Automated Essay Scoring: Exploring the Utility and Potential of the Large Language Models for the WrightRightNow Platform

Cengiz Zopluoglu, University of Oregon

Gerald Tindal, University of Oregon

BRT

behavioral research & teaching

In this technical report, we present initial findings on autoscoring student writing to allow teachers using WriteRightNow to continue assigning writing prompts without facing the burden of frequent scoring.

The University of Oregon is committed to the policy that all persons shall have equal access to its programs, facilities, and employment without regard to race, color, creed, religion, national origin, sex, age, marital status, disability, public assistance status, veteran status, or sexual orientation. This document is available in alternate formats upon request.

Automated Essay Scoring: Exploring the Utility and Potential of Large Language
Models for the WrightRightNow Platform

WriteRightNow (https://writerightnow.com) is an innovative digital platform meticulously crafted to enhance writing instruction across various curricula. Central to its design is the customization of instructional content, allowing for a multi-faceted approach that caters to diverse student needs, including those with special educational requirements and English learners. WRN fosters organizational efficiency, enabling educators to systematically guide students from initial planning stages through to the completion of polished writing assignments. Promoting a collaborative educational environment, WRN facilitates best practices among educators, while encouraging student ownership in learning through engaging review cycles. A distinctive feature of WRN is its dynamic library of writing prompts, designed to stimulate critical thinking and intellectual curiosity across a multitude of subjects. Several other features allow teachers to actively engage students in the writing process, allowing revisions with feedback, providing summative grading, and graphing both qualitative and quantitative results. The software was designed to facilitate improvement in writing by writing.

Despite its innovative features, WRN currently lacks an automated essay scoring system, a shortfall that places a significant grading burden on educators. Teachers find themselves inundated with a multitude of essays, making the grading process labor-intensive and time-consuming. This constraint detracts from the quality time that educators could invest in individualized student instruction and support, potentially compromising the overall educational experience. The integration of an automated essay scoring system within WRN could profoundly optimize this aspect of the educational process. Such a system would alleviate the grading burden on teachers, enabling a more efficient and streamlined evaluation process. It would enhance the platform's functionality, allowing educators to focus more on fostering critical thinking and writing skills in students, rather than being overwhelmed by the administrative task of grading. The incorporation of this feature is essential for bolstering WRN's efficacy and utility as a comprehensive educational tool, driving its evolution towards a more optimized and holistic writing instruction platform.

The rise of large language models (LLMs) has been nothing short of transformative, opening up new possibilities in numerous domains, including education (Bommasani et al., 2021). LLMs offer a promising solution to the challenges faced by the WRN platform in automating essay scoring. LLMs are proficient in processing and generating human-like text, enabling them to evaluate essays with a degree of nuance and understanding. Their inherent capability to analyze syntax, semantics, and context makes them particularly suited for assessing the coherence, relevance, and sophistication of student essays. Incorporating LLMs into WRN's infrastructure would enable the platform to execute automated essay scoring with reasonable accuracy and consistency, mimicking the evaluation process of human educators while significantly reducing their workload. This integration would facilitate a more dynamic and responsive grading system, allowing for real-time feedback that is instrumental in enhancing students' writing and revision processes. Moreover, LLMs can be customized to align with specific grading rubrics and criteria, ensuring that the automated scoring system is both reliable and adaptable to various educational contexts and requirements.

This technical report presents the methodology and findings of research focused on the development of an essay scoring algorithm for the WriteRightNow (WRN) platform. The purpose is to explore the feasibility and utility of automating the essay grading process, thereby enhancing efficiency and reducing the workload of educators. Through rigorous analysis, the report seeks to offer insights into the potential benefits and practical implementation of this technological advancement within WRN, emphasizing its capacity to streamline educational processes and improve the overall teaching and learning experience.

Methodology

Source dataset for developing the scoring engines

The source dataset utilized in this study is derived from a publicly available corpus, the English Language Learning Insight, Proficiency, and Skills Evaluation corpus (ELLIPSE; Crossley et al., in press). The ELLIPSE corpus is a rich and open-source collection of essays, primarily sourced from extensive state-wide and national standardized testing conducted across the United States to evaluate grade-level writing skills. This corpus offers a blend of holistic and analytic scores, coupled with a diverse array of demographic details such as gender, race/ethnicity, and economic backgrounds of the participants. The essays within the corpus exhibit a wide spectrum of word counts, reflecting a normal distribution, and are accompanied by detailed text statistics and scoring distributions. There are 6,482 observations in this dataset. Table 1 below provides a summary of demographic variables available in this dataset.

Table 1

Distribution of Demographic Variables Available in the ELLIPSE Corpus

Categories	Subcategories	Percentage (%)
Gender	Female	43.91
	Male	56.09
Grade	8	25.1
	9	0.49
	10	5.09
	11	35.17
	12	34.14
Race/Ethnicity	American Indian/Alaskan Native	0.35
	Asian/Pacific Islander	12.22
	Black/African American	7.95
	Hispanic/Latino	71.51
	Two or more races/Other	0.71
	White	7.27
Socio-Economic Status (SES)	Missing/Not specified	0.015
	Economically disadvantaged	69.53
	Not economically disadvantaged	30.45

Note: All values are represented as percentages.

Each essay in the ELLIPSE corpus was evaluated based on a comprehensive language proficiency rubric. A minimum of two experienced raters, recruited from a research university with expertise in teaching English as a second language, assessed each essay. In instances where the score difference between raters exceeded one point, a discussion ensued to reconcile discrepancies. The evaluation process culminated in the assignment of a single final score in each of the following six traits: Cohesion, Syntax, Vocabulary, Phraseology, Grammar, and Conventions. Additionally, an overarching Overall score was also assigned. Scores ranged from 1 to 5, with increments of 0.5. A summary of distribution of these assigned scores for each trait is presented in Table 2 below. Table 3 also provides the correlations among the assigned scores in different traits.

Table 2

Descriptive Statistics of Available Essay Scores in the ELLIPSE Corpus

Trait	N	Mean	SD	Min	Max	Skewness	Kurtosis
Cohesion	6482	3.10	0.64	1	5	0.11	-0.07
Syntax	6482	3.13	0.66	1	5	0.04	-0.21
Vocabulary	6482	3.03	0.64	1	5	0.13	-0.08
Phraseology	6482	3.23	0.58	1	5	0.21	0.25
Grammar	6482	3.12	0.65	1	5	0.07	-0.30
Conventions	6482	3.03	0.69	1	5	0.22	-0.38
Overall	6482	3.08	0.67	1	5	0.07	-0.24

Table 3

Correlation Matrix Among Assigned Scores in the ELLIPSE Corpus

Trait	Cohesion	Syntax	Vocabulary	Phraseology	Grammar	Conventions	Overall
Cohesion	1						
Syntax	0.69	1					
Vocabulary	0.66	0.68	1				
Phraseology	0.68	0.72	0.73	1			
Grammar	0.63	0.70	0.65	0.71	1		
Conventions	0.66	0.69	0.66	0.66	0.66	1	
Overall	0.78	0.81	0.79	0.80	0.77	0.77	1

Data Preprocessing

Input Features. We generated the numerical embeddings for each essay in the ELLIPSE corpus using two LLMs, DeBERTaV3 (He et al., 2021) and Longformer ((Beltagy et al., 2020). Numerical embeddings, generated by these two models, are high-dimensional vectors with a length of 768 and this high-dimensional vector encapsulate the semantic essence of the input text. Each dimension in the embedding carries nuanced information, collectively forming a dense representation that captures the syntactic and semantic intricacies of the textual input. In a 768-dimensional vector, each element is a numerical value contributing to a composite, continuous representation of the text, enabling the model to understand and utilize the textual information in downstream tasks such as classification, regression, or other natural language processing applications, facilitating a mathematically quantifiable and analyzable representation of linguistic data. We combined the numerical embeddings from both models for each essay, yielding an input matrix with dimensions 6482 x 1536. Each row in this input matrix represents an essay, and 1536 columns represent the numerical embeddings obtained from DeBERTaV3 and Longformer.

Outcome. The final scores assigned for each trait (cohesion, syntax, vocabulary, phraseology, grammar, conventions, and overall) was treated as an outcome to predict from the input feature matrix. In addition, we also consider creating a composite score from seven different scores using factor analysis, and this was justified due to the high correlations among these traits. The first eigenvalue extracted from the reduced correlation matrix was 4.96, indicating that the first common factor accounts about 71% of the variance. Table 4 below summarizes the output when a single factor model is fit to data from all seven traits with assigned scores. The model fit was reasonably well, and indicated that a common factor score can be formed for each essay based on the assigned scores from seven areas reported in the ELLIPSE corpus. A common factor score (CFS) was computed for each essay based on the single factor solution and appended to the dataset as the eighth outcome for the prediction task. Finally, we standardize each outcome measure such that the mean is equal to 50 and standard deviation is 10. The purpose of this standardization is to make it easier to compare the predictive accuracy results across different traits.

Table 4

Results from Fitting a Single Factor Model to Assigned Scores in ELLIPSE orpus

Traits	Standardized Factor Loadings	Standard Error
Overall	0.954	0.002
Cohesion	0.811	0.005
Syntax	0.849	0.004
Vocabulary	0.822	0.004
Phraseology	0.850	0.004
Grammar	0.813	0.005
Conventions	0.802	0.005
Model Fit		
RMSEA	0.057	
SRMR	0.012	
CFI	0.993	

Train/Test Split. The dataset was randomly split into two pieces such that a training set included 80% of the observations (*N*=5,186) and a test set included the remaining 20% of the observations (*N*=1,296).

Training Models

We implemented a cross-validated penalized regression approach, incorporating a lasso penalty, to construct scoring models for various outcomes. This approach selects the optimal lasso penalty value through a 10-fold cross-validation process, ensuring the model's robustness and accuracy. Consequently, the model that exhibited the most favorable lasso penalty emerged as the ultimate scoring model across various traits, yielding a total of eight distinct models. Seven of these models are dedicated to predicting scores in specific traits as reported in the original ELLIPSE corpus, while an eighth model is designed to predict the common factor scores we created based on a single-factor solution of these seven traits. Each model is engineered to process a numerical embedding vector derived from textual inputs, such as student essays, using DeBERTaV3 and Longformer and produce a prediction for the respective trait. All models are trained using the training set with a 10-fold cross validation, and then the performance of the optimal models is evaluated on the test set.

Results

Model Performance

Table 5 below provides a summary of results and reports the prediction accuracy as measured by the root mean squared error and correlations between observed scores and predicted scores for each trait on the test set. Figure 1 also provides a visual representation of how the predicted scores generated by scoring models relate to the observed scored generated by human raters.
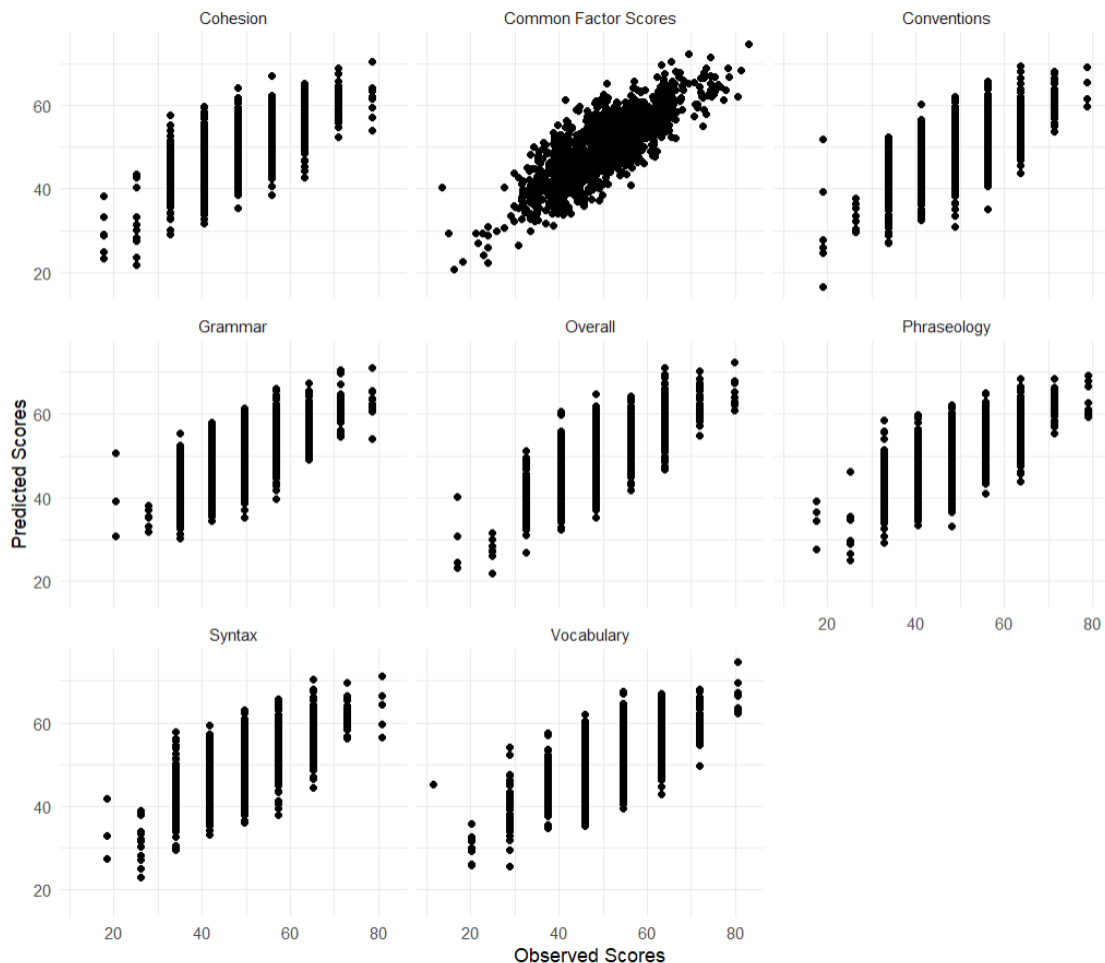
Table 5

Performance Evaluation of the Scoring Models on the Test Dataset (N=1,296)

Trait	RMSE	Correlation
Cohesion	7.35	0.69
Syntax	6.97	0.73
Vocabulary	7.30	0.70
Phraseology	6.92	0.73
Grammar	6.79	0.74
Conventions	7.04	0.73
Overall	6.27	0.79
CFS	5.87	0.82

Figure 1

Scatterplot of Observed Human Scores and Predicted Model Scores Across Different Outcomes in the ELLIPSE Corpus

Fairness Analysis

We performed a fairness analysis based on three demographic variables present in the dataset: sex, race/ethnicity, and socioeconomic status. Initially, we compared the observed scores given by human raters with the predicted scores produced by scoring models for each observation for all traits. Subsequently, we determined the average difference and its standard errors within each demographic group across all traits. Figure 2 illustrates the results of our bias analysis. A vertical line at 0 acts as a reference, indicating a point of no bias. Points on the left of this line suggest a positive bias, meaning the predicted scores are, on average, higher than the observed scores. Conversely, points on the right indicate a negative bias, where the predicted scores are lower than the observed scores on average. From the data visualized in the plot, we observed a slight bias in all categories. However, these biases were not significantly different from zero (95% CI includes zero), implying that the observed biases could have occurred due to random chance.

Figure 2

Prediction Bias by Demographic Variables Across Traits

Average Bias with 95% CI

Utility of Scoring Models

After training our scoring models using the ELLIPSE corpus, we evaluated their effectiveness and applicability to essays on the WRN platform. As of August 10, 2023, the WRN platform contained around 8,000 essays. These were submitted by 2,503 students from 267 classrooms, responding to 1,235 unique prompts over approximately four years. It's noteworthy that only 2,152 essays (26.9%) had been graded by teachers, highlighting the substantial grading workload on educators. For a more focused analysis, we limited our evaluation to essays containing a minimum of 100 characters. This approach excluded very brief essays that lacked sufficient data for reliable evaluation and scoring. Applying these restrictions reduced the number of scorable essays on the platform to 5,629. These essays were contributed by 2,165 students across 208 classrooms, with only 27.4% having been graded by teachers.

The WRN platform doesn't enforce a single grading scheme. Instead, it offers teachers the flexibility to select from nine different grading schemes when assigning writing tasks. Table 6 delineates the distribution of graded essays with at least 100 characters across each available grading scheme. A utility analysis was feasible for two grading schemes due to the availability of a substantial number of graded essays: 6 Trait and Holistic. For these two grading schemes, we applied the scoring models, which were initially trained on the ELLIPSE corpus, to evaluate the essays. Subsequently, we compared the model-generated scores with the grades assigned by teachers, as recorded in the system. A summary of our findings for each grading scheme is presented below.

Table 6

Frequency of Written Essays in the WRN Platform by Grading Scheme Chosen by Teachers

	Grade Assigned	
Grading Scheme	No	Yes
6 Trait	1906	812
Holistic	187	443
Letter Grade	187	189
Oregon Essential Skills Requirement	157	27
WRN CSSS Argumentative	94	22
WRN CCSS Informative	42	37
WRN CCSS Narrative	78	11
No Grade	907	0

Note: CCSS reflects Common Core State Standards

Holistic Grading Scheme

In the holistic grading scheme, teachers assigned scores ranging from 1 to 6 to student essays, with a detailed rubric available in Appendix. We evaluated the correlation between two sets of scores: the model-generated scores in the Common Factor Scores and the teacher-assigned holistic grades. The polyserial correlation of 0.55 was observed.

6 Trait Grading Scheme

In the 6 Trait grading scheme, teachers evaluated student essays in six distinct areas: Ideas and Content, Organization, Voice, Word Choice, Sentence Fluency, and Conventions, assigning scores from 1 to 6 in each area. A comprehensive rubric for these areas is accessible in Appendix. We conducted a comparison between the teacher-assigned grades in each area of the 6 Trait Grading Scheme and the model-generated scores across eight traits we trained models. Table 7 presents the polyserial correlations resulting from these comparisons, offering a detailed insight into the relationship between the two sets of scores.

Table 7

Polyserial Correlation between Teacher Assigned Scores for 6 Trait Categories and Model generated Scores for Traits

	Model Generated Scores							
	Cohesion	Syntax	Vocabulary	Phraseology	Grammar	Conventions	Overall	CFS
6 Trait Categories								
Ideas and Content	0.47	0.48	0.45	0.46	0.42	0.41	**0.49**	0.47
Organization	0.46	0.46	0.42	0.45	0.42	0.41	**0.48**	0.46
Voice	0.44	**0.46**	0.41	0.44	0.41	0.37	0.46	0.45
Word Choice	0.47	0.48	0.46	0.46	0.42	0.42	**0.48**	0.47
Sentence Fluency	0.51	**0.53**	0.49	0.50	0.48	0.48	0.52	0.52
Conventions	0.56	**0.58**	0.52	0.54	0.52	0.54	0.56	0.56

Note. CFS refers to the common factor scores obtained from the other seven traits.

Discussion

In this report, we presented findings from our attempt to create models capable of automatically scoring essays written by English Language Learners (ELLs). The models were initially trained using data from the ELLIPSE corpus, consisting of essays from ELLs. Our preliminary evaluations were promising, showing correlations between the human scoring and machine scoring scores ranging from 0.69 to 0.82 on a test dataset from the ELLIPSE corpus. Notably, the model that predicted a common factor score of overall writing quality across seven traits in the ELLIPSE corpus demonstrated the highest correlation and accuracy. This outcome was anticipated because the common factor scores, derived from a single-common factor model across all available seven scores, captured the most reliable variance, minimizing noise, and therefore yielding better predictions. Moreover, our fairness analysis didn't show any systematic bias against any demographic groups present in the ELLIPSE corpus dataset.

In extending our evaluation, we assessed the models' applicability and external validity by scoring essays on the WRN platform. These essays were written by students similar to the group in the ELLIPSE corpus, but in a different context. Despite having a limited number of essays with teacher-assigned grades on the WRN platform, we found a moderate average correlation of 0.49 between the teacher-assigned grades and the model-generated scores across various grading schemes.

However, this analysis comes with several challenges and limitations. Firstly, the WRN platform features a more diverse and varied prompt library compared to the ELLIPSE corpus dataset. Secondly, the WRN platform utilizes a multitude of scoring schemes and rubrics, differing from those in the ELLIPSE corpus dataset, causing a certain degree of misalignment despite some similarities. Lastly, the WRN platform's grading lacks stringent standardization without any formal training for grading, with essays typically graded by only one teacher, which could introduce more measurement errors and statistical noise. Given these considerations, we find a correlation of 0.49 between the teacher-assigned grades and machine-generated scores on an external dataset in a different context to be a promising indication of the models' potential effectiveness.

References

Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer* (arXiv:2004.05150). arXiv. http://arxiv.org/abs/2004.05150

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [Cs]*. http://arxiv.org/abs/2108.07258

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, B., Picou, A., & Boser, U. (in press). Measuring second language proficiency using the English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. *International Journal of Learner Corpus Research*.

He, P., Gao, J., & Chen, W. (2021). *DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing* (arXiv:2111.09543). arXiv. http://arxiv.org/abs/2111.09543

APPENDIX

Grading rubric for the Holistic Grading Scheme

Grade	
1	The response is rudimentary, contains little detail, and is unorganized. The response contains too little information to evaluate with few unrelated details, little structure, no transitions, and no consideration of key concepts/vocabulary or the intellectual operation in the prompt.
2	The response is missing important information (details) and is not well organized. The response begins with little direction and is not elaborated, with much detail missing and no conclusion. The response reflects only occasional reference to specific information or details that are marginally related to each other. The response has few transitions and the sentence structure fails to reflect the key concepts/vocabulary and intellectual operation in the prompt.
3	The response has most of the essential information but is roughly organized. The response begins with a vague direction and provides little elaboration, concluding with little preparation. The response provides general information that occasionally references few details but they are only somewhat related to each other. The response has rough transitions and the sentence structure is only vaguely connected to the key concepts/vocabulary and intellectual operation in the prompt.
4	The response is fairly complete with information that is detailed and organized. Though the response begins with a sense of direction, the follow up and conclusion are not tightly connected. The response includes appropriate information and some details that relate to each other. The response has broad transitions and uses sentence structure that reflects attention to key concepts/vocabulary and the intellectual operation in the prompt.
5	The response is very complete with information that is well detailed and well organized. The response begins with a clear direction that is further elaborated and ends with an appropriate conclusion. The response reflects relevant content, both in the specific information and the manner in which the details relate to each other. The response has appropriate transitions and uses excellent sentence structure to appropriately address key concepts/vocabulary as well as the intellectual operation noted in the prompt.
6	The response is complex with information that is exceedingly detailed and expertly organized. The response begins with a compelling direction that is richly elaborated and ends with a resonating conclusion. The response reflects highly appropriate content, both in the specific information and the manner in which the details build upon each other. The response uses transitions effectively and maintains perfect sentence structure to successfully address key concepts/vocabulary as well as the intellectual operation noted in the prompt.

Grading rubric for the 6 Trait Grading Scheme

Grade	Ideas and Content	Organization	Voice	Word Choice	Sentence Fluency	Conventions
1	Main ideas do little to promote the purpose or central theme. Content including key ideas and details is unclear, or missing.	Organization of content is random or unclear, and a lack of structure continually distracts the reader from the main ideas of the text. Very substantial reorganization is critical to remove unnecessary portions and to add-in necessary elements.	The reader cannot identify the author's voice, which seems to be missing or uninvolved. The author needs to make a much greater effort to apply an appropriate voice to the writing.	There is a lack of variation in word choice, and/ or limited vocabulary words.	Sentences lack fluency; writing is stilted, choppy, or awkward, which distracts the reader.	Many errors in the use of conventions require extensive proofreading and editing, and distract from the main ideas of the text.
2	The main ideas are emerging, and show some clarity. A few key ideas and details support the main ideas, but many key details are missing.	Writing demonstrates an emerging sense of organization, but most details are presented randomly, in a way that is distracting to the reader and does little to move the writing along. Substantial reorganization is recommended to remove unnecessary portions and to	The reader can begin to identify the author's voice, which seems to be emerging. The author needs to make a greater effort to apply an appropriate voice to the writing.	Writer demonstrates an emerging sense of variation in word choices/ vocabulary.	Sentence fluency is emerging; writing may be somewhat awkward, which can distract the reader.	Some errors in the use of conventions require proofreading and editing, and may, at times, distract the reader from the main ideas of the text.

		add-in necessary elements.				
3	Main ideas and content are developing, and are basically clear. Writing demonstrates a basic understanding of the topic, with some key ideas and details, though some details are missing.	Writing demonstrates a developing sense of overall organization, and has the presence of a basic structure including the main idea and some key details, which may be irrelevant at times, and/ or there may be some gaps in the information presented. Organization requires substantial editing.	Writing demonstrates a basic sense of the author's voice, but it could more strongly compel the reader. The author needs to more carefully consider elements that would support a compelling voice.	Author demonstrates basic word choices/ vocabulary, which are functional and clear, but do little to engage the reader.	Text has a developing sense of sentence fluency, but requires further editing to make writing sound more natural.	Writing demonstrates a basic command of conventions, though a few errors may distract the reader, requiring significant proofreading and editing.
4	Main ideas and content are clearly stated, and an adequate amount of supporting information is present, but there is room for clarification of details to support key ideas.	Writing demonstrates a proficient sense of overall organization, and has the presence of a solid structure including the main idea and an adequate number of relevant key details, although the writing could use more relevant, substantial support to promote the	Writing demonstrates a proficient sense of the author's voice. To improve a sense of voice, the author should carefully consider additional elements that would support a compelling voice.	The author uses appropriate word choices/ vocabulary to develop proficient writing.	Text has a proficient sense of sentence fluency, but a few sentences may require further editing to make writing sound more natural.	Proficient use of conventions, requiring minor proofreading and editing.

		main ideas. Organization requires some editing.				
5	Main idea is clearly illustrated and gives the reader a strong sense of the topic. Relevant details support a clear, focused topic, supported by key ideas that relate to the topic in a compelling way.	Strong organization facilitates the main idea supported with strongly relevant key ideas, and the writing structure effectively guides the reader through the text. Organization only requires minor editing.	Writing demonstrates a clear, strong sense of the author's voice. Key details are presented in interesting ways, thoughtfully structured to demonstrate the author's voice.	The author uses strategic, relevant word choices/ vocabulary to develop a strong, engaging written piece.	Expressive language and thoughtfully constructed sentence structure demonstrate a strong sense of sentence fluency.	Strong command of conventions leads the reader through the text, and only very minor errors are present.
6	Main idea is well-established and gives the reader an exemplary sense of the topic. Carefully chosen, relevant details support a clear, focused topic, supported by key ideas that relate to the topic in a highly compelling way.	Exemplary organization invites the reader into the main idea, supported by strongly relevant key ideas, organized in a highly logical way. The writing structure effectively guides the reader through the text. Organization requires little to no editing.	Author's voice is expertly tailored to deliver engaging content especially for a specific audience. Key details are presented in highly compelling ways, expertly tailored to demonstrate the author's voice.	The author uses highly relevant, illustrative word choices/ vocabulary to develop highly compelling writing.	Every sentence demonstrates expressive language and thoughtfully constructed sentence structure, giving an overall exemplary sense of sentence fluency.	Exemplary use of conventions, which expertly maintains the reader's focus on the main ideas. Writing is ready for publication.