned. When a CU was not detected, 0 was assigned. For example, 0 was assigned to the essay in Figure 2 for CU3, and 1 was assigned to the essay for CUs 0, 1, 2, 4 and 5. The essay quality score for this essay was 5.

Type of non-normative writing features

In this study, non-normative features were operationalized according to Johnson and VanBrackle's (2012) non-normative feature classifications of African American English (AAE) and English learners (EL). Appendix A lists the types of non-normative writing features, examples and frequencies. For example, among AAE features, deleted apostrophes were a frequent non-normative feature, particularly when students contracted a verb and *not* together. Among the EL features, the most frequent ones were related to choosing between two words (*then* in place of *than*) that had similar pronunciations. We included the combined type of AAE and EL non-normative features in this study (Table 2) because the study focus was to examine whether essays containing non-normative features were assessed regardless of non-normative features, rather than whether essays written by African American students or English learner students were assessed as fairly as essays written by students who are not African American or English learners. This focus is especially important considering that not all writing of African American students or English learners contains non-normative features.

Number of non-normative writing features

We reviewed the initial version of 307 students' essays one by one and manually coded each essay based on the *number* of non-normative features according to Appelman and

TABLE 2 Essay categorization according to non-normative writing features.

Coding criteria	Level	Frequency
Absence or presence of a non-normative feature	Absence	185
	Presence	122
Number of non-normative features	None	185
	Some	78
	Many	44
Type of non-normative features	No non-normative feature	185
	AAE type only	48
	EL type only	27
	Both AAE and EL types	47

Abbreviations: AAE refers to African American English; EL refers to English learners.

Schmierbach (2018) in which human assessors' judgment about the ability and knowledge of writers differed between *some* non-normative features versus *many* non-normative features. To classify *some* versus *many* non-normative features, we counted the number of non-normative features in each essay, including repeated ones of the same kind. We then calculated the mean ($M = 3.50$) and standard deviation ($SD = 2.67$) of the number of non-normative features in 122 essays that contained at least one non-normative feature. We coded essays containing non-normative features less than the mean (ie, 1 to 3) as essays with *some* non-normative features and essays containing more than the mean (ie, 4 and above) as essays with *many* non-normative features.

Prior physics knowledge

Students' prior physics knowledge was measured through a test assessing physics content knowledge related to the instructional unit in which PyrEval was implemented. The test consisted of 11 multiple choice items asking students to indicate accuracy of statements such as "The greater the height of the initial drop, the lower the potential energy at the top."

Data analysis methods

To address RQ1, two researchers performed manual coding together on all initial essays that had at least one non-normative feature ($n = 122$), which was 39.74% of the total 307 essays. First, we coded all sentences that contained non-normative writing features (see Appendix A). Second, we coded sentences that contained main ideas/content units (CUs; see Table 1 and Figure 2). Our manual coding criteria for this phase were the list of propositions for each CU along with sample sentences representing each CU that served as input in developing the wise crowd model for PyrEval (eg, Figure 3). We evaluated each sentence by comparing it against the list. For instance, CU1 was assessed using the following sample sentences:

- This suggests that the initial drop should be higher than the hill for the car to make it over the hill.
- The height of the initial drop was higher than the hills and loop afterwards.
- When you add a hill, you want the height of that to be less than the initial drop.
- In order for the car to reach back over the hill after the initial drop, the hill height has to be lower than the initial drop height.

- The hill cannot be higher than the initial drop.

The following sentences from student essays, for example, were identified to include a CU:

- "The hill height needs to be less than the height of the initial drop otherwise there will not be enough energy and it wont make it over the hill." (Essay 001)
- "If you want a hill, it needs to be lower than the initial drop." (Essay 023)
- "Any hills on this coaster need to be lower then the initial drop height for the car to go over." (Essay 088)
- "when you are deciding you hill height you need to make sure that it isnt greater than your drop height or the car will not make it up." (Essay 116)

Discrepancies occasionally arose between our manual assessment and NLP assessment in determining the most relevant sentence for each CU, which is discussed in detail in the results section below.

Figure 4 shows an example essay in which manual coding was done to underline the parts containing CUs and colored text containing non-normative features. That is, underlined text contains CUs and red text contains non-normative features.

Third, we examined each essay to see in what ways the underlined sentences that contained non-normative writing features (ie, the sentences containing both CUs and non-normative features) were assessed by PyrEval. We analyzed the log output of PyrEval for each essay. Each log output lists which CUs were matched to sentences (or segments where a sentence was segmented into two or more). For example, Figure 3 is part of a log output showing a sentence that contained non-normative grammar and was assessed by PyrEval as a sentence explaining CU2. We compared the NLP assessment results from each log output (eg, Figure 3) to the essay manual coding results (eg, Figure 4). We also examined cosine similarity values associated with CUs that PyrEval listed per sentence/segment.

To address RQ2, we ran a series of two-way repeated measures analysis of variance (ANOVAs) with essay quality scores as an outcome variable, time as a within-subjects measure and non-normative features in initial essays as a between-subjects factor to examine

The drop height for the roller coaster will be 5 meters. This is because the taller the initial drop height the more PE there will be. Having more PE means we will also have more total energy and KE. This is because energy cant be created or destroyed according to the Law of Conservation of Energy. For example when the initial height was 3 meters the PE at the top was 1466J with the KE being 0J. Then at the bottom the KE was 1456J and PE was 1. Next, we tried 4 meters and got 1955J of PE and 0J of KE at the top, then at the bottom 1954J of KE and 1J PE. Finally we did 5 meters and got a PE of 2443J at the initial drop and KE with 1J PE at the bottom and 2442J of KE. This shows like that I said, the more height there is the more energy there will be, making it so the roller coaster can make it until the end.

For the hill height we decided on 3 meters. This is because the hill height has to be less then the initial drop. It needs to be less then the initial drop because if its to high there wont be enough energy to get over the hill, as energy cant be created. When we tested 5.01 meters the cart was not able to go over. At the top of the hill the PE was 2448J, which is greater than the energy given at five meters, making it so the cart coldnot go over the hill. We decided to go with 3 meters because at the top of the hill the PE was 1471J and KE was 975J with a total of 2446J, making the cart able to go over the hill.

We also learned that the greater the cars mass the more PE and KE it would have. For example when we used 40kg for weight the PE was 1955J, at the initial drop, and the 1951J of KE at the bottom of the drop. On the other hand a 60kg cart had 2933J of PE at the initial drop, 2927J of KE at the bottom and the total energy being 2933J.

FIGURE 4 An example essay (Essay 003) manually coded to examine whether or not PyrEval actually identified main ideas/content units (CUs) regardless of non-normative writing features.

the main effects of time between initial and revised essays and the presence and number of non-normative features, as well as interaction effects.

To address RQ3, we ran a generalized linear mixed effect model (GLMM) to investigate how the type of non-normative features in students' initial essays explain essay quality scores in their revised essays. In the model, we included the fixed effects of time, prior physics knowledge score and the type of non-normative features in initial essays. We also included teacher and school district variables in the model (in the parentheses below) as random effects to control for the potential impact of teacher and school district variance. As shown in Appendix A, essays from school district 2 had more non-normative features. The model specification was as follows: essay quality score change ~ time + prior physics knowledge score + the type of non-normative features in initial essays + (teacher) + (school district).

RESULTS

Performance of PyrEval in assessing essays containing non-normative writing features

Manual coding and examination results indicated that PyrEval assessed student essays according to science concepts and relations (ie, main ideas; content units) regardless of non-normative writing features. Details are explained below.

In comparing the NLP assessment and manual assessment of the 122 initial essays that contained at least one non-normative writing feature, we found that PyrEval correctly identified the main ideas/content units (CUs) within the sentences containing both CUs and non-normative writing features 90% of the time. In fact, we found no discrepancies between PyrEval assessment and manual assessment on sentences/segments that contained both main ideas and non-normative features in 105 of the 122 initial essays (indicated as 100% in Appendix B). For example, in Essay 003 (Figure 5), PyrEval assessment and manual assessment both identified five green-highlighted sentences/segments as covering five CUs.

In 17 essays (indicated as <100% on Appendix B), we found discrepancies between the NLP assessment and manual assessment of CUs on sentences/segments that contained both main ideas and non-normative features.

For example, Essay 013 (indicated as 67% on Appendix B) on Figure 6 included three sentences/segments that were identified by manual assessment as covering three CUs but only two of them were identified by PyrEval. It should be noted that our comparison between manual assessment and the NLP assessment was centered on sentences/segments that contained non-normative grammar (see the data analysis methods section). We further investigated these 17 essays to see whether these discrepancies were related to the use of non-normative features in the writing or other factors. As a part of our examination of these discrepancies, we analyzed cosine similarity value outputs in addition to the log outputs that PyrEval generated. Cosine similarity value outputs list segments of each sentence that were segmented along with a cosine similarity value per CU. Cosine similarity value outputs also list sentences without segmentation along with a cosine similarity value per CU. When a cosine similarity value is lower than 0.5, it is listed as *nan* (ie, *n*ot *a n*umber) because 0.5 was set to be a threshold through extensive experiments prior to the present study. When cosine similarity values are lower than 0.5 for all six CUs, the sentence/segment is not listed on cosine similarity value outputs. In the case of Essay 013 on Figure 6, the manual assessment matched the yellow-highlighted sentence (Sentence 8) to CU1, but PyrEval assessment matched the green-highlighted sentence (Sentence 5) to CU1, as shown on Figure 7. Thus, we additionally examined cosine similarity value outputs for Essay 013 (Figure 8).

The drop height for the roller coaster will be 5 meters. This is because the taller the initial drop height the more PE there will be. Having more PE means we will also have more total energy and KE. This is because energy cant be created or destroyed according to the Law of Conservation of Energy. For example when the initial height was 3 meters the PE at the top was 1466J with the KE being 0J. Then at the bottom the KE was 1456J and PE was 1. Next, we tried 4 meters and got 1955J of PE and 0J of KE at the top, then at the bottom 1954J of KE and 1J PE. Finally we did 5 meters and got a PE of 2443J at the initial drop and KE with 1J PE at the bottom and 2442J of KE. This shows like that I said, the more height there is the more energy there will be, making it so the roller coaster can make it until the end.

For the hill height we decided on 3 meters. This is because the hill height has to be less then the initial drop. It needs to be less then the initial drop because if its to high there wont be enough energy to get over the hill, as energy cant be created. When we tested 5.01 meters the cart was not able to go over. At the top of the hill the PE was 2448J, which is greater than the energy given at five meters, making it so the cart coldnot go over the hill. We decided to go with 3 meters because at the top of the hill the PE was 1471J and KE was 975J with a total of 2446J, making the cart able to go over the hill.

We also learned that the greater the cars mass the more PE and KE it would have. For example when we used 40kg for weight the PE was 1955J, at the initial drop, and the 1951J of KE at the bottom of the drop. On the other hand a 60kg cart had 2933J of PE at the initial drop, 2927J of KE at the bottom and the total energy being 2933J.

FIGURE 5 An example essay (Essay 003) showing no discrepancy between PyrEval assessment results and manual assessment results.

A recommendation is to make the initial drop as high as you can and the hill as small as you can. This is because it well make the roller coaster faster. The coaster we have at the moment is very fast do to how high up the initial drop is compared to the hill. The height of the initial drop should be taller then the hills height because it will make the coaster fast. We know this because the coaster data shows that when at the top of the initial drop there is more PE then there is at the top of the hill. When we have a initial drop height of 5 meters the PE is 2440 and when you get to the top of the hill the PE is 984. When going down the drop the PE degrees and turns into KE then when going up the hill that same KE turns back into PE. The height of the initial drop has to be taller then the height of the hill. This is because the coaster car has friction and that friction makes the coaster car not have enough KE to get up the hill. The more mass you have on the coaster car the better. This is because when going down the drop or hill it is going to go fast do to the gravity acting on it is more because of the mass it is easier to pull it down the drops.

FIGURE 6 An example essay (Essay 013) showing a discrepancy in assessment results between the manual coding and PyrEval.

```
Sentence: 5 | Segmentation: 0

        Segment ID: 0 | Content Unit: 1 [Weight: 5]
        Segment: ..................... We know this because the coaster data shows that when at the top
                 ..................... of the initial drop there is more PE then there is at the top of
                 ..................... the hill.
        Content Unit: ................ (1) This suggests that the initial drop should be higher than the
                      ................       hill for the car to make it over the hill.
                      ................ (2) the height of the initial drop was higher than the hills and
                      ................       loop afterwards
                      ................ (3) When you add a hill you want the height of that to be less
                      ................       than the initial drop.
                      ................ (4) In order for the car to reach back over the hill after the
                      ................       initial drop, the hill height has to be lower than the initial
                      ................       drop height.
                      ................ (5) The hill can not be higher than the initial drop,
```

FIGURE 7 Part of a log output for Essay 013 that PyrEval generated showing Sentence 5 assessed as content unit 1 (CU1).

Among the 17 essays in which PyrEval did not identify a CU in the sentences/segments that manual assessment did, we found that in three of the essays, the NLP tool matched the sentence/segment that had a higher cosine similarity value to a CU, instead of the sentence/

Content Unit	0	1	2	3	4	5
15&5&0&0	nan	0.6809090811433610	nan	nan	nan	nan
15&5&1&0	nan	0.6763941672494770	nan	nan	nan	0.510604426603533
15&5&2&0	nan	0.6922087407053340	nan	nan	nan	nan
15&5&3&1	nan	0.6763941672494770	nan	nan	nan	0.510604426603533
15&8&0&0	nan	0.784338880729767	nan	nan	nan	0.5153981506199200

FIGURE 8 Part of cosine similarity value outputs for Essay 013 showing Sentence 5 (in green) and Sentence 8 (in yellow).

segment(s) that had a lower cosine similarity value. This pattern (Pattern 1) is explained by the fact that matching sentences/segments to a CU relies on cosine similarity values (Singh et al., 2022). Similarity of meaning that is closer to 1, not 0, is generally matched to a CU. However, in 13 of 17 essays, we found a reversed pattern (Pattern 2) in which the sentence/segment that had a lower cosine similarity value was matched to a CU, as in Essay 013 above (Figure 8). None of these patterns was related to the use of non-normative writing features, as detailed below. That is, PyrEval did not penalize these essays for containing non-normative features. Two sub-patterns are described below that explain why the sentence/segment with a higher cosine similarity value was not matched to a CU.

Pattern 2A

When there were more than one sentence/segment that covered a CU, PyrEval matched the sentence/segment that had a lower number of cosine similarity values for CUs instead of the sentence/segment(s) that had a higher number of cosine similarity values for CUs. For example, in Essay 013 above, PyrEval matched Sentence 5 (in green) to CU1 instead of Sentence 8 (in yellow) that we identified during manual assessment. As shown in Figure 8, Sentence 5 was listed with only one cosine similarity value whereas Sentence 8 was listed with two cosine similarity values. That is, Sentence 5 had one significant cosine similarity value (above the threshold, 0.5) which was with CU1. Sentence 8 had two significant cosine similarity values, which were with CU1 and CU5. This means Sentence 5 was clearer on CU1 than Sentence 8 in which CU1 was explained along with a related idea, CU5.

Pattern 2B

PyrEval matched the sentence that had a lower number of segmentations to a CU instead of the sentence that had a higher number of segmentations. For example, in Essay 012 (Figure 9), PyrEval matched Sentence 12 (in green) to CU5 (Figure 10) instead of Sentence 10 (in yellow) that was identified during manual assessment. Sentence 12 was not segmented whereas Sentence 10 was segmented (Figure 11). This result can be attributed to differences between the length of Sentence 12 (25 words) and Sentence 10 (33 words). Longer sentences tend to cause errors in sentence segmenting during the search for matching text for CUs. Even though students had a pop-up reminder asking them to make each sentence <25 words as they wrote essays, some sentences in submitted essays contained more than 25 words.

Close to but still below 0.5

A final reason PyrEval did not detect a CU in essays, when manual assessment identified them, was because the essay only had sentences/segments that were assigned cosine

A recommendation that I have for the Gonzales family is that the hill height must be smaller than the initial drop because otherwise the car won't make it over the hill and make it to the end of the ride. In one of my experiments, when the hill height was at 4.00 and the initial drop was at 4.00 the car didn't make it over the hill and got stuck at the bottom between the hill and initial drop. Also making the mass bigger will increase the amount of potential and kinetic energy in the ride. When I put the car mass 40 kg, the potential energy was at 1953 at the top of the initial drop and at the bottom of the drop the kinetic energy was at 1953. But when you put the car mass up to 60 kg, the potential energy was at 2930 at the top of the initial drop and at the bottom of the drop it was at 2930 kinetic energy. The potential energy will always add up to equal the kinetic energy at the top and bottom of the initial drop, unless friction was added. But the total energy will always depend on the potential and kinetic energy. The bigger the height of the drop will also make it so that their will be more potential energy at the top of the initial drop and more kinetic energy at the bottom. At the top of the hill the higher up it is will mean that their will be more potential energy stored up in that car. The more potential energy stored up will make it have more kinetic energy while going down the initial drop into the start of the hill. When the potential energy is stored up at the top of the drop the car starts slowly going down the initial drop. The potential energy will gradually start to turn into kinetic energy and at the end of the initial drop the energy will almost always be fully transferred to the kinetic energy from the potential energy. Lastly, having the initial drop higher than the hill height is the best idea for the riders enjoyment and safety so that the car won't get stuck between the initial drop and the hill.

FIGURE 9 An example essay (Essay 012) showing a discrepancy in assessment results between the manual coding and PyrEval.

```
Sentence: 12 | Segmentation: 0

        Segment ID: 0 | Content Unit: 5 [Weight: 5]
        Segment: ..................... The more potential energy stored up will make it have more kinetic
                 ..................... energy while going down the initial drop into the start of the
                 ..................... hill.
        Content Unit: ..............  (1) When another roller coaster was dropped from a lower height
                      ..............      such as meters, the potential energy was only calculated to
                      ..............      Joules.
                      ..............  (2) the higher the drop the more potential energy will have at the
                      ..............      top KE.
                      ..............  (3) The taller the initial drop, the more potential energy the
                      ..............      roller coaster has.
                      ..............  (4) So the higher the drop the more potential energy.
                      ..............  (5) At the initial drop, the potential energy depends on the
                      ..............      height.
```

FIGURE 10 Part of a log output for Essay 012 that PyrEval generated showing Sentence 12 assessed as Content Unit 5.

Content Unit	0	1	2	3	4	5
14&10&0&0	0.6016076690613120	0.5889325553783160	nan	0.51345549086377	nan	0.6958369122677760
14&10&1&0	0.6089617333421480	0.6202323451622600	nan	0.5166591986045390	nan	0.7401808506695320
14&12&0&0	0.6949909850275570	0.6409286746487740	nan	0.5259603616022400	nan	0.6667829902621100

FIGURE 11 Part of cosine similarity value outputs for Essay 012 showing Sentence 10 (in yellow) and Sentence 12 (in green).

similarity values that were lower than 0.5. For example, there was one essay (Essay 112) where manual assessment matched one sentence that contained non-normative grammar ("The law conservation of energy states energy can't be gained or lost in a closed system") to CU2 but PyrEval did not. The reason why PyrEval did not match this sentence to CU2 was because the cosine similarity for CU2 was 0.49, which was below the threshold, 0.5., but close enough to 0.5 for manual assessment to identify. In another essay (Essay 015) including a similar sentence that contained non-normative grammar, CU2 was matched ("The Law of conservation of energy law states that no energy is lost or gain while an object is moving").

In sum, among the 17 essays in which sentences/segments containing non-normative writing features were not matched to CUs that manual assessment identified, we found no case in which PyrEval penalized for having non-normative features. In other words, none of the discrepancies in these 17 essays resulted from a failure of PyrEval assessment to identify the sentences/segments with CUs due to non-normative writing features. The patterns that emerged from our comparison analysis between our manual assessment and PyrEval assessment in all these 17 essays were not related to non-normative writing features. This means that PyrEval did not penalize these 17 essays for having non-normative features. Thus, also given that PyrEval did not penalize the other 105 of the 122 essays that contained at least one non-normative writing feature, as reported earlier, PyrEval assessed essays regardless of non-normative features.

Difference in essay quality change according to presence or absence of non-normative writing features

We ran a two-way repeated measures ANOVA with the presence of non-normative features as a between-subjects factor. The sphericity assumption was met, so uncorrected within-subjects effects were used. The results indicated that there was a significant increase in essay quality scores between initial essays ($M = 4.44$, $SD = 1.36$) and revised essays ($M = 5.20$, $SD = 1.10$), $F(1, 305) = 155.81$, $p < 0.001$, Cohen's $d = 0.714$. This essay quality improvement was not statistically different between students whose initial essays contained non-normative features and students whose initial essays contained no non-normative features. That is, there was no significant main effect of the presence of non-normative features on essay quality improvement, $p = 0.656$ (see Table 3). Particularly, the difference in essay quality improvement between the two groups was equal to 0.03 standard deviation (Cohen's $d = 0.03$). There was no significant interaction effect between the presence of non-normative features in essays and time on essay quality improvement ($p = 0.500$). See Table 3 and Figure 12. In sum, students' essay quality significantly improved from their initial to revised essays, after receiving feedback from PyrEval, regardless of the presence ($M = 4.783$, $SE = 0.101$) or absence ($M = 4.841$, $SE = 0.082$) of non-normative features in initial essays.

TABLE 3 Two-way repeated measures ANOVA results with between-subject effect of the presence of non-normative features on essay quality change from initial essays to revised essays.

	df	*F*	Sig.	Effect size (Cohen's *d*)
Between-subject effects				
Presence of non-normative features in initial essays	1	0.198	0.656	0.03
Error	305			
Within-subject effects				
Time	1	155.81	<0.001	0.714
Time* Presence of non-normative features in initial essays	1	0.455	0.500	0.03
Error	305			

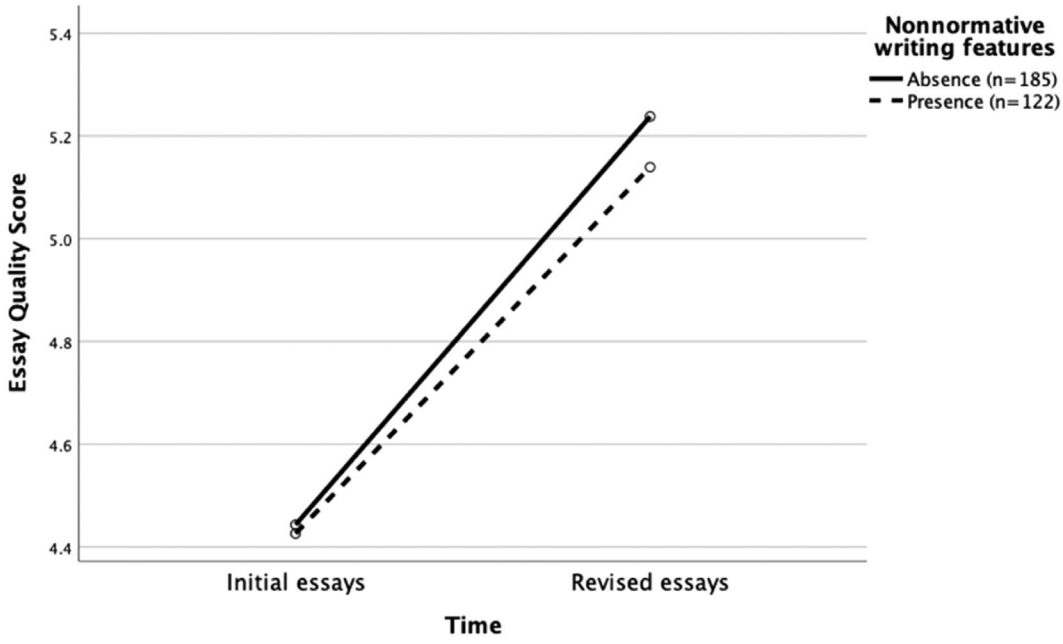

FIGURE 12 Essay quality change from initial to revised essays of students whose initial essays had non-normative writing features versus students whose initial essays had no non-normative writing features.

Difference in essay quality change according to the number of non-normative writing features

We ran a two-way repeated measures ANOVA with the number of non-normative features as a between-subjects factor. The sphericity assumption was met, so uncorrected within-subjects effects were used. The results indicated that students' essay quality significantly increased from their initial to revised essays as shown in the first analysis results above. This essay quality improvement was not statistically different among students whose initial essays contained no, some, or many non-normative features. That is, there was no significant main effect of the number of non-normative features on essay quality improvement, $p=0.800$. There was no significant interaction effect between the number of non-normative features in essays and time (from initial to revised) on essay quality improvement, $p=0.763$. <0.1% of the change in essay quality score could be accounted for by the number of non-normative features ($\eta^2<0.001$; see Table 4 and Figure 13). So, the average quality of essays with no non-normative feature ($M=4.841$, $SE=0.082$), some non-normative features ($M=4.821$, $SE=0.126$) and many non-normative features ($M=4.716$, $SE=0.168$) was not statistically different. In sum, regardless of how many non-normative features were in initial essays, essay quality significantly improved from initial essays to revised essays after receiving feedback from PyrEval.

THE ROLE OF TYPE OF NON-NORMATIVE FEATURES IN ESSAY QUALITY IMPROVEMENT

The generalized linear mixed effect model (GLMM) analysis results indicated that there was a significant fixed effect of time ($\beta=0.762\pm0.097$, $t=7.810$, $p<0.001$) on essay quality (Table 5). This reflects the fact that essay scores significantly improved from initial essays to

TABLE 4 Two-way repeated measures ANOVA results with between-subject effect of the number of non-normative features on essay quality change from initial essays to revised essays.

	df	*F*	Sig.	Effect size
Between-subject effects				
Number of non-normative features in initial essays	2	0.223	0.800	<0.001[a]
Error	304			
Within-subject effects				
Time	1	113.567	<0.001	0.610[b]
Time* Number of non-normative features in initial essays	2	0.270	0.763	0.002[a]
Error	304			

[a]Partial eta squared effect size.
[b]Cohen's *d* effect size.

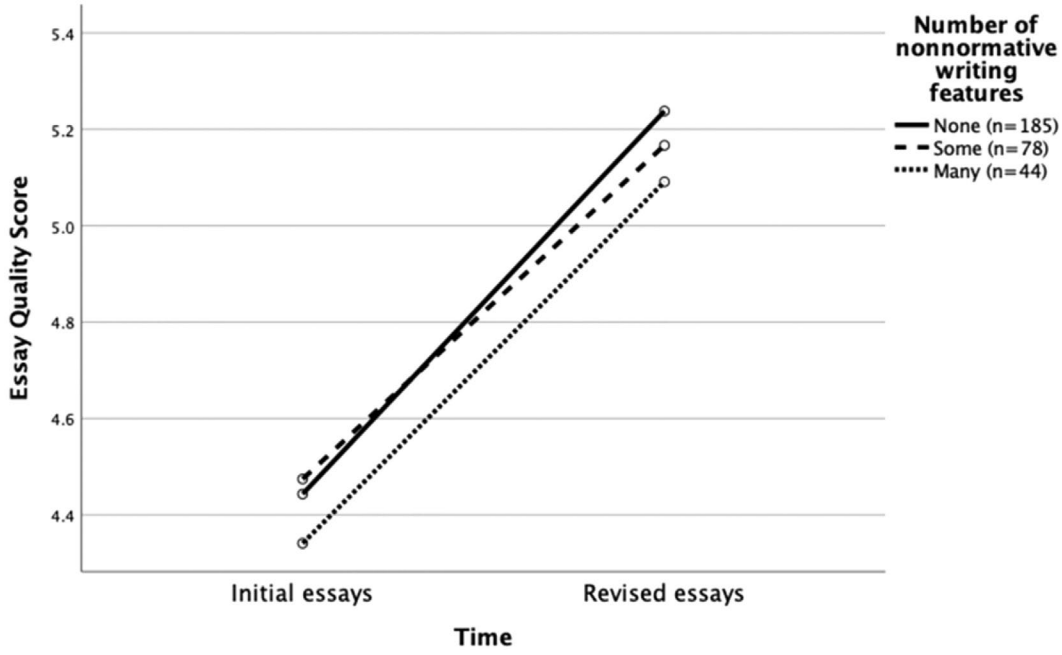

FIGURE 13 Essay quality change from initial to revised essays whose initial essays had *no* non-normative writing feature, *some* non-normative features and *many* non-normative features.

TABLE 5 Fixed effects of variables on essay quality change.

Variable	Coefficient[a]	*SE*	*t*	Sig.
Time	0.762	0.097	7.810	<0.001
Prior knowledge	0.038	0.026	1.436	0.152
Types of non-normative features = AAE	0.035[b]	0.141	0.247	0.805
Types of non-normative features = EL	−0.207[b]	0.177	−1.168	0.243
Type of non-normative features = mixed	0.093[b]	0.143	0.650	0.516
$R^2 = 0.58$				

[a]Probability distribution: Gamma; Link Function: Log.
[b]Base level: types of non-normative features = none.

revised essays after receiving feedback from PyrEval. No significant effect was found in other variables. That is, the type of non-normative features in initial essays was not a significant predictor for essay quality change. Prior physics knowledge was not a significant variable in explaining essay quality change. Neither the teacher variable nor the school district variable had random effects.

DISCUSSION

The performance of PyrEval in assessing essays containing non-normative writing features

Manual coding and examination results indicated that PyrEval did not penalize student essays that contained non-normative writing features. Essays were assessed according to whether science concepts and relations were explained or not, regardless of non-normative writing features. As noted earlier, non-normative features in the present study refer to non-normative grammatical patterns that do not hinder communication about the content of explanations but are often assessed lower without valuing and respecting linguistic diversity (Johnson & VanBrackle, 2012). For example, the sentence shown in Figure 3 included non-normative features ("can convert to") and yet was assessed in terms of main ideas/content units as done in other sentences without non-normative grammar. This finding is important because, as argued in Kucirkova et al. (2021), automated assessment tools in science learning contexts should not create biases or demands related to language backgrounds of students. Considering automated assessment results that varied according to whether students were English learners or not (Liu et al., 2016; Loukina et al., 2019; Wilson et al., 2024), the present study offers implications for equitable automated assessments.

As described in the method section, PyrEval assesses the coverage of important ideas in essays by comparing their propositions to the weighted content (ie, content units) in the model content (Gao et al., 2019; Passonneau et al., 2018). Thus, how many of the main ideas (Table 1) were present mattered, but how the content was expressed did not. Not only historical essay data but also essay data from the implementation of the first version of PyrEval in two school districts produced "multiple human models" (Nenkova et al., 2007, p. 2). That is, essay data from the two participating school districts were used in developing PyrEval for the present study. This means that reference models used in adapting and refining PyrEval included student writing that contained not only normative writing features but also non-normative writing features. The training data were likely to be representative of student populations for which PyrEval was used. While doing so, PyrEval captured variations in student writing and became less biased towards non-normative writing features. When training data did not include variations in scientific explanations using "different words or grammatical structures to express similar ideas" (Liu et al., 2016, p. 229), automated assessments were harsh on English learners' explanations, because, as Liu et al. (2016) noted:

> synonyms may not receive the same score because one word was well-represented in the training data and the other was not. Misspelling is another typical type of variation that may not be adequately modeled by scoring engines … Improvements in NLP are needed to deal with synonyms, misspellings, and odd grammatical structures.
>
> (p. 229)

Our finding can be attributed to another aspect of PyrEval adaptation and refinement process. Multiple rounds of reference essay collection and modification in wise crowd content models were done by our interdisciplinary team members, rather than simply mapping an assessment rubric onto state standards. Dixon-Román et al. (2020) reported on an automated essay feedback tool serving as "racializing assemblages" (Dixon-Román et al., 2020, p. 236) that was developed exclusively based on state standards and sample essay data that did not include marginalized ways of expression. In the present study, while our co-designer teachers' engagement in this multi-year process was crucial and their pedagogical goal for science teaching was not separable from state standards, we as a team constantly questioned what should be essential in scientific explanations when evaluating and refining the performance of PyrEval. For instance, part of our core conversations were about how diverse essays, including those written in *non-normative* ways, should be credited for their inclusion of main ideas.

Another contribution of our finding is from the actual examination of non-normative writing features in student essays. We investigated in what ways PyrEval worked to assess essays that contained non-normative features. While limited, there are some studies reporting the performance of automated tools on assessing the writing of English learners or other students with marginalized backgrounds (Lee et al., 2019; Litman et al., 2021; Liu et al., 2016; Loukina et al., 2019; Wilson et al., 2024). While these studies provide strong evidence for the need to eliminate linguistic discrimination perpetuated by automated assessment tools, demographic data were used in their analysis rather than actual non-normative writing features. For example, Lee et al. (2019) mentioned a minimal role that non-normative features played in their automated assessment tool when student scientific explanations contained ideal content, but no further detail was provided. Ha and Nehm (2016) reported some impact of misspelled words on automated assessment accuracy, but other non-normative writing features were not examined and what happened in the assessment of writing with misspelled words was not fully explained. Given the current literature, our finding about the ways in which PyrEval assessed essays that contained non-normative writing features uniquely contributes to automated assessment research. Our finding is also timely and important considering growing concerns about AI reproducing inequity in education and ongoing research in improving AI fairness. For example, as noted earlier in our conceptual framework, linguistic discrimination against non-normative English use through discriminatory AI (Brandt & Hazel, 2024; Cunningham et al., 2024; Jeon et al., 2024; Martin & Wright, 2023; Ngueajio & Washington, 2022) creates linguistic oppression in science and education (Martin & Wright, 2023; Payne et al., 2024). We hope that our finding inspires AI researchers to attend to linguistic discrimination in their efforts related to AI fairness. For example, our list of non-normative features could be used as a basis for AI researchers to experiment with their AI models using/creating text and speech corpuses that contain such features and thereby minimizing linguistic discrimination in AI.

Our finding is not surprising considering the design of PyrEval, but it also provides directions for further development towards improved feedback. For example, when essays include multiple sentences/segments in which cosine similarity values are assigned to one same CU, the feedback can invite students to revise their essays not to repeat text related to the CU. When essays include sentences that were segmented multiple times, the feedback can ask students to make the sentence into two or more. Although our purpose of using PyrEval is not to improve language fluency per se, these kinds of feedback can help students learn to communicate scientific explanations more effectively. In so doing, students can revisit and improve their understanding of relations among scientific concepts.

Number of non-normative writing features and essay quality change

Repeated measures ANOVA results indicated that students' essay quality significantly improved from their initial to revised essays, after receiving feedback from PyrEval, regardless of the presence or absence of non-normative writing features in initial essays. The number of non-normative features also did not matter. In other words, all groups of students, including those whose initial essays contained many non-normative features (more than 4) exhibited statistically significant improvement in their essay quality when revised. As noted earlier, essay quality was measured by a total score for the inclusion of explanations of the six key science concepts and their relations; that is, content units (CUs).

The positive impact of NLP-based automated assessment and feedback on student science writing has been reported in several studies (eg, Boda et al., 2021; Gerard & Linn, 2022; Lee et al., 2019; Tansomboon et al., 2017). None of these studies detailed how non-normative writing features such as subject-verb disagreement were handled. Still, especially transparent feedback in Tansomboon et al. (2017), which included an explanation of how automated assessment worked for students, is relevant to our finding. In the present study, students were provided CU-level information about the accuracy of PyrEval (explained elsewhere; Kim et al., 2024). Students were not discouraged by feedback from PyrEval when their essays did not cover certain CUs because the feedback was presented with a question mark rather than point deduction (see Figure 2). Not being penalized for non-normative writing features may have increased students' trust in PyrEval and thereby their engagement in revisions as students whose essays had no non-normative features did. Further research is needed.

Type of non-normative writing features and essay quality improvement

None of the non-normative writing features (ie, AAE and EL; see Table 2) in student essays was a predictor of essay quality improvement in the present study. Considering numerous studies reporting human assessors' bias or inaccuracy due to non-normative grammar (eg, Appelman & Schmierbach, 2018; Johnson & VanBrackle, 2012), this finding showcases the potential of using AI to ignore grammatical and other mechanical errors in science writing assessment and "make a reasonable inference that the student 'knows' the response at a given performance level" (Shermis, 2015, p. 49) without linguistic discrimination. As discussed earlier, marginalized ways of expression should be valued in automated assessment and feedback tools (Dixon-Román et al., 2020).

We also found no statistically significant relation between the type of non-normative writing features and essay quality improvement, regardless of school districts and teachers. As shown in Appendix A, student essays from school district 2 contained significantly more AAE and EL grammatical features, $t\,(20.126) = -2.674$, $p = 0.015$, Cohen's $d = 0.087$. Nonetheless, GLMM results revealed that essay quality significantly improved regardless of school district and teacher. In addition, student prior physics knowledge was not a significant predictor for essay quality improvement. This finding is important considering that the impact of automated writing feedback on students with low prior knowledge varied across schools in Tansomboon et al. (2017). As described earlier, our prior knowledge assessment covered physics content in the unit and directly related to CUs. Considering automated feedback that responded to the presence and absence of CUs in student essays, students received feedback on their physics knowledge that was integrated into their writing. Thus, the feedback accounted for variations in student understanding. This may have contributed to the finding that prior knowledge was not a significant predictor for

essay quality improvement; that is, regardless of their prior knowledge and (non)normative writing features, students included more CUs in their revised essays after receiving automated feedback. These findings should be further studied to trace individual progress from prior knowledge to initial essays, to automated feedback and to revisions, as well as subsequent knowledge progression.

LIMITATION OF THE STUDY

The present study focused on non-normative writing features and how they were assessed by PyrEval, rather than which racial or linguistic groups of students received (un)biased assessment on their essays. Not all English learners use EL non-normative features and not all African American students use AAE non-normative features (Latimer-Hearn, 2020). Still, some may wonder whether EL non-normative features were made by English learners and AAE non-normative features were made by African American students. Future research could study racial or linguistic groups, in addition to the actual examination of non-normative writing features. Albeit out of the study scope, interviews with student participants, especially those whose essays contained many non-normative writing features and improved when revised, may have provided rich data about how they perceived feedback from PyrEval and engaged in their essay revisions.

CONCLUSION

The present study offers possibilities towards eliminating linguistic discrimination in science writing assessment. As criticized in our conceptual framework, languagelessness (Rosa, 2016) is problematic in science (Lynch et al., 2021) and science education (Barton & Tan, 2009; Lee, 2005; Lyon et al., 2012). PyrEval can be used as a reflective tool for human assessors to check their own bias and inaccuracy. Human assessors' bias or inaccuracy towards non-normative grammar have been studied for a long time (eg, Appelman & Schmierbach, 2018; Johnson & VanBrackle, 2012). How human bias or inaccuracy is manifested varies. For example, in Liu et al. (2016), human assessors were lenient on explanations containing non-normative features and scored them higher than automated assessment did. In Wilson et al. (2024), both human and automated assessments led to lower scoring of English learners' scientific writing than that of non-English learners, but automated assessment using an analytic approach was even harsher on the writing of English learners. This research points to the potential of leveraging the strengths of collaboration between human assessors and AI to reduce linguistic discrimination.

Attention to linguistic discrimination that AI may practice is an ethical responsibility that should continue. As extensively discussed in the literature on AI ethics (Ayling & Chapman, 2022; Bleher & Braun, 2023; Borenstein & Howard, 2021; Hagendorff, 2020; Heilinger, 2022; Hickok, 2021; Huang et al., 2023; Morley et al., 2023), algorithms are made by humans who are inherently biased. Unintentional discrimination that various biases of AI cause due to training data has long been documented (eg, Chan, 2023; Deshpande et al., 2020; Dusi et al., 2024; Ferrer et al., 2021; Henderson et al., 2018; Howard & Borenstein, 2018; Jenks, 2024; Lauer, 2021; Morley et al., 2020; O'Connor & Liu, 2024). Vigilant consideration of AI bias and discrimination is necessary among not only designers and developers but also educators, as identified in the need for AI ethics education (Chee et al., 2024; Garrett et al., 2020). We hope that the present study inspires such consideration to materialize among AI researchers and educators.

ACKNOWLEDGEMENTS

This work was supported by grants 2010351 and 2010483 from the National Science Foundation (USA). Any opinions, findings or conclusions are those of the authors and do not necessarily represent official positions of the National Science Foundation.

CONFLICT OF INTEREST STATEMENT

There was no conflict of interest to report.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ETHICS STATEMENT

Ethical approval for this research was obtained from the authors' institutional review board.

ORCID

ChanMin Kim https://orcid.org/0000-0001-9383-8846
Eunseo Lee https://orcid.org/0000-0003-2948-3328
Mahsa Sheikhi Karizaki https://orcid.org/0009-0001-3167-0950
Dana Gnesdilow https://orcid.org/0000-0001-8977-6187

REFERENCES

Abdilla, A., Arista, N., Baker, K., Benesiinaabandan, S., Brown, M., Cheung, M., Coleman, M., Cordes, A., Davison, J., Duncan, K., Garzon, S., Harrell, D. F., Jones, P.-L., Kealiikanakaoleohaililani, K., Kelleher, M., Kite, S., Lagon, O., Leigh, J., Levesque, M., … Whaanga, H. (2020). *Indigenous protocol and artificial intelligence position paper.* https://doi.org/10.11573/SPECTRUM.LIBRARY.CONCORDIA.CA.00986506

Alim, H. S. (2010). Critical language awareness. In N. H. Hornberger & S. McKay (Eds.), *Sociolinguistics and language education* (Vol. *1*, pp. 205–231). Multilingual Matters.

Almusharraf, N., & Alotaibi, H. (2023). An error-analysis study from an EFL writing context: Human and automated essay scoring approaches. *Technology, Knowledge and Learning*, *28*(3), 1015–1031. https://doi.org/10.1007/s10758-022-09592-z

Amano, T., Rios Rojas, C., Boum, Y., II, Calvo, M., & Misra, B. B. (2021). Ten tips for overcoming language barriers in science. *Nature Human Behaviour*, *5*(9), 1119–1122. https://doi.org/10.1038/s41562-021-01137-1

Appelman, A., & Schmierbach, M. (2018). Make no mistake? Exploring cognitive and perceptual effects of grammatical errors in news articles. *Journalism and Mass Communication Quarterly*, *95*(4), 930–947. https://doi.org/10.1177/1077699017736040

Ayling, J., & Chapman, A. (2022). Putting AI ethics to work: Are the tools fit for purpose? *AI and Ethics*, *2*(3), 405–429. https://doi.org/10.1007/s43681-021-00084-x

Barton, A. C., & Tan, E. (2009). Funds of knowledge and discourses and hybrid space. *Journal of Research in Science Teaching*, *46*(1), 50–73. https://doi.org/10.1002/tea.20269

Bearman, M., & Ajjawi, R. (2023). Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, *54*(5), 1160–1173. https://doi.org/10.1111/bjet.13337

Bleher, H., & Braun, M. (2023). Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics*, *29*(3), 21. https://doi.org/10.1007/s11948-023-00443-3

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.485

Boda, P. A., Bathia, S., & Linn, M. C. (2021). Longitudinal impact of interactive science activities: Developing, implementing, and validating a graphing integration inventory. *Journal of Research in Science Teaching*, *58*(2), 225–248. https://doi.org/10.1002/tea.21653

Borenstein, J., & Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, *1*(1), 61–65. https://doi.org/10.1007/s43681-020-00002-7

Boulanger, D., & Kumar, V. (2020). SHAPed automated essay scoring: Explaining writing features' contributions to English writing organization. In V. Kumar & C. Troussas (Eds.), *Intelligent tutoring systems* (pp. 68–78). Springer International Publishing. https://doi.org/10.1007/978-3-030-49663-0_10

Brandt, A., & Hazel, S. (2024). Towards interculturally adaptive conversational AI. *Applied Linguistics Review*, *16*. https://doi.org/10.1515/applirev-2024-0187

Chan, A. (2023). GPT-3 and InstructGPT: Technological dystopianism, utopianism, and "contextual" perspectives in AI ethics and industry. *AI and Ethics*, *3*(1), 53–64. https://doi.org/10.1007/s43681-022-00148-6

Chee, H., Ahn, S., & Lee, J. (2024). A competency framework for AI literacy: Variations by different learner groups and an implied learning pathway. *British Journal of Educational Technology*. https://doi.org/10.1111/bjet.13556

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, *10*(1), 1. https://doi.org/10.18608/jla.2023.7801

Correnti, R., Matsumura, L. C., Wang, E. L., Litman, D., & Zhang, H. (2022). Building a validity argument for an automated writing evaluation system (eRevise) as a formative assessment. *Computers and Education Open*, *3*, 100084. https://doi.org/10.1016/j.caeo.2022.100084

Cunningham, J., Blodgett, S. L., Madaio, M., Daumé Iii, H., Harrington, C., & Wallach, H. (2024). Understanding the impacts of language technologies' performance disparities on African American language speakers. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: ACL 2024* (pp. 12826–12833). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.761

Deshpande, K. V., Pan, S., & Foulds, J. R. (2020). Mitigating demographic bias in AI-based resume filtering. *Adjunct publication of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 268–275). https://doi.org/10.1145/3386392.3399569

Dixon-Román, E., Nichols, T. P., & Nyame-Mensah, A. (2020). The racializing forces of/in AI educational technologies. *Learning, Media and Technology*, *45*(3), 236–250. https://doi.org/10.1080/17439884.2020.1667825

Duschl, R. A., & Osborne, J. (2002). Supporting and promoting argumentation discourse in science education. *Studies in Science Education*, *38*(1), 39–72. https://doi.org/10.1080/03057260208560187

Dusi, M., Arici, N., Emilio Gerevini, A., Putelli, L., & Serina, I. (2024). Discrimination bias detection through categorical association in pre-trained language models. *IEEE Access*, *12*, 162651–162667. https://doi.org/10.1109/ACCESS.2024.3482010

Edelblut, P. (2020). Realizing the promise of AI-powered, adaptive, automated, instant feedback on writing for students in grade 3–8 with an IEP. In R. A. Sottilare & J. Schwarz (Eds.), *Adaptive instructional systems* (pp. 283–292). Springer International Publishing. https://doi.org/10.1007/978-3-030-50788-6_21

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Faulkner-Bond, M., & Sireci, S. G. (2015). Validity issues in assessing linguistic minorities. *International Journal of Testing*, *15*(2), 114–135. https://doi.org/10.1080/15305058.2014.974763

Ferrer, X., van Nuenen, T., Such, J. M., Coté, M., & Criado, N. (2021). Bias and discrimination in AI: A cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80. https://doi.org/10.1109/MTS.2021.3056293

Gao, Y., Sun, C., & Passonneau, R. J. (2019). Automated pyramid summarization evaluation. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, 404–418. https://doi.org/10.18653/v1/K19-1038

Garrett, N., Beard, N., & Fiesler, C. (2020). More than "if time allows": The role of ethics in AI education. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 272–278. https://doi.org/10.1145/3375627.3375868

Gerard, L., Kidron, A., & Linn, M. C. (2019). Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning*, *14*(3), 291–324. https://doi.org/10.1007/s11412-019-09298-y

Gerard, L., & Linn, M. C. (2022). Computer-based guidance to support students' revision of their science explanations. *Computers & Education*, *176*, 104351. https://doi.org/10.1016/j.compedu.2021.104351

Goldshtein, M., Alhashim, A. G., & Roscoe, R. D. (2024). Automating bias in writing evaluation: Sources, barriers, and recommendations. In *The Routledge international handbook of automated essay evaluation*. Routledge.

Ha, M., & Nehm, R. H. (2016). The impact of misspelled words on automated computer scoring: A case study of scientific explanations. *Journal of Science Education and Technology*, *25*(3), 358–374. https://doi.org/10.1007/s10956-015-9598-9

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, *30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Heilinger, J.-C. (2022). The ethics of AI ethics: A constructive critique. *Philosophy and Technology*, *35*(3), 61. https://doi.org/10.1007/s13347-022-00557-9

Henderson, P., Sinha, K., Angelard-Gontier, N., Ke, N. R., Fried, G., Lowe, R., & Pineau, J. (2018). Ethical challenges in data-driven dialogue systems. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 123–129. https://doi.org/10.1145/3278721.3278777

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI and Ethics*, *1*(1), 41–47. https://doi.org/10.1007/s43681-020-00008-1

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics*, *24*(5), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819. https://doi.org/10.1109/TAI.2022.3194503

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies*, *5*(1), 1–17.

Jank, I. (2017, March 20). A new method for testing the role of linguistic discrimination in pedagogical evaluation. *4th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM2017*. https://doi.org/10.5593/SGEMSOCIAL2017/HB31/S10.002

Jenks, C. J. (2024). Communicating the cultural other: Trust and bias in generative AI and large language models. *Applied Linguistics Review*, *16*. https://doi.org/10.1515/applirev-2024-0196

Jeon, J., Lee, S., & Coronel-Molina, S. M. (2024). Rethinking AI: Bias in speech-recognition chatbots for ELT. *ELT Journal*, *78*(4), 435–445. https://doi.org/10.1093/elt/ccae035

Johnson, D., & VanBrackle, L. (2012). Linguistic discrimination in writing assessment: How raters react to African American "errors," ESL errors, and standard English errors on a state-mandated writing exam. *Assessing Writing*, *17*(1), 35–54. https://doi.org/10.1016/j.asw.2011.10.001

Karizaki, M. S., Gnesdilow, D., Puntambekar, S., & Passonneau, R. J. (2024). How well can you articulate that idea? Insights from automated formative assessment. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education* (pp. 225–233). Springer Nature. https://doi.org/10.1007/978-3-031-64299-9_16

Kim, C., Puntambekar, S., Lee, E., Gnesdilow, D., Karizaki, M. S., & Passonneau, R. J. (2024). NLP-enabled automated feedback about science writing. In R. Lindgren, T. I. Asino, E. A. Kyza, C. K. Looi, D. T. Keifert, & E. Suárez (Eds.), *Proceedings of the 18th International Conference of the Learning Sciences (ICLS)* (pp. 2431–2432). International Society of the Learning Sciences. https://doi.org/10.22318/icls2024.712073

Krajcik, J., & McNeill, K. L. (2015). Designing and assessing scientific explanation tasks. In R. Gunstone (Ed.), *Encyclopedia of science education* (pp. 285–291). Springer Netherlands. https://doi.org/10.1007/978-94-007-2150-0_48

Kucirkova, N., Gerard, L., & Linn, M. C. (2021). Designing personalised instruction: A research and design framework. *British Journal of Educational Technology*, *52*(5), 1839–1861. https://doi.org/10.1111/bjet.13119

Latimer-Hearn, D. (2020). Don't get it twisted- hear my voice. *ASHA Leader*, *25*(1), 54–59. https://doi.org/10.1044/leader.FTR2.25012020.54

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics*, *1*(1), 21–25. https://doi.org/10.1007/s43681-020-00013-4

LaVoie, N., Parker, J., Legree, P. J., Ardison, S., & Kilcullen, R. N. (2020). Using latent semantic analysis to score short answer constructed responses: Automated scoring of the consequences test. *Educational and Psychological Measurement*, *80*(2), 399–414. https://doi.org/10.1177/0013164419860575

Lee, A., Luco, A. C., & Tan, S. C. (2023). A human-centric automated essay scoring and feedback system for the development of ethical reasoning. *Educational Technology & Society*, *26*(1), 147–159. https://doi.org/10.30191/ETS.202301_26(1).0011

Lee, H.-S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education*, *103*(3), 590–622. https://doi.org/10.1002/sce.21504

Lee, O. (2005). Science education with English language learners: Synthesis and research agenda. *Review of Educational Research*, *75*(4), 491–530. https://doi.org/10.3102/00346543075004491

Lin, Y.-T., Hung, T.-W., & Huang, L. T.-L. (2021). Engineering equity: How AI can help reduce the harm of implicit bias. *Philosophy and Technology*, *34*(1), 65–90. https://doi.org/10.1007/s13347-020-00406-7

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). A fairness evaluation of automated methods for scoring text evidence usage in writing. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 255–267). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_21

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233. https://doi.org/10.1002/tea.21299

Loukina, A., Madnani, N., & Zechner, K. (2019). The many dimensions of algorithmic fairness in educational applications. In H. Yannakoudakis, E. Kochmar, C. Leacock, N. Madnani, I. Pilán, & T. Zesch (Eds.), *Proceedings*

of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications (pp. 1–10). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4401

Lynch, A. J., Fernández-Llamazares, Á., Palomo, I., Jaureguiberry, P., Amano, T., Basher, Z., Lim, M., Mwampamba, T. H., Samakov, A., & Selomane, O. (2021). Culturally diverse expert teams have yet to bring comprehensive linguistic diversity to intergovernmental ecosystem assessments. *One Earth*, *4*(2), 269–278. https://doi.org/10.1016/j.oneear.2021.01.002

Lyon, E. G., Bunch, G. C., & Shaw, J. M. (2012). Navigating the language demands of an inquiry-based science performance assessment: Classroom challenges and opportunities for English learners. *Science Education*, *96*(4), 631–651. https://doi.org/10.1002/sce.21008

Mahboob, A., & Szenes, E. (2010). Linguicism and racism in assessment practices in higher education. *Linguistics and the Human Sciences*, *3*(3), 325–354. https://doi.org/10.1558/lhs.v3i3.325

Martin, J. L., & Wright, K. E. (2023). Bias in automatic speech recognition: The case of African American language. *Applied Linguistics*, *44*(4), 613–630. https://doi.org/10.1093/applin/amac066

McNeill, K. L., & Berland, L. (2017). What is (or should be) scientific evidence use in K-12 classrooms? *Journal of Research in Science Teaching*, *54*(5), 672–689. https://doi.org/10.1002/tea.21381

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society*, *38*(1), 411–423. https://doi.org/10.1007/s00146-021-01308-8

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., & Floridi, L. (2020). The ethics of AI in health care: A mapping review. *Social Science & Medicine*, *260*, 113172. https://doi.org/10.1016/j.socscimed.2020.113172

Nenkova, A., Passonneau, R., & McKeown, K. (2007). The Pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing*, *4*(2), 4. https://doi.org/10.1145/1233912.1233913

Ngueajio, M. K., & Washington, G. (2022). Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques: A literature review. In J. Y. C. Chen, G. Fragomeni, H. Degen, & S. Ntoa (Eds.), *HCI International 2022—Late breaking papers: Interacting with eXtended reality and artificial intelligence* (Vol. *13518*, pp. 421–440). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21707-4_30

O'Connor, S., & Liu, H. (2024). Gender bias perpetuation and mitigation in AI technologies: Challenges and opportunities. *AI & Society*, *39*(4), 2045–2057. https://doi.org/10.1007/s00146-023-01675-4

Passonneau, R. J., Poddar, A., Gite, G., Krivokapic, A., Yang, Q., & Perin, D. (2018). Wise crowd content assessment and educational rubrics. *International Journal of Artificial Intelligence in Education*, *28*(1), 29–55. https://doi.org/10.1007/s40593-016-0128-6

Payne, A. L., Austin, T., & Clemons, A. M. (2024). Beyond the front yard: The dehumanizing message of accent-altering technology. *Applied Linguistics*, *45*(3), 553–560. https://doi.org/10.1093/applin/amae002

Peters, U. (2023). Linguistic discrimination in science: Can English disfluency help debias scientific research? *International Studies in the Philosophy of Science*, *36*(1), 61–79. https://doi.org/10.1080/02698595.2023.2251676

Qin, F., Li, K., & Yan, J. (2020). Understanding user trust in artificial intelligence-based educational systems: Evidence from China. *British Journal of Educational Technology*, *51*(5), 1693–1710. https://doi.org/10.1111/bjet.12994

Rosa, J. D. (2016). Standardization, racialization, languagelessness: Raciolinguistic ideologies across communicative contexts. *Journal of Linguistic Anthropology*, *26*(2), 162–183. https://doi.org/10.1111/jola.12116

Rosa, J. D. (2019). "They're bilingual … That means they don't know the language": The ideology of languagelessness in practice, policy, and theory. In J. D. Rosa (Ed.), *Looking like a language, sounding like a race: Raciolinguistic ideologies and the learning of latinidad* (pp.124-143). Oxford University Press. https://doi.org/10.1093/oso/9780190634728.003.0005

Rosa, J. D., & Burdick, C. (2017). Language ideologies. In O. García, N. Flores, & M. Spotti (Eds.), *The Oxford handbook of language and Society* (pp. 103-124). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190212896.013.15

Rutt, A. A., & Mumba, F. (2022). Pre-service teachers enactment of language- and literacy-integrated science instruction in linguistically diverse science classrooms. *Journal of Research in Science Teaching*, *59*(4), 619–655. https://doi.org/10.1002/tea.21739

Sandoval, W. A., & Millwood, K. A. (2005). The quality of students' use of evidence in written scientific explanations. *Cognition and Instruction*, *23*(1), 23–55. https://doi.org/10.1207/s1532690xci2301_2

Shermis, M. D. (2015). Contrasting state-of-the-art in the machine scoring of short-form constructed responses. *Educational Assessment*, *20*(1), 46–65. https://doi.org/10.1080/10627197.2015.997617

Singh, P., Passonneau, R. J., Wasih, M., Cang, X., Kim, C., & Puntambekar, S. (2022). Automated support to scaffold students' written explanations in science. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.), *Artificial intelligence in education* (pp. 660–665). Springer International Publishing. https://doi.org/10.1007/978-3-031-11644-5_64

Sumner, J. L. (2018). The gender balance assessment tool (GBAT): A web-based tool for estimating gender balance in syllabi and bibliographies. *PS: Political Science & Politics*, *51*(2), 396–400. https://doi.org/10.1017/S1049096517002074

Tansomboon, C., Gerard, L. F., Vitale, J. M., & Linn, M. C. (2017). Designing automated guidance to promote productive revision of science explanations. *International Journal of Artificial Intelligence in Education*, *27*(4), 729–757. https://doi.org/10.1007/s40593-017-0145-0

Wang, Q. (2022). The use of semantic similarity tools in automated content scoring of fact-based essays written by EFL learners. *Education and Information Technologies*, *27*(9), 13021–13049. https://doi.org/10.1007/s10639-022-11179-1

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2024). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*, *61*(1), 38–69. https://doi.org/10.1002/tea.21864

Yang, K., Raković, M., Li, Y., Guan, Q., Gašević, D., & Chen, G. (2024). Unveiling the tapestry of automated essay scoring: A comprehensive investigation of accuracy, fairness, and generalizability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(20), 22466–22474. https://doi.org/10.1609/aaai.v38i20.30254

Zhai, X., Haudek, K. C., Stuhlsatz, M. A. M., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation*, *67*, 100916. https://doi.org/10.1016/j.stueduc.2020.100916

How to cite this article: Kim, C., Passonneau, R. J., Lee, E., Sheikhi Karizaki, M., Gnesdilow, D., & Puntambekar, S. (2025). NLP-enabled automated assessment of scientific explanations: Towards eliminating linguistic discrimination. *British Journal of Educational Technology*, *00*, 1–33. https://doi.org/10.1111/bjet.13596

APPENDIX A

LIST OF NON-NORMATIVE WRITING FEATURES MANUALLY CODED

Type	Non-normative features	Example from student essays	District 1 (224 essays)		District 2 (83 essays)	
			n	%	*n*	%
African American English (AAE) features	Apostrophe deletion*	So if the initial drop *isnt* higher than the hill, the coaster *wont* have the energy it needs to get the car over the hill	23	10.3	32	38.6
	Non-normative apostrophe use* (plural -s in place of apostrophe -s; apostrophe -s in place of third person -s)	For KE the higher the drop the more KE the cart *get's* because it's moving faster and further	14	6.3	16	19.3
	Non-normative use of voice*	When the energy is being *transfer* the energy cant be *crates* or *destroy*, but it an be *transfer* in to kinetic energy or potential energy	12	5.4	4	4.8
	Lack of subjunctive format*	If I had a cart mass of 80 and a cart mass of 5 the cart *will* still both have the same PE and the same KE according to my data. If you had a cart with the same mass but similar hill heights then the total energy *will* be almost the same each time	0	0.0	3	3.6
	Non-normative use of subject-verb agreement*	One last detail on the coaster *are* the hills. For example in my 1st trial I *has* a release height of 2 m	23	10.3	23	27.7
	Unconjugated or deleted copula*	You should have a decently high initial drop because it adds more potential energy which leads to Kinetic energy and *it would more thrill* if there was more speed	1	0.4	1	1.2
	Word choice (their instead of there)	The bigger the height of the drop will also make it so that *their* will be more potential energy at the top of the initial drop and more kinetic energy at the bottom	2	0.9	4	4.8
	Word choice* (your instead of you're; you instead of your)	Then when *your* going down the initial drop the PE levels will go down and the KE levels will rise and when *your* going up the hill the PE levels go up and the KE levels will go down	0	0.0	4	4.8
	Non-normative noun singular or plural choice*	What i mean by that is instead of them going around the track only one time, the Gonzales family could make it go around *two time* instead. So that is my *ideas* to the Gonzales family to make their roller coasters better	0	0.0	1	1.2
	Non-normative verb form*	I hope you *chose* my design, I have *wrote* my design process below	12	5.4	11	13.3

Type	Non-normative features	Example from student essays	District 1 (224 essays)		District 2 (83 essays)	
			n	%	*n*	%
English Learners (EL) features	Non-normative article use*	You need to make sure you have *a* initial drop height taller than the hill height if you want the cart to get over	13	5.8	17	20.5
	Non-normative preposition use*	The KE will depend *about* how much potential energy it gave to the Kinetic energy	8	3.6	11	13.3
	Non-normative use of infinitive to*	I have some recommendations for you the first one is *make* your initial drop higher than your hill	3	1.3	1	1.2
	Word choice (are instead of our)	Meanwhile with *are* second choice being four, it starts at 1955 and ends at 1835, making it slow, and boring	3	1.3	0	0.0
	Word choice (do instead of due)	This is because when going down the drop or hill it is going to go fast *do* to the gravity acting on it	1	0.4	1	1.2
	Word choice (effect instead of affect)	Initial drop height can *effect* what happens to the KE, PE, and total energy	10	4.5	17	20.5
	Word choice (then instead of than)	If the hill is smaller *then* the initial drop the cart will be able to make it over the hill	21	9.4	24	28.9
	Word choice (to instead of too)	When I create my roller coaster I am going to make my hill steeper, but not *to* steep because I do not want there to be *to* much force that my car falls of the track	5	2.2	2	2.4
	Total		145		171	

Note: 1. Features with an asterisk (*) indicate non-normative features identified in Johnson and VanBrackle (2012). 2. Word choices listed as EL non-normative features result from confusion caused by the similar pronunciation of two different words. 3. Word choice listed as AAE non-normative features results from the non-normative use of possessive case.

APPENDIX B

RESULTS FROM COMPARISON BETWEEN MANUAL AND PYREVAL ASSESSMENTS OF SENTENCES CONTAINING BOTH NON-NORMATIVE WRITING FEATURES AND MAIN IDEAS

Essay	Number of sentences containing both main ideas and non-normative features that the manual assessment identified	Number of sentences containing both main ideas and non-normative features that PyrEval assessment identified	Interrater agreement (%)	Patterns observed from PyrEval log output and cosine similarity value examinations on essays with <100% interrater agreement
001	5	4	80	Pattern 2A
002	2	2	100	
003	5	5	100	
004	0	0	100	
005	2	2	100	
006	1	1	100	
007	2	1	50	Pattern 2A
008	1	1	100	
009	1	1	100	
010	2	2	100	
011	1	1	100	
012	2	1	50	Pattern 2B
013	3	2	67	Pattern 2A
014	0	0	100	
015	1	1	100	
016	4	4	100	
017	3	3	100	
018	3	3	100	
019	1	1	100	
020	2	1	50	Pattern 2B
021	1	1	100	
022	3	1	33	Pattern 2A
023	1	1	100	
024	2	2	100	
025	1	1	100	
026	5	5	100	
027	2	2	100	
028	0	0	100	
029	0	0	100	
030	3	2	67	Pattern 2B
031	0	0	100	
032	1	1	100	
033	0	0	100	

Essay	Number of sentences containing both main ideas and non-normative features that the manual assessment identified	Number of sentences containing both main ideas and non-normative features that PyrEval assessment identified	Interrater agreement (%)	Patterns observed from PyrEval log output and cosine similarity value examinations on essays with <100% interrater agreement
034	0	0	100	
035	1	1	100	
036	0	0	100	
037	2	1	50	Pattern 2A
038	0	0	100	
039	0	0	100	
040	0	0	100	
041	2	2	100	
042	0	0	100	
043	1	1	100	
044	0	0	100	
045	0	0	100	
046	0	0	100	
047	1	1	100	
048	0	0	100	
049	0	0	100	
050	0	0	100	
051	0	0	100	
052	3	3	100	
053	2	2	100	
054	1	1	100	
055	0	0	100	
056	0	0	100	
057	0	0	100	
058	2	2	100	
059	1	1	100	
060	0	0	100	
061	2	2	100	
062	0	0	100	
063	0	0	100	
064	0	0	100	
065	1	1	100	
066	0	0	100	
067	1	1	100	
068	0	0	100	
069	0	0	100	
070	1	0	0	Pattern 2A
071	1	1	100	

Essay	Number of sentences containing both main ideas and non-normative features that the manual assessment identified	Number of sentences containing both main ideas and non-normative features that PyrEval assessment identified	Interrater agreement (%)	Patterns observed from PyrEval log output and cosine similarity value examinations on essays with <100% interrater agreement
072	0	0	100	
073	1	1	100	
074	0	0	100	
075	0	0	100	
076	0	0	100	
077	0	0	100	
078	1	1	100	
079	0	0	100	
080	0	0	100	
081	0	0	100	
082	1	0	0	Pattern 1
083	0	0	100	
084	0	0	100	
085	1	1	100	
086	0	0	100	
087	1	0	0	Pattern 2A
088	1	0	0	Pattern 2A
089	1	1	100	
090	1	1	100	
091	0	0	100	
092	0	0	100	
093	1	1	100	
094	0	0	100	
095	0	0	100	
096	0	0	100	
097	0	0	100	
098	1	1	100	
099	1	1	100	
100	1	0	0	Pattern 1
101	1	1	100	
102	1	1	100	
103	1	1	100	
104	1	1	100	
105	0	0	100	
106	1	1	100	
107	1	1	100	
108	1	0	0	Pattern 2A
109	0	0	100	

Essay	Number of sentences containing both main ideas and non-normative features that the manual assessment identified	Number of sentences containing both main ideas and non-normative features that PyrEval assessment identified	Interrater agreement (%)	Patterns observed from PyrEval log output and cosine similarity value examinations on essays with <100% interrater agreement
110	1	1	100	
111	1	1	100	
112	1	0	0	Cosine similarity value = 0.49
113	0	0	100	
114	0	0	100	
115	1	0	0	Pattern 2B
116	2	2	100	
117	0	0	100	
118	0	0	100	
119	4	3	75	Pattern 1
120	0	0	100	
121	0	0	100	
122	1	1	100	

Note: In 105 essays, no discrepancy was found between PyrEval assessment and manual assessment on sentences/segments that contained both main ideas and non-normative features. In 17 essays, discrepancies were found between PyrEval assessment and manual assessment on sentences/segments that contained both main ideas and non-normative features.