

When Language Becomes a Shortcut in Automated Short-Answer Scoring

Xiaomeng Xiong Corinne Huggins-Manley Jinnie Shin Christan Grant

University of Florida

xiaomengxiong@ufl.edu amanley@coe.ufl.edu

jinnie.shin@coe.ufl.edu christan@ufl.edu

Abstract

Automated short-answer scoring should primarily reflect the substantive information expressed in a response rather than how that information is worded. This study examines whether automated short-answer scoring (ASAS) models are influenced by linguistic form when score-relevant content is preserved. We propose a controlled rewriting framework that systematically varies linguistic complexity while preserving the information targeted by the scoring rubric, enabling within-response tests of construct-irrelevant linguistic influence. Using 743 responses from ASAP-SAS Set 1, we generated three controlled rewrite levels and evaluated 6 fine-tuned encoders and 11 LLMs. Fine-tuned encoders showed consistently positive associations, particularly for lower-scoring responses, suggesting reliance on construct-irrelevant linguistic cues consistent with shortcut learning. These associations persisted after adjustment for observed linguistic changes, indicating that the pattern may involve higher-level linguistic properties.

LLMs were more heterogeneous. GPT-4.1, GPT-5.1, and Nemotron-3-Super showed negative associations with increased linguistic complexity, contrasting with the commonly reported tendency to favor verbosity. Among other models with relatively high agreement with human ratings, Llama-3.3 showed a positive association, whereas Claude Sonnet 4.6, Gemma-4, and GPT-OSS-120B showed no significant association. These divergent patterns, together with their differing responses to adjustment for observed linguistic changes, suggest that different LLMs may rely on different scoring signals.

1 Introduction

In short-answer scoring, the target construct is the correctness of substantive meaning, not writing style (Burrows et al., 2015). This study examines the construct validity of automated short-answer

scoring (ASAS) by testing whether construct-irrelevant linguistic variation influences predicted scores. From Messick's construct validity perspective, scores should primarily reflect the target psychological trait or ability being measured rather than extraneous factors; systematic influence from non-target factors may therefore threaten construct validity (Messick, 1995). In automated scoring, although models can achieve high agreement with human ratings, it remains unclear whether their decisions are truly based on the intended measurement construct. *The Standards for Educational and Psychological Testing* further emphasizes that score interpretations should be supported by evidence related to the target construct (American Educational Research Association et al., 2014). Therefore, agreement with human ratings alone is insufficient to support claims of construct validity, as it does not directly validate the inferences underlying score interpretation (Kane, 2013).

If model predictions vary systematically across linguistically different but semantically equivalent responses, this suggests that models may rely on surface-level linguistic features that are construct-irrelevant. In other words, models may exploit features that are predictive in the training data but not aligned with the intended construct, consistent with the phenomenon of shortcut learning observed in deep learning systems (Geirhos et al., 2020). Such behavior introduces construct-irrelevant variance into test scores, raising concerns about construct validity and fairness (Johnson and McCaffrey, 2023).

This study introduces a controlled perturbation framework that systematically rewrites the same response while preserving semantic content and progressively increasing linguistic complexity. As illustrated in Figure 1, the framework proceeds in three stages: rewriting each response into controlled linguistic variants, scoring all variants with a range of automated systems, and estimating within-response differences in predicted scores. The scor-

ing systems include multiple encoder-based models and generative AI models.

By holding semantic content constant by construction, this within-response design enables a direct test of whether variation in linguistic form influences predicted scores, thus offering a stricter test of construct validity in ASAS. Moreover, the framework provides a generalizable tool for interpretability and potential fairness evaluations in ASAS, where linguistic features are irrelevant to the construct being assessed.

2 Related Work

Automated scoring systems have gradually evolved from feature-engineering-based approaches to Transformer-based pretrained language models, including both encoder-based architectures and generative LLMs, which are more effective at modeling contextual semantics and long-range dependencies. However, despite substantial improvements in performance metrics, prior research has shown that these models still exhibit limitations in robustness and construct alignment. Some prior work has evaluated model behavior by constructing adversarial examples (e.g., Lottridge et al., 2021; Kabra et al., 2022; Kumar et al., 2020). For instance, Lottridge et al. (2021) examined several gaming response types, including shuffled word order, grammatically correct but nonsensical (Babel) essays, off-topic essays, and duplicated text and found that both feature-based and deep-learning (BERT) engines were not robust to these manipulations, with response length contributing substantially to predicted scores. However, these manipulations either alter the meaning of the response or produce degraded, clearly anomalous text that human raters would readily flag. They therefore test whether models detect aberrant inputs, rather than whether linguistic form influences scores within the range of fluent, meaning-preserving responses. The latter has received little systematic attention.

Fairness is a further concern. Existing research on fairness in automated scoring has largely focused on group-level disparities, typically relying on background variables (e.g., Andersen et al. (2025), Johnson and McCaffrey (2023)). While informative, they rely on observational group comparisons and cannot disentangle construct-relevant from construct-irrelevant sources of variation at the design level.

To address these gaps, the present study adopts

a controlled perturbation design, in which linguistic expression is systematically manipulated while holding semantic content constant. This approach constructs counterfactual responses that are meaning-equivalent but differ in form, enabling within-response comparisons that fix construct-relevant factors by design. As a result, it directly tests whether linguistic expression, a construct-irrelevant factor, systematically influences model scores, without relying on statistical estimation of latent ability. This design extends the analysis of fairness from group-level disparities to within-response linguistic variation, as well as from correlational comparisons of scores across groups to causal comparisons of scores within responses, providing a more direct identification of construct-irrelevant variance.

If such effects are observed, they would suggest that ASAS systems exhibit implicit preferences for certain linguistic resources, favoring responses that are more complex or more "standard" in form. This issue is not merely a matter of test-taking strategies or "gaming" behavior, but rather reflects construct-irrelevant variance in the scoring process. It also raises concerns about measurement fairness, as disparities in access to linguistic resources across students may translate into systematic differences in model-assigned scores that are supposed to reflect content knowledge only.

3 Methodology

This study applies controlled linguistic rewrites to the same set of student responses and compares how different automated short-answer scoring models behave across the resulting language variants. The research questions are as follows:

RQ1. How is rewrite intensity associated with within-response score changes across models?

RQ2. How does this association change after accounting for observed linguistic changes?

RQ3. Does the association between rewrite intensity and score change vary by the original human score?

3.1 Dataset and Analytic Sample

Data were drawn from Set 1 of the ASAP-SAS dataset (Barbara et al., 2012), which contains 1,672 student responses scored on a 0 to 3 scale. The H1 score distribution was 380 responses scored 0 (22.7%), 429 scored 1 (25.7%), 524 scored 2 (31.3%), and 339 scored 3 (20.3%). Dataset docu-

mentation identifies H1 as the final score and H2 as a second rating for inter-rater reliability; H1 therefore served as the primary human reference in this study.

The item is situated in a science context involving an acid rain experiment and asks students to identify information missing from the procedure that would be needed to replicate the experiment. The item-specific rubric assigns scores according to the number of relevant pieces of information identified. Responses identifying three, two, one, or little to no relevant information receive scores of 3, 2, 1, and 0, respectively. Thus, scoring for this item is based directly on the experimental information contained in the response rather than on linguistic complexity or writing style.

To support a complete within-response comparison, we retained only responses for which the original response and all three rewrite versions, LOW, MEDIUM, and HIGH, met the prespecified quality criteria. All 1,672 responses entered the rewriting procedure. If any rewrite failed the quality checks or if the three rewrite versions did not satisfy the prespecified ordering of rewrite intensity, the response was excluded from the final rewrite analysis sample. A total of 743 responses met all retention criteria, representing 44.4% of the original sample. Their H1 score distribution was 195 responses scored 0 (26.2%), 188 scored 1 (25.3%), 226 scored 2 (30.4%), and 134 scored 3 (18.0%).

Retained responses and responses not included in the final rewrite analysis sample differed little in H1 score (1.40 vs. 1.56; standardized difference = -0.15), but retained responses were shorter on average (42.5 vs. 51.5 words; standardized difference = -0.41). Accordingly, we interpret the findings as model behavior observed among responses that met the prespecified rewrite quality criteria rather than as findings that necessarily generalize to all responses in Set 1.

3.2 Controlled Linguistic Rewriting

The framework generates graded linguistic perturbations under explicit constraints designed to preserve substantive meaning. For each student response, three rewritten versions with different levels of intensity (LOW, MEDIUM, HIGH) are constructed to examine the sensitivity of automated scoring models to surface linguistic features.

In preprocessing, only formatting noise (e.g., bullet markers) is removed, as preliminary experiments showed that such structural artifacts con-

strain effective sentence-level rewriting.

For each original response, three rewritten versions are generated independently using GPT-4.1-mini (non-chained generation) to avoid dependency across levels and error propagation. The model is explicitly instructed to act as a linguistic rewriter, modifying only language while preserving all original information under a set of global constraints (see Appendix C). Specifically, the prompt prohibited adding or removing information, changing any numerical information, or replacing unclear expressions with inferred information.

The three rewrite levels are clearly defined and required to satisfy an ordered intensity structure ($LOW < MEDIUM < HIGH$). At the LOW level, rewriting involves minimal edits such as correcting spelling, grammar, and punctuation, while largely preserving the original sentence structure, corresponding to local polishing. At the MEDIUM level, the rewrite introduces noticeable sentence-level changes, improving clarity and fluency through moderate syntactic reorganization and more precise lexical choices. At the HIGH level, the rewrite involves substantial structural transformation, adopting more formal and academic expressions and producing denser information representation, for example through subordination and nominalization.

To stabilize rewriting behavior across levels, multiple few-shot examples are incorporated into the prompt, explicitly illustrating the mapping from original responses to each rewrite level (see Appendix C). Table 1 presents outputs generated by the proposed framework.

Level	Response
Original	They should label the materials. Makes sure the scale is calibrated. Make sure you have a few days to do this lab.
LOW	They should label the materials. Make sure the scale is calibrated. Also, ensure you have a few days to complete this lab.
MEDIUM	They should label the materials, ensure the scale is calibrated, and allow a few days to complete the lab.
HIGH	The materials should be clearly labeled, and the scale must be properly calibrated. Additionally, it is essential to allocate several days to complete this laboratory procedure.

Table 1: Example of controlled linguistic perturbations across intensity levels.

The item-specific rubric assigns scores accord-

ing to the number of relevant missing information points identified in the response rather than the quality of linguistic expression itself. Therefore, provided that these score-relevant information points are preserved, changes in wording, syntax, or formality should not alter the score assigned under the original rubric. To further limit potential changes in substantive meaning and regulate rewrite intensity, each rewritten response was required to satisfy the following quality-control constraints:

- (1) all numerical information must be preserved exactly, to avoid altering the factual content of the response;
- (2) semantic similarity had to meet the level-specific minimum threshold, computed using the GTE sentence-embedding model (Li et al., 2023). The thresholds were 0.92, 0.90, and 0.88 for LOW, MEDIUM, and HIGH, respectively. Rewrites falling below the corresponding threshold were rejected and regenerated;
- (3) rewrite intensity is controlled via a change ratio to ensure clear separation across the LOW, MEDIUM, and HIGH levels;
- (4) length and newly introduced wording are bounded relative to the original, excluding degenerate cases such as large omissions or substantial expansion.

The generation follows an iterative generate–evaluate–regenerate procedure: each rewrite is automatically checked against these constraints, and if any is violated, targeted feedback is provided and the rewrite is regenerated, for up to six iterations.

3.3 ASAS Models

We evaluated 17 automated short-answer scoring models, including six encoder models with task-specific fine-tuning and 11 large language models (LLMs) that scored responses directly using the scoring rubric (Table 2).

The encoder models were trained using the 1,672 original responses and their H1 scores. The generated LOW, MEDIUM, and HIGH rewrites were used only for evaluation and were not included in model training. To obtain out-of-sample predictions for the 743 responses in the final analytic sample, we used five-fold cross-fitting, with the 743 responses stratified by H1. In each fold, the original responses in the current evaluation fold were removed from the 1,672 labeled original responses, and the remaining responses were used for training and validation. Ten percent of the remaining

responses were held out as a validation set for early stopping. The trained model then scored the original responses in the evaluation fold and their three rewritten versions. Predictions from the five folds were combined to obtain out-of-sample scores for all 743 original responses and 2,229 rewrites. This procedure ensured that each original response and its rewritten versions were scored by a model that had not been trained on that original response. Encoder models used a maximum input length of 512, a learning rate of 2×10^{-5} , and up to six training epochs with early stopping.

The generative large language models used a common zero-shot scoring procedure. For each response, the model received the item prompt together with the scoring rubrics provided by ASAP-SAS and assigned a score from 0 to 3. The scoring prompt did not include scored examples. Original responses and all three rewrite versions were scored independently using the same prompt. The complete scoring prompt is provided in Appendix D.

All LLMs were run under deterministic or near-deterministic settings. For models that supported explicit temperature control, temperature was set to 0. GPT-4.1 and GPT-5.1 used the default interface settings. Internal reasoning was disabled for the Nemotron models, while the GPT-OSS models used the low reasoning effort setting.

3.4 Statistical Analysis

To identify the effect of rewriting, we estimate within-response models comparing each rewritten response to its original version. For each response i , the change in predicted score is defined as

$$\Delta Y_i^{(l)} = Y_i^{(l)} - Y_i^{(0)}.$$

We estimate three specifications that share a common baseline and separately introduce linguistic and interaction terms:

(1) *Baseline*:

$$\begin{aligned} \Delta Y_i^{(l)} = & \beta_0 + \beta_1 \text{RewriteLevel}_l + \beta_2 \text{HumanScore}_i \\ & + u_{0i} + u_{1i} \text{RewriteLevel}_l + \epsilon_i^{(l)} \end{aligned} \quad (1)$$

(2) *+ Linguistic features*:

$$\begin{aligned} \Delta Y_i^{(l)} = & \beta_0 + \beta_1 \text{RewriteLevel}_l + \beta_2 \text{HumanScore}_i \\ & + \sum_k \delta_k \text{Ling}_{ik}^{(l)} + u_{0i} \\ & + u_{1i} \text{RewriteLevel}_l + \epsilon_i^{(l)} \end{aligned} \quad (2)$$

(3) + *Interaction*:

$$\begin{aligned}\Delta Y_i^{(l)} = & \beta_0 + \beta_1 \text{RewriteLevel}_l + \beta_2 \text{HumanScore}_i \\ & + \beta_3 (\text{RewriteLevel}_l \times \text{HumanScore}_i) \\ & + u_{0i} + u_{1i} \text{RewriteLevel}_l + \epsilon_i^{(l)}\end{aligned}\quad (3)$$

where $\Delta Y_i^{(l)}$ is the change in predicted score for response i at rewrite level l relative to its original version; RewriteLevel_l indexes rewrite intensity (low, medium, high); $\text{Ling}_{ik}^{(l)}$ denotes the k -th linguistic feature; HumanScore_i is the human-assigned score of the original response; u_{0i} is a response-level random intercept; u_{1i} is a response-level random slope allowing the rewrite effect to vary across responses; and $\epsilon_i^{(l)}$ is the residual.

Specification (1) estimates the overall rewrite effect; Specification (2) adds linguistic features to assess whether this effect can be explained by observable surface characteristics; and Specification (3) adds an interaction with the human score to test whether the rewrite effect varies with response quality. Moreover, the linguistic features capture rewrite-induced changes and surface-level textual properties, including sentence length, readability, lexical diversity, syntactic composition, and punctuation. To reduce multicollinearity among these features, we computed variance inflation factors (VIF) and iteratively removed the feature with the highest VIF until all retained features had VIF below 10.

4 Evaluation

4.1 Validation of Rewrite Manipulation

The effectiveness of the rewriting procedure is evaluated using both semantic similarity and a range of linguistic features. Semantic similarity remains consistently high and decreases gradually with rewrite intensity (LOW = 0.990, MEDIUM = 0.979, HIGH = 0.952), as shown in the distributions in Figure 2: low-level rewrites concentrate near 1.0, while medium and high levels show progressively broader, lower distributions. Even at the highest level, similarity remains high overall, indicating that the rewriting procedure produced progressively stronger linguistic changes while remaining within the prespecified semantic constraints.

Changes in linguistic features generally increased with rewrite intensity (Table 3). Higher-intensity rewrites showed longer average word length, greater readability difficulty, higher lexical diversity, and more subordination. In contrast,

total word count changed very little, with the HIGH condition increasing by only 0.32 words on average, indicating that the rewriting was not primarily achieved by simply expanding response length. Several lexical and syntactic changes were concentrated in the HIGH condition, further supporting the intended increase in linguistic modification across the three rewrite levels.

4.2 Agreement with Human Ratings

Before examining the effects of rewriting, we first evaluated each model’s agreement with human ratings on the original responses (Table 4), using agreement between the two human raters (H1–H2) as a reference. The encoder models showed consistently high agreement, with QWK values ranging from 0.845 to 0.867, and small SMDs. LLMs showed substantially greater variation, with QWK values ranging from 0.418 to 0.803. Nemotron-3-Super performed nearly identically to the best proprietary model, Claude Sonnet 4.6, while Llama-3.3 and Gemma-4 both exceeded GPT-5.1 and GPT-4.1. The lowest agreement was observed for Nemotron-3-Nano, followed by Granite-3.3.

4.3 Main Effects of Rewriting

The encoder models showed consistently positive responses to increasing rewrite intensity (Table 5 and Table 6). In the baseline specification, all Rewrite Level coefficients were positive, ranging from 0.013 for MPNet to 0.044 for ELECTRA. Five of the six coefficients were statistically significant, with MPNet as the only exception.

LLMs showed substantially greater heterogeneity in both direction and magnitude. GPT-5.1 had the largest negative coefficient at -0.088 . GPT-4.1, Nemotron-3-Super, Mistral-Small-3.1, and Granite-3.3 also showed significant negative coefficients, ranging from -0.018 to -0.043 . In contrast, Llama-3.3 showed the largest significant positive coefficient at 0.057. The coefficients for the remaining five models, Claude Sonnet 4.6, Gemma-4, GPT-OSS-120B, GPT-OSS-20B, and Nemotron-3-Nano, were not statistically significant. Overall, the range of coefficients was substantially wider for LLMs (-0.088 to 0.057) than for fine-tuned encoder models (0.013 to 0.044).

4.4 Linguistic Adjustment

We then added 14 linguistic change measures computed relative to each response’s original version. These were retained from 17 candidate measures

through iterative VIF screening using a threshold of 10.

For all encoder models, the Rewrite Level coefficients remained positive and statistically significant after the linguistic change measures were added, ranging from 0.040 to 0.076. The coefficients were generally larger than in the baseline models. For example, the BERT coefficient increased from 0.024 to 0.076.

LLMs showed greater heterogeneity. The significant negative associations observed for GPT-4.1, GPT-5.1, Mistral-Small-3.1, and Granite-3.3 in the baseline models were attenuated and no longer statistically significant after the linguistic change measures were added. In contrast, the Rewrite Level coefficient for Nemotron-3-Super changed from a significant negative association of -0.028 to a significant positive association of 0.043 . GPT-OSS-20B and Llama-3.3 also showed significant positive coefficients of 0.070 and 0.053 , respectively. The remaining models did not show statistically significant Rewrite Level coefficients.

4.5 Heterogeneity in Within-Response Effects

All six encoder models showed significant negative Rewrite Level $\times$ Human Score interactions, ranging from -0.028 to -0.071 . This pattern indicates that the positive association between rewrite intensity and score change tended to weaken as the original human score increased. LLMs showed no consistent interaction pattern. Only Nemotron-3-Nano and Llama-3.3 showed significant positive interactions, with coefficients of 0.053 and 0.036 , respectively, in the opposite direction from the encoder models.

Because scores were bounded between 0 and 3, we conducted a sensitivity analysis to reduce direct floor and ceiling constraints. The interaction models were re-estimated using only responses for which each scoring model assigned an original predicted score of 1 or 2, allowing movement in both directions (Table 7). Among the encoder models, the negative interactions remained similar for DistilBERT, BERT, RoBERTa, and ELECTRA, while the coefficients for MPNet and DeBERTa-v3 were more substantially attenuated. LLMs remained heterogeneous, with the positive interactions for Nemotron-3-Nano and Llama-3.3 persisting in the interior sample.

Overall, direct floor and ceiling constraints may account for part of the interaction pattern observed among the encoder models, but they do not fully

explain the differences across scoring models.

5 Discussion

The results indicate that agreement with human ratings on original responses does not by itself characterize how scoring models respond to controlled linguistic rewriting.

Encoder-based models showed consistent positive rewrite effects that were stronger for lower-scoring responses. The results indicate that encoder predictions reflect not only content correctness but also construct-irrelevant linguistic form, functioning as a kind of shortcut signal, consistent with prior evidence that fine-tuned scoring models can exploit spurious linguistic cues (Filighera et al., 2024). Because the positive association persisted after adjustment for length, readability, and lexical diversity, it does not appear to be fully attributable to these observed surface features and may involve higher-level properties such as discourse organization or overall fluency.

In contrast, LLMs showed substantially greater heterogeneity in both direction and magnitude, suggesting that generative scorers do not form a behaviorally uniform category. GPT-4.1, GPT-5.1, and Nemotron-3-Super responded negatively to rewriting. This contrasts with prior reports of verbosity bias, where longer or more verbose responses can receive more favorable judgments (Zheng et al., 2023; Dubois et al., 2024). This pattern suggests that some LLMs may instead penalize unnecessary complexity: when additional lexical and syntactic complexity introduces no new score-relevant information, they may favor more direct short-answer responses. However, this response was model-dependent. Llama-3.3 responded positively to rewriting. Notably, it achieved QWK comparable to or higher than the proprietary GPT models, suggesting that these contrasting rewrite responses are not simply explained by overall scoring performance but may reflect differences in the scoring signals used by different models.

By combining controlled perturbation with within-response comparisons, this study provides a more direct way to examine construct-irrelevant influences on automated scores, reveals distinct response patterns across fine-tuned encoder and generative models, and offers a framework for assessing construct validity in automated short-answer scoring.

References

- American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.
- Nico Andersen, Julia Mang, Frank Goldhammer, and Fabian Zehner. 2025. Algorithmic fairness in automatic short answer scoring. *International Journal of Artificial Intelligence in Education*, pages 1–38.
- Barbara, Ben Hamner, Jaison Morgan, lynnvandev, and Mark Shermis. 2012. The hewlett foundation: Short answer scoring. <https://kaggle.com/competitions/asap-sas>. Kaggle.
- Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25(1):60–117.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Anna Filighera, Sebastian Ochs, Tim Steuer, and Thomas Tregel. 2024. Cheating automatic short answer grading with the adversarial usage of adjectives and adverbs. *International journal of artificial intelligence in education*, 34(2):616–646.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.
- Matthew S Johnson and Daniel F McCaffrey. 2023. Evaluating fairness of automated scoring in educational measurement. In *Advancing natural language processing in educational assessment*, pages 142–164. Routledge.
- Anubha Kabra, Mehar Bhatia, Yaman Kumar Singla, Junyi Jessy Li, and Rajiv Ratn Shah. 2022. Evaluation toolkit for robustness testing of automatic essay scoring systems. In *Proceedings of the 5th Joint International Conference on Data Science & Management of Data (9th ACM IKDD CODS and 27th COMAD)*, pages 90–99.
- Michael T Kane. 2013. Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1):1–73.
- Yaman Kumar, Mehar Bhatia, Anubha Kabra, Jessy Junyi Li, Di Jin, and Rajiv Ratn Shah. 2020. Calling out bluff: attacking the robustness of automatic scoring systems with simple adversarial testing. *arXiv preprint arXiv:2007.06796*.
- Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.
- Sue Lottridge, Ben Godek, Amir Jafari, and Milan Patel. 2021. Comparing the robustness of deep learning and classical automated scoring approaches to gaming strategies. Technical report, Cambium Assessment Inc.
- Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9):741.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Framework Overview

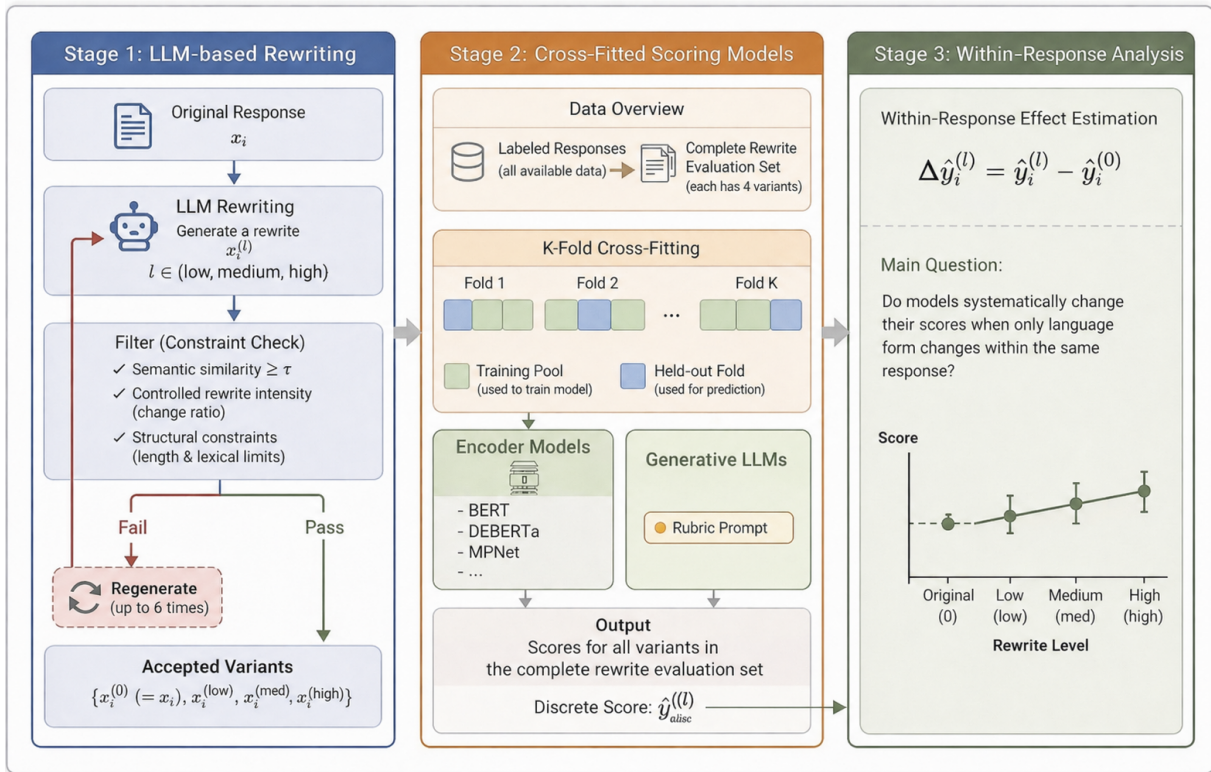


Figure 1: Overview of the controlled perturbation framework. The diagram was created with the assistance of an AI image-generation tool (Gemini’s Nano Banana image model).

B Figures and Tables

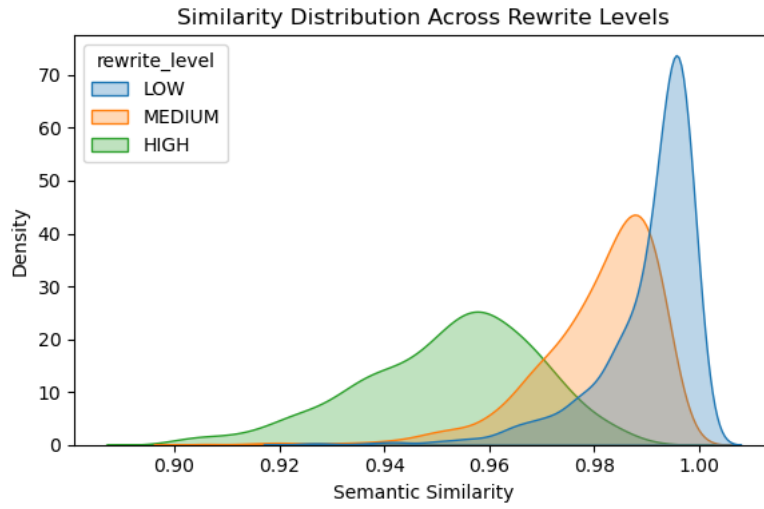


Figure 2: Distribution of semantic similarity scores across rewrite levels.

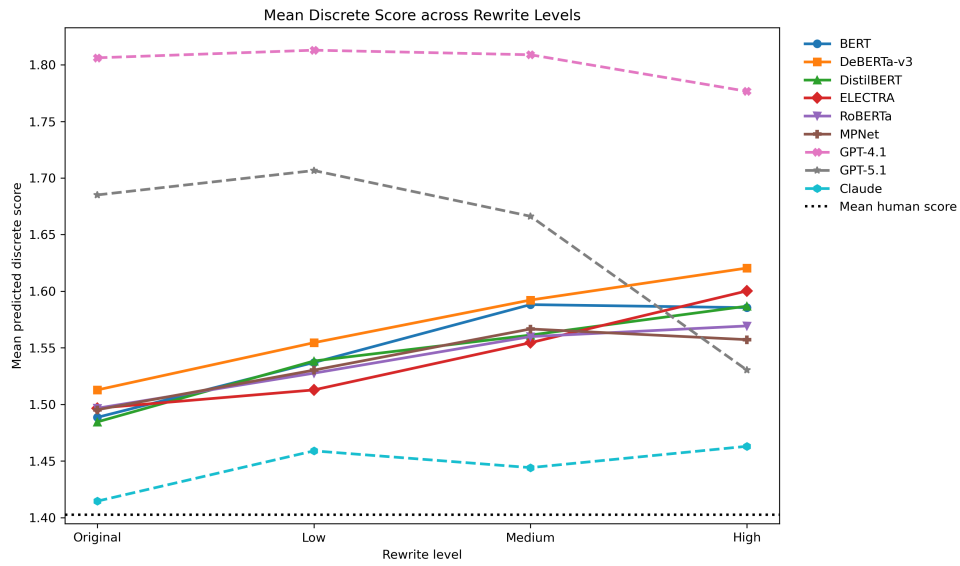


Figure 3: Mean predicted discrete scores across rewrite levels for all models.

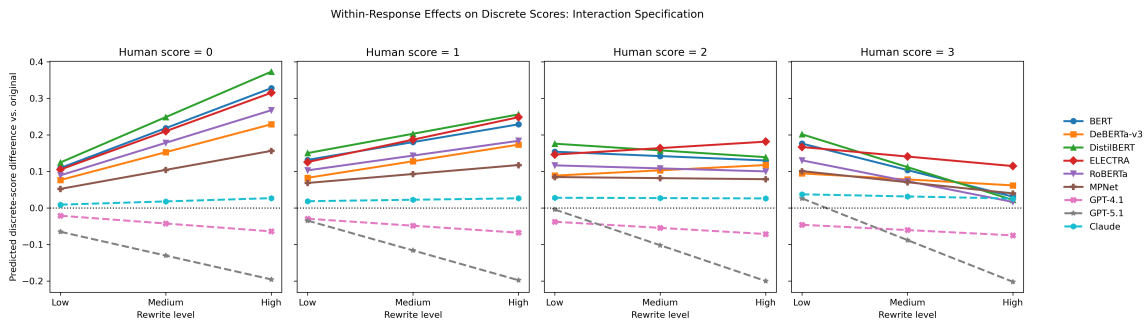


Figure 4: Within-response rewrite effects by human score level.

Table 2: Models used in the study

Model	Parameters	Architecture	Developer
<i>Fine-tuned encoders</i>			
microsoft/deberta-v3-base	183M	Encoder-only	Microsoft
roberta-base	125M	Encoder-only	Meta
bert-base-uncased	110M	Encoder-only	Google
google/electra-base-discriminator	110M	Encoder-only	Google
microsoft/mpnet-base	110M	Encoder-only	Microsoft
distilbert-base-uncased	66M	Encoder-only	Hugging Face
<i>Generative LLMs: proprietary</i>			
GPT-5.1	–	–	OpenAI
GPT-4.1	–	–	OpenAI
Claude Sonnet 4.6	–	–	Anthropic
<i>Generative LLMs: open weights</i>			
Nemotron-3-Super-120B-A12B	120B / 12B active	Decoder (hybrid Mamba-Transformer MoE)	NVIDIA
GPT-OSS-120B	117B / 5.1B active	Decoder (MoE)	OpenAI
Llama-3.3-70B-Instruct	70B	Decoder (dense)	Meta
Nemotron-3-Nano-30B-A3B	31.6B / 3.2B active	Decoder (hybrid Mamba-Transformer MoE)	NVIDIA
Gemma-4-31B-IT	31B	Decoder (dense)	Google
Mistral-Small-3.1	24B	Decoder (dense)	Mistral AI
GPT-OSS-20B	21B / 3.6B active	Decoder (MoE)	OpenAI
Granite-3.3-8B-Instruct	8B	Decoder (dense)	IBM
<i>Rewrite model</i>			
GPT-4.1 mini	–	–	OpenAI

Note. “–” indicates that parameter counts or architectural details have not been publicly disclosed.

Table 3: Mean change in linguistic features by rewrite level

Feature	Raw change			SD units		
	Low	Medium	High	Low	Medium	High
Average word length	0.10	0.39	1.08	0.19	0.74	2.03
Flesch reading ease	-2.39	-13.31	-42.81	-0.11	-0.62	-1.98
Gunning Fog	0.28	2.12	8.55	0.05	0.41	1.65
Characters	1.30	15.08	49.87	0.04	0.47	1.54
Flesch-Kincaid grade	0.26	1.92	6.96	0.06	0.43	1.54
Adjectives	-0.02	-0.03	1.54	-0.01	-0.02	1.00
Nouns	-0.28	0.29	1.66	-0.15	0.15	0.87
Lexical TTR	0.01	0.03	0.06	0.16	0.47	0.80
Type-token ratio	0.01	0.03	0.05	0.15	0.47	0.79
Verbs	-0.17	-0.40	-1.68	-0.07	-0.17	-0.72
MTLD	1.94	6.36	14.46	0.09	0.29	0.65
Punctuation	0.90	1.01	1.07	0.53	0.59	0.62
Sentences	-0.09	-0.18	-0.45	-0.11	-0.22	-0.55
Average sentence length	-0.08	0.39	4.24	-0.01	0.05	0.49
Adverbs	-0.00	0.18	0.51	-0.00	0.16	0.47
Subordination	-0.02	0.17	0.50	-0.02	0.15	0.45
Words	-0.67	-0.48	0.32	-0.16	-0.11	0.08

Each value is the mean change from the response’s own original version ($n = 743$ per cell). SD units divide that change by the standard deviation of the same feature across all rewritten versions, so features measured on different scales can be compared. Negative values for Flesch reading ease indicate text that is harder to read.

Table 4: Agreement with human ratings

Model	QWK _{orig}	Exact	SMD	QWK _{rew}
<i>Human reference</i>				
Human-human (H1-H2)	0.945	0.888	0.000	—
<i>Fine-tuned encoders</i>				
RoBERTa	0.867	0.740	0.089	0.844
DeBERTa-v3	0.856	0.735	0.104	0.835
BERT	0.851	0.705	0.081	0.835
MPNet	0.851	0.708	0.088	0.847
ELECTRA	0.847	0.712	0.089	0.838
DistilBERT	0.845	0.703	0.077	0.818
<i>Generative LLMs: proprietary</i>				
Claude Sonnet 4.6	0.803	0.635	0.011	0.797
GPT-5.1	0.762	0.583	0.266	0.783
GPT-4.1	0.751	0.563	0.380	0.755
<i>Generative LLMs: open weights</i>				
Nemotron-3-Super-120B	0.802	0.635	0.075	0.791
Llama-3.3-70B	0.792	0.630	0.001	0.780
Gemma-4-31B	0.792	0.622	0.264	0.784
GPT-OSS-120B	0.754	0.600	0.093	0.757
GPT-OSS-20B	0.687	0.538	0.270	0.706
Mistral-Small-3.1	0.685	0.552	-0.105	0.682
Granite-3.3-8B	0.545	0.499	-0.386	0.494
Nemotron-3-Nano-30B	0.418	0.389	-0.208	0.592

Note. QWK_{orig}, exact agreement, and SMD are computed on the 743 original responses. QWK_{rew} is computed on the 2,229 rewritten responses using H1, the human score assigned to the corresponding original response, as the reference.

Table 5: Within-Response Effects: Fine-Tuned Encoder Models

	(1) Baseline	(2) + Linguistic	(3) + Interaction
Panel A: BERT			
Rewrite Level	0.024* (0.010)	0.076*** (0.017)	0.109*** (0.016)
Rewrite Level $\times$ H1			−0.061*** (0.009)
Panel B: DeBERTa-v3			
Rewrite Level	0.033*** (0.008)	0.043* (0.014)	0.076*** (0.013)
Rewrite Level $\times$ H1			−0.031*** (0.007)
Panel C: RoBERTa			
Rewrite Level	0.021* (0.008)	0.043* (0.015)	0.089*** (0.013)
Rewrite Level $\times$ H1			−0.049*** (0.007)
Panel D: DistilBERT			
Rewrite Level	0.024* (0.010)	0.050* (0.018)	0.124*** (0.016)
Rewrite Level $\times$ H1			−0.071*** (0.009)
Panel E: ELECTRA			
Rewrite Level	0.044*** (0.009)	0.040* (0.016)	0.105*** (0.015)
Rewrite Level $\times$ H1			−0.044*** (0.008)
Panel F: MPNet			
Rewrite Level	0.013 (0.008)	0.040* (0.014)	0.052*** (0.013)
Rewrite Level $\times$ H1			−0.028*** (0.007)

Note. Entries report fixed-effect coefficients from mixed-effects models with random intercepts and random slopes for Rewrite Level by response ID. Columns correspond to Baseline, +Linguistic, and +Interaction specifications. The dependent variable is the change in predicted score relative to the original response, and each specification includes 2,229 rewritten responses from 743 original responses. Specification (2) includes 14 linguistic change scores retained after VIF selection. * $p < .05$, ** $p < .01$, and *** $p < .001$, using Benjamini–Hochberg adjusted p values.

Table 6: Within-Response Effects: Large Language Models

	(1) Baseline	(2) + Linguistic	(3) + Interaction
<i>Proprietary models</i>			
Panel A: GPT-4.1			
Rewrite Level	−0.018* (0.007)	−0.004 (0.014)	−0.021 (0.012)
Rewrite Level × H1			0.002 (0.007)
Panel B: GPT-5.1			
Rewrite Level	−0.088*** (0.009)	−0.019 (0.018)	−0.065*** (0.015)
Rewrite Level × H1			−0.016 (0.009)
Panel C: Claude Sonnet 4.6			
Rewrite Level	0.002 (0.008)	0.029 (0.016)	0.009 (0.014)
Rewrite Level × H1			−0.005 (0.008)
<i>Open-weight models</i>			
Panel D: GPT-OSS-120B			
Rewrite Level	0.009 (0.010)	0.001 (0.020)	0.022 (0.017)
Rewrite Level × H1			−0.010 (0.010)
Panel E: GPT-OSS-20B			
Rewrite Level	0.027 (0.014)	0.070* (0.026)	0.041 (0.022)
Rewrite Level × H1			−0.010 (0.013)
Panel F: Nemotron-3-Super-120B			
Rewrite Level	−0.028** (0.009)	0.043* (0.017)	−0.047** (0.015)
Rewrite Level × H1			0.014 (0.008)
Panel G: Nemotron-3-Nano-30B			
Rewrite Level	−0.020 (0.012)	0.026 (0.022)	−0.094*** (0.020)
Rewrite Level × H1			0.053*** (0.011)
Panel H: Llama-3.3-70B			
Rewrite Level	0.057*** (0.010)	0.053* (0.018)	0.006 (0.016)
Rewrite Level × H1			0.036*** (0.009)
Panel I: Gemma-4-31B			
Rewrite Level	0.004 (0.007)	0.024 (0.013)	0.007 (0.012)
Rewrite Level × H1			−0.002 (0.007)
Panel J: Mistral-Small-3.1			
Rewrite Level	−0.026* (0.010)	0.006 (0.019)	−0.003 (0.016)
Rewrite Level × H1			−0.016 (0.009)
Panel K: Granite-3.3-8B			
Rewrite Level	−0.043* (0.016)	−0.032 (0.030)	−0.072* (0.027)
Rewrite Level × H1			0.021 (0.015)

Note. Entries report fixed-effect coefficients from mixed-effects models with random intercepts and random slopes for Rewrite Level by response ID. Columns correspond to Baseline, +Linguistic, and +Interaction specifications. The dependent variable is the change in predicted score relative to the original response, and each specification includes 2,229 rewritten responses from 743 original responses. Specification (2) includes 14 linguistic change scores retained after VIF selection. * $p < .05$, ** $p < .01$, and *** $p < .001$, using Benjamini–Hochberg adjusted p values.

Table 7: Interaction with human score: full and interior-prediction samples

Model	Full	<i>n</i>	Interior	<i>n</i>
<i>Fine-tuned encoders</i>				
BERT	-0.061***	2229	-0.049*	1140
DeBERTa-v3	-0.031***	2229	-0.005	1083
RoBERTa	-0.049***	2229	-0.040*	1206
DistilBERT	-0.071***	2229	-0.071***	1191
ELECTRA	-0.044***	2229	-0.034	1176
MPNet	-0.028***	2229	-0.012	1122
<i>Generative LLMs: proprietary</i>				
GPT-4.1	0.002	2229	0.031	957
GPT-5.1	-0.016	2229	-0.006	1071
Claude Sonnet 4.6	-0.005	2229	0.026	1137
<i>Generative LLMs: open weights</i>				
GPT-OSS-120B	-0.010	2229	0.017	1104
GPT-OSS-20B	-0.010	2229	0.030	894
Nemotron-3-Super-120B	0.014	2229	0.068**	1080
Nemotron-3-Nano-30B	0.053***	2229	0.057***	2070
Llama-3.3-70B	0.036***	2229	0.091***	1179
Gemma-4-31B	-0.002	2229	0.030	1023
Mistral-Small-3.1	-0.016	2229	0.008	1740
Granite-3.3-8B	0.021	2229	0.059*	1377

Note. Entries are Rewrite Level $\times$ Human Score coefficients. Interior samples retain responses with original predicted scores of 1 or 2. * $p < .05$, ** $p < .01$, *** $p < .001$, Benjamini–Hochberg adjusted.

C Rewriting Prompt

The full rewriting prompt used to generate the LOW, MEDIUM, and HIGH rewrites is shown below.

[TASK PROMPT]

A group of students wrote the following procedure for their investigation.

Procedure:

1. Determine the mass of four different samples.
2. Pour vinegar in each of four separate, but identical, containers.
3. Place a sample of one material into one container and label. Repeat with remaining samples, placing a single sample into a single container.
4. After 24 hours, remove the samples from the containers and rinse each sample with distilled water.
5. Allow the samples to sit and dry for 30 minutes.
6. Determine the mass of each sample.

The students data are recorded in the table below.

Sample	Starting Mass (g)	Ending Mass (g)	Difference in Mass (g)
Marble	9.8	9.4	-0.4
Limestone	10.4	9.1	-1.3
Wood	11.2	11.2	0.0
Plastic	7.2	7.1	-0.1

After reading the groups procedure, describe what additional information you would need in order to replicate the experiment. Make sure to include at least three pieces of information.

[GLOBAL REWRITE INSTRUCTION]

You are a highly careful linguistic rewriter.

Your job is to improve language ONLY.

You must preserve the student's original information and intended meaning.

STRICT RULES:

- Do NOT add new information.
- Do NOT remove information.
- Do NOT change any numbers.
- Do NOT replace unclear content with guessed new content.
- Keep unusual words if their meaning is unclear.
- If formatting is awkward, you may improve the formatting while preserving the same content.
- If the original response is written as bullets, fragments, or loosely separated sentences, you may rewrite it as fluent prose.
- You may merge or split sentences if needed for clarity and flow, as long as all original information is preserved.

LEVEL CONTRAST REQUIREMENT:

- The three rewrite levels must be clearly distinguishable in how much they change the wording.
- LOW should remain close to the original and mainly make light local improvements.
- MEDIUM should show noticeable rephrasing and clearer sentence-level rewriting.
- HIGH should show substantial restructuring and a much more polished academic style.
- LOW, MEDIUM, and HIGH should be visibly different to a human reader, while preserving the same information.

[LEVEL-SPECIFIC INSTRUCTIONS]

LEVEL = LOW

Target behavior:

- Light polishing only
- Correct spelling, grammar, punctuation, and wording errors
- Keep the original structure mostly unchanged
- Make local improvements rather than global restructuring
- Preserve the student's original meaning and general tone

- Mildly improve overly informal wording where appropriate, but do not substantially raise the level of formality

LEVEL = MEDIUM

Target behavior:

- Improve clarity, flow, and sentence structure
- Use more precise and moderately more formal vocabulary where appropriate
- Replace informal, vague, or repetitive wording with clearer phrasing
- The rewrite should go beyond simple grammar correction
- Minor edits alone are NOT sufficient
- The sentence should be noticeably rephrased compared to the original
- Reorganize wording and sentence structure clearly, not just by substituting a few words
- You may merge or split sentences if needed
- Preserve all original information

LEVEL = HIGH

Target behavior:

- Strong language enhancement
- Use sophisticated, precise, and academically appropriate vocabulary
- Replace informal, vague, simplistic, or repetitive wording with more formal and information-dense phrasing
- Prefer structural rewriting over surface-level editing
- The rewrite should not closely mirror the original sentence structure
- The rewrite should present the same information in a more mature, concise, and formally structured way
- Compress repeated phrasing when possible while preserving all original information
- Use nominalization, subordination, apposition, or denser sentence structure where appropriate
- You may merge or split sentences if needed
- You may convert list-like phrasing into fluent prose if all information is preserved
- Preserve all original information exactly

[FEW-SHOT EXEMPLAR]

EXAMPLE TYPE: procedural_explanation

EXAMPLE ORIGINAL:

I would need to know what step are testing, what the independant variable is, what size be samples they are talking about. I would also need to know what other variables are controlled.

EXAMPLE LOW:

I would need to know what step is being tested, what the independent variable is, and what size the samples are that they are referring to. I would also need to know what other variables are controlled.

EXAMPLE MEDIUM:

I would need to know which step is being tested, what the independent variable is, and the size of the samples being referred to, as well as which other variables are controlled.

EXAMPLE HIGH:

Replicating the experiment would require identification of the step being tested, the independent variable, the size of the samples being referenced, and the additional variables that are controlled.

[REGENERATION PROMPT]

The previous rewrite did not fully satisfy the requirements.

LEVEL:
[LEVEL]

ORIGINAL RESPONSE:
[ORIGINAL]

PREVIOUS REWRITE:

[CURRENT]

ISSUES TO FIX:

[FAIL REASONS]

Please rewrite it again.

Additional guidance:

- The new version should show clearer improvement in wording and sentence structure.
- Do not rely only on minor grammar correction.
- Use more precise and more formal phrasing where appropriate.
- At MEDIUM and HIGH levels, the rewrite should sound noticeably more polished than the original.
- At HIGH level, prioritize stronger sentence restructuring and more information-dense expression.

Requirements:

- Preserve all original information
- Keep all numbers unchanged
- Do not add or delete content
- You may merge or split sentences if needed
- You may convert list-like structure into fluent prose if all information is preserved
- Return ONLY the rewritten response

D LLM Scoring Prompt

[GENERAL SCORING RUBRIC]

General scoring rubric for science open-ended items (0-3):

Score 3:

The response is an excellent answer to the question. It is correct, complete, and appropriate and contains elaboration, extension, and/or evidence of higher-order thinking and relevant prior knowledge. There is no evidence of misconceptions. Minor errors will not necessarily lower the score.

Score 2:

The response is a proficient answer to the question. It is generally correct, complete, and appropriate, although minor inaccuracies may appear. There may be limited evidence of elaboration, extension, higher-order thinking, and relevant prior knowledge, or there may be significant evidence of these traits but other flaws may be more than minor.

Score 1:

The response is a marginal answer to the question. While it may contain some elements of a proficient response, it is inaccurate, incomplete, and/or inappropriate. There is little evidence, if any, of elaboration, extension, higher-order thinking, or relevant prior knowledge. There may be evidence of significant misconceptions.

Score 0:

The response, though possibly on topic, is an unsatisfactory answer to the question. It may fail to address the question, or it may address the question in a very limited way. There may be no evidence of elaboration, extension, higher-order thinking, or relevant prior knowledge. There may be evidence of serious misconceptions.

[ITEM-SPECIFIC RUBRIC]

Rubric for Acid Rain

Possible correct responses include additional information such as:

- how much vinegar was used in each container
- what type of vinegar was used
- what materials to test
- what size or surface area of materials should be used
- how long each sample was rinsed in distilled water
- what drying method to use
- what size or type of container to use
- other acceptable responses

Item-specific 0-3 scoring rule:

- Score 3: describes three additional pieces of information needed to accurately replicate the experiment
- Score 2: describes two additional pieces of information needed to accurately replicate the experiment
- Score 1: describes one additional piece of information needed to accurately replicate the experiment
- Score 0: describes little or no accurate or relevant information from the acid rain investigation