Psychometric Analysis of LLM Differential Sensitivity to Word Predictability using Explanatory MFRM

Wesley Morris[1*], Langdon Holmes[1] and Scott Crossley[1]

[1] *Vanderbilt University, 2201 West End Ave. Nashville TN 37203, United States*

Abstract

Language Models (LLMs) can achieve high reliability but their internal scoring logic remains opaque, raising concerns about construct validity and generalizability. This study applies a rigorous psychometric framework – Explanatory Many-Facet Rasch Measurement (eMFRM) – to evaluate whether LLMs are differentially sensitive to measures of word predictability calculated by their un-fine-tuned base models. Results reveal that the LLM rater demonstrated a small but significantly greater sensitivity to traditional corpus-based association measures than human raters, and that humans and machine raters are equally sensitive to word predictability. This provides evidence that representation LLMs may not be influenced by their training sets when predicting human phraseology scores.

Keywords

Explanatory MFRM, LLMs, Automatic Essay Scoring, Differential Sensitivity, Phraseology

1. Introduction and Background

The landscape of language evaluation is undergoing a profound shift with the advent of Large Language Models (LLMs) as automated raters. While generative LLMs such as GPT are more widely known, smaller embedding LLMs, such as BERT, can achieve high agreement with human raters on benchmark datasets [18, 23] when fine-tuned on human-scored samples,. Yet this reliability often masks a critical question: What construct is the model actually measuring? LLMs are notoriously inscrutable "black boxes" whose parameters encode complex statistical patterns from large training corpora. Although they predict human scores accurately, the rationale for any specific score remains opaque [8]. This opacity comes with significant risk: an LLM rater may achieve high reliability by latching onto surface-level linguistic features correlated with, but not constitutive of, the target construct (e.g. using text length to predict essay quality). The LLM rater may therefore fail to generalize or be vulnerable to adversarial gaming [7, 10]. The gap between reliability and interpretability points to a need for psychometric frameworks that can evaluate how LLMs arrive at scores, not only whether their scores align with humans.

LLM Psychometrics: Applying Assessment Theories to Evaluate LLMs and LLM-Supported Learners, LAK 2026 Workshop
* Corresponding author.
wesley.g.morris@vanderbilt.edu (W. Morris); langdon.holmes@vanderbilt.edu (L. Holmes);
scott.crossley@vanderbilt.edu (S. Crossley)
0000-0001-6316-6479 (W. Morris); 0000-0003-4338-4609 (L. Holmes); 0000-0002-5148-0273 (S. Crossley)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4330/paper8.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-15

This study builds on work of De Boeck and Wilson [5] and Kim and Wilson [13] to use Explanatory Many-Facet Rasch Measurment (eMFRM) to explore LLM scoring mechanisms. Many-Facet Rasch Measurement (MFRM) [16] approaches explicitly model systematic rater effects, such as severity or leniency, as distinct "facets" of the measurement process. This allows for the partitioning of score variance, acknowledging that a score is jointly determined by the quality of the text and the idiosyncrasies of the rater. An explanatory extension of MFRM, eMFRM decomposes rater and text facets into vectors of explanatory covariates with random effects for each of the facets. It moves from asking *if* a rater is lenient to *why* a rater is lenient by modeling how specific rater characteristics (e.g. rater experience, training, or in our case, rater type: human vs. machine) interact with specific text features (e.g. indices of linguistic features). The key parameters are these interaction effects, which quantify differential sensitivity – the degree to which individual raters are differentially sensitive to specific linguistic features [17].

We apply eMFRM to the evaluation of phraseological competence (e.g., the appropriate use of collocations and formulaic language) in second language (L2) writers. Phraseological competence refers to a writer's mastery of conventional, multi-word expressions such as collocations (*strong coffee*), lexical bundles (*on the other hand*), and idioms (*miss the boat*). It is a key indicator of advanced L2 proficiency and writing quality [2], because it reflects a native-like, automatized knowledge of how words combine within specific genres and contexts [19]. In fact, human judgements of collocation accuracy have been found to account for 84%-89% of the variance in holistic scores of lexical proficiency for L2 learners [3].

Phraseological competence is closely related to the statistical predictability of words – a property that is central to the training of representation-based LLMs. Specifically, LLMs use masked language modeling where one or more tokens are masked during pretraining, and the model generates a probability distribution across its vocabulary for the token in the masked position. Because LLM learning is based on word distributions, we hypothesize that LLM raters will be differentially sensitive to word predictability relative to human raters, who may employ a broader, more pragmatic evaluative schema.

This study applies a psychometric statistical framework to evaluate LLMs. Our research questions asks if representation LLM raters exhibit significantly different sensitivity than human raters to word predictability as measured by masked language modeling by the LLM base models? We predict that LLM raters will display greater sensitivity to mean text-level word predictability as measured by their base models compared to human raters. Our approach can provide concrete, model-based evidence that LLMs raters may operationalize constructs in ways fundamentally shaped by their architecture and training, offering a path toward a more transparent and valid AI-powered assessment.

2. Methods

2.1. Data

This study uses the English Language Learner Insight, Proficiency, and Skills Evaluation (ELLIPSE) corpus [4], comprising 6,345 persuasive essays written by English language learners. For this analysis, we focus only on the phraseology dimension of the analytic rubric used in the original data collection (see Appendix A). The rubric defines phraseological competence across a 5-point scale, emphasizing the use of lexical bundles, collocations, and idioms. The essays

were scored by a total of 27 expert raters, with each essay receiving scores from two of the raters. In total, the dataset comprises a total of 12,690 individual ratings with each rater scoring a mean of 514 essays (SD=287.9, min=22, max=1,090). Because each essay has only 2 ratings, the rater-by-text matrix is sparse. However, the data is missing by design, which allows parameter estimation of text and rater-level parameters within a Rasch framework. Although the rubric scale has a minimum of 1, in 36 cases raters assigned essays scores of 0. After removing these scores of 0 from the dataset, we were left 12,690 scores of 6,345 essays.

2.2. Large Language Model (LLM) Rater

To create the LLM rater facet, we used a representation-based model architecture: ModernBERT-base [22]. ModernBERT-base is a performant, efficient bidirectional encoder which allows for longer context lengths than its predecessors in the BERT [6] family of models. The fine-tuning and scoring procedure was designed to mirror the human rating design. The data was randomly split into ten exclusive, non-overlapping subsets. For each of the ten ModernBERT instances, a model was fine-tuned to predict the human-assigned phraseology score using 9 of the 10 subsets (~5710 essays). Hyperparameter tuning was based on default specifications and each model was trained for two epochs with a batch size of 8, with 100 warmup steps, a weight decay of 1e-6, and a learning rate of 8e-5 using mean squared error (MSE) as a loss function. They then predicted scores from the subset that was withheld from training. All models achieved MSE of between 0.38 and 0.44 (M=0.41) on their held-out test sets. To match the ordinal structure of the original dataset, scores were rounded to the nearest integer with a minimum of one and a maximum of 5. The final dataset comprised 6,345 essays, each with its original two human ratings and one newly generated LLM rating, for a total of 19,035 observations nested within texts and crossed by rater type. Final inter-rater agreement between the humans and the models was measured using Cohen's weighted kappa against the human consensus, calculated by rounding the mean human score to the nearest integer.

2.3. Text-level Covariates: Measures of Word Predictability

We calculated two metrics of word predictability and association strength for each essay to serve as explanatory text covariates. The first, mean word probability (LLM-Predictability) developed by Holmes et al. [11] used the base (un-finetuned) ModernBERT model to compute the probability of each word in its textual content using the masked language modeling objective. This probability is computed by masking each word in the text, one at a time, and generating a probability distribution for each masked position across the model's vocabulary. The word probability for each observed word is simply its probability in the generated distribution. The mean probability across all words in an essay is taken as a global index of phraseological conventionality, where higher word predictability indicates more predictable, and thus potentially more conventional, word combinations. Previous research has shown that linguistic indices generated in this way predict holistic essay scores for language learners [11, 12].

Additionally, we included a traditional measure of association strength – Mean Pointwise Mutual Information (PMI) – as a measure of phraseological conventionality that is unassociated with the LLM being used to assign the score. PMI quantifies the strength of association between adjacent pairs of words as the logarithm of the ratio of their observed co-occurrence frequency

to their expected frequency under statistical independence. High PMI values indicate that the constituent words co-occur more often than would be expected by chance [15]. We used the Tool for the Automatic Analysis of Lexical Sophistication (TAALES) [14] to calculate PMI using the Corpus of Contemporary American English (COCA) magazine subcorpus [21] as a reference.

2.4. Statistical Method: The Doubly-Explanatory MFRM

The core analysis employed a Bayesian generalized linear mixed model (GLMM) with a partial credit specification, implementing a doubly explanatory MFRM as described by Kim and Wilson [13]. The model treats the polytomous phraseology score (1-5) as an ordinal outcome. The outcome $Y_{tr} \in \{1, 2, 3, 4, 5\}$ is the phraseology score assigned to text t by rater r. The probability that the score is at or above category m is modeled with a cumulative logit link. The linear predictor is:

$$\eta_{trm} = \sum_{j=0}^{J} \theta_j Z_{tj} + \beta X_r + \sum_{j=1}^{J} \omega_j Z_{tj} X_r + \tau_m + \varepsilon_t + \varepsilon_r \quad (1)$$

Where θ coefficients represent effects for each of $\mathcal{J}$ measures (i.e. LLM-predictability and PMI) of phraseological conventionality and Z is a dummy-coded vector which indicates which measure's parameter is being estimated, β represents the effect of rater type (human vs. machine) and X is a binary dummy coded vector. The ω coefficients are the key parameters of interest. They estimate the difference in the slope (sensitivity) for each predictability feature between LLM raters and human raters. A positive ω indicates the LLM rater increases the score more per unit increase in that feature than the human rater does. τ_m is a fixed threshold parameter for the cumulative logits at threshold m. The model also estimates random effects including $\varepsilon_t \sim N(0, \sigma_t^2)$ which is a random intercept for text, capturing residual text quality after accounting for the predictability features, while $\varepsilon_r \sim N(0, \sigma_r^2)$ is a random intercept for rater, capturing residual rater severity after accounting for rater type.

The model was estimated using a Bayesian framework implemented in Python with the Bambi interface to PyMC [20]. Weakly informative, regularizing priors were placed on the interaction terms (ω) to mitigate overfitting and stabilize estimation. Parameters were estimated using Hamiltonian Monte Carlo via the NumPyro backend in Bambi, allowing GPU acceleration. Four Markov chains were run for 2,000 post-warmup iterations each, following 1,000 adaptation iterations, with a target acceptance rate of 0.95 to improve sampling efficiency. Convergence was assessed using $\hat{R}$ statistics and effective sample size. Hypothesis testing was conducted by examining the 94% Highest Posterior Density (HPD) intervals for the interaction parameters (ω). Evidence for differential sensitivity was found if an HPD interval for a key ω parameter excludes zero. This analysis directly tests whether LLM rater's scoring function is differentially sensitive to statistical word predictability compared to that of human raters.

3. Results

3.1. Model Training and Index Calculation

We found that the LLM rater achieved human-like interrater reliability, with QWK to the human consensus score of 0.52, slightly higher than the human-human reliability of 0.51.

Despite the moderate level of agreement [1], some variance still existed between raters. Our next question is whether the variance from the human and LLM raters is differentially predicted by word predictability measures (i.e., ModernBERT word predictability or PMI predictability).

3.2. Doubly-explanatory Many Facet Rasch Measurement

The doubly-explanatory MFRM was estimated successfully with all convergence diagnostics indicating stable posterior distributions (all $\hat{R} < 1.01$, all ESS > 1,000). Although the results provide some quantifiable evidence of differential sensitivity to traditional measures of word association, the results indicate no significant differential sensitivity of the LLM-based measure to word predictability (see Table 1).

The analysis revealed strong, consistent main effects for the linguistic predictors such that a text's LLM-calculated word predictability was a positive predictor of its phraseology score. Similarly, stronger pointwise mutual information (PMI) from the COCA magazine corpus was associated with higher scores. There was no evidence of an average difference in severity between the rater types, as the credible interval for the human rater effect included zero. A small but significant interaction was found for the PMI feature (see Figure 1). With 97% of the posterior density below zero, there is some evidence that human raters may be less sensitive to this corpus-derived association measure than the LLM raters. Critically, no reliable differential sensitivity was found for the LLM-calculated predictability measure, suggesting that human and LLM raters were similarly sensitive to the predictability metric derived from the LLM's own base architecture. The random effects identified substantial residual variance attributable to differences between texts, indicating that the two predictability features do not fully capture text quality. Meaningful variance was also attributed to individual rater idiosyncrasies, confirming persistent differences in severity or leniency. The threshold parameters (e.g., threshold 1-2: β = -9.49) were well-estimated and showed the expected monotonic increase, confirming the ordered nature of the rating scale.

Table 1
Results of Doubly-Explanatory Many-Facet Rasch Measurement

	Effect	β	SD	HDI 3%	HDI 97%
Threshold Intercepts	threshold 1-2	-9.49	0.303	-10.051	-8.908
	threshold 2-3	-2.792	0.24	-3.248	-2.338
	threshold 3-4	2.109	0.24	1.673	2.576
	threshold 4-5	6.981	0.255	6.517	7.471
Fixed Effects	Rater [Human]	0.231	0.285	-0.328	0.74
	LLM Predictability	1.737	0.046	1.65	1.821
	COCA Magazine MI	0.295	0.041	0.217	0.37
	LLM Pred : Rater [human]	0.058	0.043	-0.024	0.138
	COCA Magazine MI : Rater [human]	-0.076	0.04	-0.155	-0.004
Random Effects	filename sigma	1.648	0.034	1.585	1.713
	rater sigma	0.856	0.11	0.651	1.049

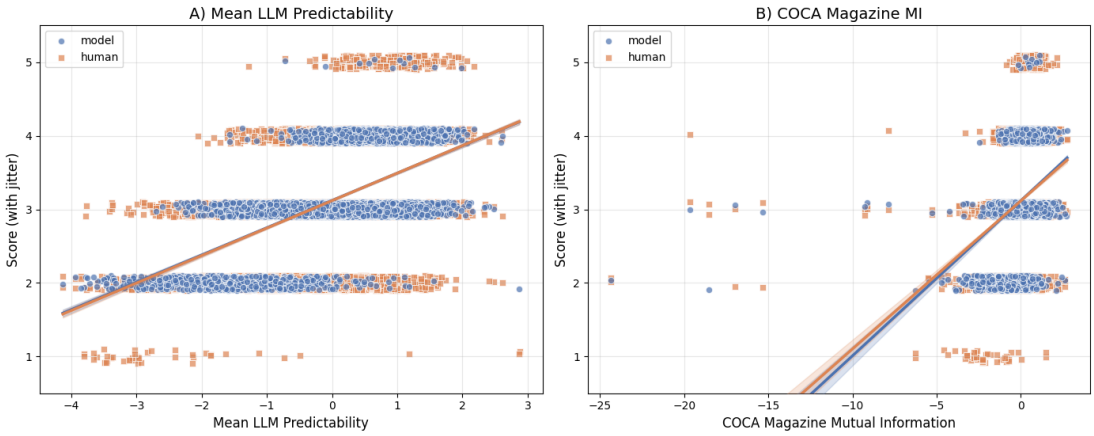


Figure 1: Word Association Measures with Linear Regression

4. Discussion and Conclusion

This study applied Explanatory Many-Facet Rasch Measurement (eMFRM) as a psychometric tool to evaluate a representation-based LLM rater, moving beyond simple agreement metrics to test a hypothesis about its internal scoring logic. As predicted based on previous work [3, 11, 12], both measures of word predictability/association had significant positive main effects on phraseology scores. The results of the interaction effects show a nuanced validation argument. The significant interaction for corpus-based PMI confirms that the LLM rater operationalized phraseological competence slightly differently than human raters, placing disproportionate weight on formal collocational association strength. This constitutes measurable differential sensitivity and potential construct-irrelevant variance, suggesting the model may have learned a heuristic anchored to surface-level co-occurrence statistics from its training data.

However, the null interaction for the LLM's own predictability metric is a critical, counter-intuitive finding. It indicates that the fine-tuning process successfully calibrated the model's use of its core architectural signal – word predictability – to align with human sensitivity for the task of assessing phraseological use. This alignment is a positive sign for the validity of this specific automated scoring system, suggesting it responds to this feature in a human-like way. Nevertheless, the differential sensitivity to PMI reveals that alignment is incomplete, as LLMs appear to be slightly more sensitive to pointwise mutual information than human raters.

In conclusion, this work demonstrates that eMFRM can provide a powerful framework for LLM psychometrics by transforming the "black box" into a statistical rater whose scoring behavior can be decomposed, compared to human benchmarks, and rigorously evaluated for construct relevance. The findings argue for a shift in validation practice: instead of asking only if an AI agrees with humans, we can use models like eMFRM to investigate how and why LLM raters may agree or disagree with human raters. For practitioners and developers, this method offers a pathway to diagnose specific biases, guide targeted model refinement, and ultimately build more transparent and trustworthy AI-assisted assessment systems. Future work should extend this approach to generative raters and other textual constructs to further generalize this methodology.

5. Appendices

Appendix A: Scoring Rubric for the Ellipse Dataset

English Proficiency Scoring Rubric for English Language Learners

	HOLISTIC	ANALYTIC					
	Overall	**Cohesion**	**Syntax**	**Vocabulary**	**Phraseology**	**Grammar**	**Conventions**
5	Native-like facility in the use of language with syntactic variety, Appropriate word choice and phrases; well-controlled text organization; precise use of grammar and conventions; rare language inaccuracies that do not impede communication.	Text organization consistently well controlled using a variety of effective linguistic features such as reference and transitional words and phrases to connect ideas across sentences and paragraphs; appropriate overlap of ideas.	Flexible and effective use of a full range of syntactic structures including simple, compound, and complex sentences; There may be rare minor and negligible errors in sentence formation.	Wide range of vocabulary flexibly and effectively used to convey precise meanings; skillful use of topic-related terms and less common words; rare negligible inaccuracies in word use.	Flexible and effective use of a variety of phrases, such as idioms, collocations, and lexical bundles, to convey precise and subtle meanings; rare minor inaccuracies that are negligible.	Command of grammar and usage with few or no errors.	Consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors nonexistent or negligible.
4	Facility in the use of language with syntactic variety and range of words and phrases; controlled organization; accuracy in grammar and conventions; occasional language inaccuracies that rarely impede communication.	Organization generally well controlled; a range of cohesive devices used appropriately such as reference and transitional words and phrases to connect ideas; generally appropriate overlap of ideas	Appropriate use of a variety of syntactic structures, such as simple, compound, and complex sentences; occasional errors or inappropriatenes s in sentence formation.	Sufficient range of vocabulary to allow flexibility and precision; appropriate use of topic-related terms and less common lexical items	Appropriate use of a variety of phrases, such as idioms, collocations, and lexical bundles; occasional inaccuracies and colloquialisms.	Minimal errors in grammar and usage.	Generally consistent use of appropriate conventions to convey meaning; spelling, capitalization, and punctuation errors few and not distracting.
3	Facility limited to the use of common structures and generic vocabulary; organization generally controlled although connection sometimes absent or unsuccessful; errors in grammar and syntax and usage. Communication is impeded by language inaccuracies in some cases.	Organization generally controlled; cohesive devices used but limited in type; Some repetitive, mechanical, or faulty use of cohesion use within and/or between sentences and paragraphs.	Simple, compound, and complex syntactic structures present although the range may be limited; some apparent errors in sentence formation, especially in more complex sentences.	Minimally adequate range of vocabulary for the topic; no precise use of subtle word meanings; topic related terms only used occasionally; attempts to use less common vocabulary but with some inaccuracy	Evident use of phrases such as idioms, collocations, and lexical bundles but without much variety; some noticeable repetitions and misuses.	Some errors in grammar and usage.	Developing use of conventions to convey meaning; errors in spelling, capitalization, and punctuation that are sometimes distracting.
2	Inconsistent facility in sentence formation, word choice, and mechanics; organization partially developed but may be missing or unsuccessful. Communication impeded in many instances by language inaccuracies.	Organization only partially developed with a lack of logical sequencing of ideas; some basic cohesive devices used but with inaccuracy or repetition.	Some sentence variation used; many sentence structure problems.	Narrow range of vocabulary to convey basic and elementary meaning; topic related terms used inappropri ately; errors in word formation and word choice that may distort meanings	Narrow range of phrases, such as collocations and lexical bundles, used to convey basic and elementary meaning; many repetitions and /or misuses of phrases.	Many errors in grammar and usage.	Variable use of conventions; spelling, capitalization, and punctuation errors frequent and distracting.
1	A limited range of familiar words or phrases loosely strung together; frequent errors in grammar (including syntax) and usage. Communication impeded in most cases by language inaccuracies.	No clear control of organization; cohesive devices not present or unsuccessfully used; presentation of ideas unclear.	Pervasive and basic errors in sentence structure and word order that cause confusion; basic sentences errors common.	Limited vocabulary often inappropriately used; limited control of word choice and word forms; little attempt to use topic-related terms	Memorized chunks of language, or simple phrasal patterns predominate; many repetitions and misuses of phrases.	Errors in grammar and usage throughout.	Minimal use of conventions; spelling, capitalizatio n, and punctuation errors throughout.

Declaration on Generative AI

During the preparation of this work, the authors used Claude 4.8 Opus in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content

References

[1] Afrizal Doewes, Nughthoh Arfawi Kurdhi, and Akrati Saxena 2023. Evaluating Quadratic Weighted Kappa as the Standard Performance Metric for Automated Essay Scoring. (July 2023). https://doi.org/10.5281/ZENODO.8115784.

[2] Cowie, A.P. and Howarth, P. 1996. Phraseological Competence and Written Proficiency. *Language and education : papers from the annual meeting of the British Association for Applied Linguistics held at the University of Southampton, September 1995*. Multilingual Matters Ltd. 80–93.

[3] Crossley, S.A., Salsbury, T. and Mcnamara, D.S. 2014. Assessing Lexical Proficiency Using Analytic Ratings: A Case for Collocation Accuracy. *Applied Linguistics*. (Feb. 2014), amt056. https://doi.org/10.1093/applin/amt056.

[4] Crossley, S.A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A. and 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*. (in press).

[5] De Boeck, P. and Wilson, M. eds. 2004. *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer New York.

[6] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.

[7] Doewes, A. and Pechenizkiy, M. 2021. On the Limitations of Human-Computer Agreement in Automated Essay Scoring. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*. (2021).

[8] Ferrara, S. and Qunbar, S. 2022. Validity Arguments for AI-Based Automated Scores: Essay Scoring as an Illustration. *Journal of Educational Measurement*. 59, 3 (Sept. 2022), 288–313. https://doi.org/10.1111/jedm.12333.

[9] Han, B.-C. 2017. *Psychopolitics: neoliberalism and new technologies of power*. Verso.

[10] Hannah, L., Jang, E.E., Shah, M. and Gupta, V. 2023. Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits. *Language Assessment Quarterly*. 20, 4–5 (Oct. 2023), 399–420. https://doi.org/10.1080/15434303.2023.2288253.

[11] Holmes, L., Crossley, S., Morris, W. and Choi, J.S. under review. Assessing Second Language Proficiency using Automatic Measures: A Case for Word Predictability. *Language Learning & Technology*. (under review).

[12] Hu, J. and Cong, Y. 2025. Modeling Chinese L2 Writing Development: The LLM-Surprisal Perspective. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics* (Albuquerque, New Mexico, USA, 2025), 172–183.

[13] Kim, J. and Wilson, M. 2020. Polytomous Item Explanatory Item Response Theory Models. *Educational and Psychological Measurement*. 80, 4 (Aug. 2020), 726–755. https://doi.org/10.1177/0013164419892667.

[14] Kyle, K. and Crossley, S.A. 2015. Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *TESOL Quarterly*. 49, 4 (Dec. 2015), 757–786. https://doi.org/10.1002/tesq.194.

[15] Kyle, K. and Eguchi, M. 2021. 6 Automatically Assessing Lexical Sophistication Using Word, Bigram, and Dependency Indices. *Perspectives on the L2 Phrasicon*. S. Granger, ed. Multilingual Matters. 126–151.

[16] Linacre, J. 1989. *Many-faceted Rasch measurement*. University of Chicago.

[17] Morris, W., Holmes, L., Choi, J.S. and Crossley, S. 2025. Uncovering Differential Sensitivity Toward Linguistic Features of Cohesion in Large Language Models. *Artificial Intelligence in Education*. A.I. Cristea, E. Walker, Y. Lu, O.C. Santos, and S. Isotani, eds. Springer Nature Switzerland. 227–234.

[18] Pack, A., Barrett, A. and Escalante, J. 2024. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*. 6, (June 2024), 100234. https://doi.org/10.1016/j.caeai.2024.100234.

[19] Paquot, M. 2019. The phraseological dimension in interlanguage complexity research. *Second Language Research*. 35, 1 (Jan. 2019), 121–145. https://doi.org/10.1177/0267658317694221.

[20] Salvatier, J., Wiecki, T.V. and Fonnesbeck, C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*. 2, (Apr. 2016), e55. https://doi.org/10.7717/peerj-cs.55.

[21] The Corpus of Contemporary American English (COCA): 2008. *https://www.english-corpora.org/coca/*.

[22] Warner, B., Chaffin, A., Clavié, B., Weller, O., Hallström, O., Taghadouini, S., Gallagher, A., Biswas, R., Ladhak, F., Aarsen, T., Cooper, N., Adams, G., Howard, J. and Poli, I. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv.

[23] Zhong, W. 2024. Effectiveness of finetuning pretrained BERT and deBERTa for automatic essay scoring. *Applied and Computational Engineering*. 52, 1 (Mar. 2024), 87–95. https://doi.org/10.54254/2755-2721/52/20241321.

A. Online Resources

Code for this study can be found at https://github.com/wesleygmorris/llm-psychometrics/