

Assessing the Reliability and Construct Representation of LLM-based Language Proficiency Scores

Langdon Holmes

Vanderbilt University

Nashville, USA

langdon.holmes@vanderbilt.edu

Joon Suh Choi

The Learning Agency

joon@the-learning-agency.com

Scott Crossley

Vanderbilt University

Nashville, USA

scott.crossley@vanderbilt.edu

Wesley Morris

Vanderbilt University

Nashville, USA

wesley.g.morris@vanderbilt.edu

Abstract

We used confirmatory factor analysis to assess the reliability and construct representation of an LLM-based measurement instrument of language proficiency. LLMs were at least as reliable as human raters and loaded onto the same underlying factor, though analyses indicated a less than perfect alignment between LLM and human raters.

1 Introduction

The assessment of second language (L2) writing performances often relies on the use of expert human raters to assign scores against a rating scale (Barkaoui, 2011). These scores can provide holistic measures of writing quality as well as analytic ratings of specific language attributes such as vocabulary, syntax, and grammar/mechanics (Hunter et al., 1996). However, human ratings are inherently subjective and can be unreliable (Barkaoui, 2011; Jølle and Skar, 2020). Various approaches have been used to mitigate this subjectivity and increase the reliability of human ratings, but they can add costs, and collecting expert ratings of L2 discourse is already resource intensive (Lumley and McNamara, 1993; Wind and Peterson, 2018).

As a result, automated scoring models have been developed that predict human scores using surface-level linguistic features. While such features have been shown to be useful measures of argumentation, organization, and vocabulary use (Deane, 2013), scoring models trained on surface-level features may struggle to adequately measure higher level discourse traits such as coherence, text organization, and clarity of ideas, indicating that some construct-relevant language features are not adequately captured. Within these models, there are also concerns surrounding bias across certain demographics, such as gender, language background, nationality, and language proficiency (Ramineni et al., 2012) and the potential for scoring models

to be gamed by students (Perelman, 2020). Despite these concerns, uptake in the use of automatic scoring systems is increasing, and many large-scale writing assessments are moving forward with integrating automatic scoring systems. This movement in the direction of fully automated scoring systems has become even more pronounced with the development of large language models (LLMs) that can more consistently make human-like decisions, although research in this area is just beginning (Laurer et al., 2024; Morris et al., 2023). The present study asks whether a finetuned LLM produces proficiency scores that agree with human raters (RQ1) and whether those scores measure the same underlying construct as human ratings (RQ2).

2 L2 Proficiency Assessment

Researchers typically view language proficiency as a multi-dimensional construct, with frameworks such as complexity, accuracy, and fluency (CAF, Larsen-Freeman, 1978) informing benchmarks like the Common European Framework of Reference for Languages, to which many large-scale assessments are aligned (Housen et al., 2012; Housen and Kuiken, 2009). Fully automated language assessments such as the Duolingo English Test (DET, Cardwell et al., 2023) and the Pearson Test of English Language Learning (Pearson TELL, Hidalgo et al., 2022) use proprietary natural language processing (NLP) algorithms to score spoken and written language samples, but none of the reviewed tests describe the use of large language models (LLMs) in their scoring procedures. Duolingo and ETS have both signaled that the use of LLMs for scoring would require strong validity evidence, with ETS stating that such use would require a “strong body of evidence to support its use” (Hao et al., 2024; Burstein, 2024, p. 25). These comments highlight the need for rigorous evaluations of the scores produced by LLMs.

2.1 Large Language Models in Language Assessment

In language research, representation LLMs have been finetuned to assess human judgements of learner language proficiency, to synthesize human ratings, and score essays. For example, [Crossley and Holmes \(2022\)](#) used LLMs to model English as a Second Language (ESL) learners' lexical sophistication, finding that LLMs were better aligned with human rater scores than more traditional, linguistic feature-based approaches. [Yee et al. \(2023\)](#) explored the potential for LLMs to classify texts according to structure, content, and sentiment, finding that they could use human annotators to label 1% of their data and then finetune an LLM that was capable of reliably annotating the remainder. [Morris et al. \(2025\)](#) found that several linguistic features, including discourse-level cohesion features, were more strongly associated with essay scores produced by finetuned LLMs than the scores produced by human raters, taken as evidence of "differential sensitivity," indicating that finetuned LLMs may weight discourse properties differently than trained human raters. [Mayfield and Black \(2020\)](#) compared finetuned LLMs to machine learning models that use document embeddings or n-grams as input features to score persuasive essays, finding that human interrater reliability was an "approximate upper bound on performance" for both automatic scoring methods.

3 Current Study

Research into the use of LLMs as automatic scoring systems is nascent, and it is not clear whether LLMs can contribute to automated assessment of language proficiency. Research generally indicates that automated scoring models, whether based on linguistic features models or LLMs, have demonstrated reliability when compared with human raters. However, [Attali \(2007\)](#) argues that agreement in observed scores is insufficient to assess the construct validity of automated scores because automated scores may suffer from construct underrepresentation or construct-irrelevant variance. This is to say that the source of disagreement is important for determining whether two sets of scores should have the same interpretation. Random measurement error can be corrected by averaging additional measurements, but systematic error indicates that the scores are measuring different constructs. Thus, establishing that auto-

rated scores have the same interpretation as scores produced by human raters requires decomposing sources of error and analyzing the construct representation of each set of scores.

Following [Kane's \(2013\)](#) argument-based approach to validity, the scoring inference for LLM-based assessment of language proficiency holds across rating methods only if substituting human raters with automated systems introduces random rather than systematic differences in scores.

The goal of this study is to examine the potential for LLMs to assign language proficiency scores to L2 learners in a manner similar to expert raters that is both reliable and valid. We first finetune two instances of DeBERTa-v3-large ([He et al., 2023](#)), a 435M-parameter open-weight encoder-only model, on different sets of human ratings. We then generate proficiency score predictions on a held-out test set and compare the models' predictions to both sets of human ratings. To make this comparison, we first test the interrater reliability between the LLM-based ratings and the human ratings to assess the reliability of the scoring methods. We then construct a confirmatory factor analysis (CFA) model to test whether the LLM-based ratings and human ratings measure the same underlying L2 proficiency construct. CFA separates measurement error from explainable variance. This provides a method for assessing construct representation in the presence of measurement error. This statistical approach has been successfully employed across a rich body of scale validation research (e.g., [Brown, 2015](#); [Cheung et al., 2023](#)). This study is guided by the following research questions:

1. Does an LLM finetuned on human ratings of L2 proficiency produce proficiency scores in agreement with human ratings?
2. Do LLM-produced proficiency scores measure the same underlying construct of language proficiency as human ratings?
 - (a) Do these measurement methods demonstrate adequate construct representation?
 - (b) Is there any evidence of divergence between these measurement methods?

4 Methods

4.1 Data

We used the full version of the open-source English Language Learning Insight, Proficiency, and

Skill Evaluation corpus (ELLIPSE, Crossley et al., 2024), which provides human ratings of proficiency for L2 English learners based on argumentative essays. The full ELLIPSE dataset contains all of the data prior to pruning and score averaging, and comprises 8,890 L2 English essays. Essays were written as part of state-wide and national standardized testing for the 8th–12th grade levels in the United States. All essays were written by L2 English learners the majority whom identified as Hispanic (72%) and as economically disadvantaged (70%). Each essay was written in response to one of 29 different prompts and learners were allowed 25–30 minutes to respond to the prompt on a computer.

Each essay was scored by two expert raters using an analytic rubric. The scoring rubric consisted of a holistic score, measuring the writer’s language proficiency, and six analytic traits related to cohesion, syntax, vocabulary, phraseology, grammar, and conventions. All traits were evaluated on a 5-point Likert-style scale. Several of these traits, particularly cohesion and text organization, are discourse-level properties of writing. The current study focuses on the holistic scores in the ELLIPSE dataset, which measured “a 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.” The full rubric is available in Crossley et al. (2024). There were 26 raters who scored 634 essays on average ($SD = 469.11$). The median and mode holistic score for both sets of raters was 3. The raters disagreed by plus or minus one for about 50% of the writing samples, and agreed on the exact score for the rest. There were two essays that resulted in a disagreement of more than one score point. The two scores for each essay were randomly assigned to a Rater 1 set or a Rater 2 set to prevent any individual rater from being overrepresented in either set.

4.2 LLM Selection

For our LLM, we selected DeBERTa because it was the largest freely available encoder-only model (He et al., 2023) at time of analysis. Our first step was to tokenize the texts using DeBERTa’s vocabulary, which is a closed set of 128,000 tokens. We limited the maximum sequence length to 512 to minimize computational costs. After tokenization, we found that approximately 40% of the ELLIPSE writing samples were longer than 510 tokens. Thus, we

truncated all texts to 510 tokens by trimming the right-hand side before adding in the classification and end-of-sequence tokens required by DeBERTa. We partitioned the data so that the performance of our model could be verified on a held-out portion of the data. We used 70% of the data for training, 15% for development, and 15% for testing. The development set was used for evaluating performance during model development, as described in Appendix A. We report results on the held-out testing set of 973 essays. We finetuned two DeBERTa models, one on each set of rater scores. The models were finetuned to minimize mean squared error, which is the standard choice of loss function for data with continuous labels and has also been shown to be effective for ordinal data (Gaudette and Japkowicz, 2009; Morris et al., 2024). The models output continuous scores, which we then converted back to ordered ranks such that the resulting distribution of ordinal scores matched the score distribution of the training data.

4.3 Analyses

In the present study, we tested whether human judgements of L2 proficiency measure the same underlying construct as LLM-based scoring models. For this, we used CFA, a type of structural equation model that tests hypotheses about relationships among measurements and constructs. In a test of construct representation, the primary advantage of CFA compared to a simpler approach such as an observed correlation is that CFA separates the latent construct of interest from measurement error.

To test the validity of our LLM-based scoring models, we specified two CFA models using the Lavaan (Rosseel, 2012) package in R and fit them to the held-out testing portion of the ELLIPSE data. The first was a model where both sets of human ratings and both language models loaded onto a single latent factor representing L2 proficiency. We followed up this analysis with a more stringent test of construct validity, which assessed whether the latent construct of L2 proficiency measured by human and language models was the same. This allowed us to derive a value for the “true, *unobservable* relation” between two constructs, accounting for random error (Bagozzi, 1981, p. 378). To test this hypothesis, we specified a second model with two latent factors, one for each scoring method. We then examined the covariance $\psi_{human, aes}$ between the two latent factors, fixing the variance to 1.0, which makes $\psi_{human, aes}$ a simple correlation.

Since this was a one-tailed hypothesis, if the 90% confidence interval for the estimated correlation included 1.0, then we would fail to reject the null hypothesis that the constructs are the same. If so, model specification with a single latent factor for L2 proficiency would be preferred. Rejection of the null hypothesis would suggest that there is some difference between the constructs being measured by human raters and our LLM raters.

To ensure proper model estimation, we addressed model identifiability (Brown, 2015), which is crucial for obtaining meaningful CFA results. For both models, we implemented two key constraints: first, we fixed the variance for each factor to 1.0, and second, we constrained the factor loadings for scores of the same type to be equal. These equality constraints were theoretically justified—“Rater 1” and “Rater 2” represent random collections of human scores expected to contribute equally to their underlying factor, and similarly, the LLMs trained on these respective collections should contribute equally to their factor. Through these specifications, both models achieved over-identification, and each sub-model within the two-factor model was independently just-identified. This identification strategy allowed us to reliably estimate each latent construct despite having only two indicators per factor in the second model, thereby enabling a robust comparison of the construct measured by human raters versus our LLM-based scoring models.

For both models, we report model fit statistics and standardized factor loadings. We used least squares estimators instead of maximum likelihood because it is appropriate for ordinal data (Li, 2016). L2 proficiency was measured on a 5-point ordinal scale. We fit using unweighted least squares (ULS), which does not require estimating a weight matrix and has been shown to yield less biased and more precise estimates than diagonally weighted least squares for ordinal indicators with few categories (Forero et al., 2009). We report standard model fit statistics for CFA but note that many of these statistics are influenced by the estimation method (Xia and Yang, 2019). Furthermore, we do not report the χ^2 test because the test statistic does not follow a χ^2 distribution when the model is fit with ULS (Blunch, 2008).

The standardized root mean square residual (SRMR) is the most robust metric of model fit when working with ordinal data and least squares estimators (Shi and Maydeu-Olivares, 2020). We

additionally report the root mean square error of approximation (RMSEA), and the comparative fit index (CFI). We relied on several rules of thumb to guide our interpretation of these statistics. A well-specified model should have SRMR < 0.06, RMSEA < 0.06, and CFI > 0.95 (Hu and Bentler, 1999). In addition to goodness of fit indices, we report the average variance extracted (AVE) and standardized factor loadings. All factor loadings should have AVE > 0.5, indicating that the latent factor explains at least 50% of the variance in the observed variables (Fornell and Larcker, 1981). Likewise, all standardized factor loadings should be greater than 0.5, or ideally greater than 0.7 (Hair, 2014). These tests assess the strength of the relationship between indicators and the latent construct of interest, language proficiency.

We also report a true-score correlation on the held-out testing set to facilitate comparisons with previous work on rating methods. The true-score correlation is calculated from the observed correlation, r_{xy} , and the reliabilities of each measure, r_{xx} and r_{yy} :

$$\hat{\rho}_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \cdot r_{yy}}} \quad (1)$$

We used a Pearson’s correlation to calculate the observed correlation in the mean scores for each rating method. We calculated the reliabilities as intra-class correlation (ICC) coefficients of absolute agreement, using a one-way ANOVA for human ratings (since each essay was scored by a different, randomly assigned pair of raters) and a two-way ANOVA for LLM ratings (since the same two rating models scored every sample).

5 Results

5.1 Interrater Reliability

Table 1 reports the rater-to-rater, model-to-rater, and model-to-model correlations on the held-out testing partition of the ELLIPSE corpus, using Kendall’s τ and quadratic weighted Kappa (QWK). The rater-to-rater correlation was $\tau = 0.53$. Model-to-rater estimates had confidence intervals that overlapped with the rater-to-rater interval when Rater 2 was involved but not when Rater 1 was involved, suggesting that the model-rater correlation was significantly different from the rater-rater correlation only for Rater 1. The model-to-model correlation was substantially higher at $\tau = 0.85$, indicating that the scoring models were more consistent than

humans even though they were trained on different sets of rater scores.

5.2 Construct Validity

One-factor Model. To address research question two, we constructed a one-factor CFA model depicted in Figure 1. The fit indices indicated good model fit, with CFI = 0.998, SRMR = 0.032, and RMSEA = 0.042 (90% CI = [0.018, 0.068]). The test of close fit (null hypothesis that RMSEA $\leq$ 0.05) yielded $p = 0.66$, so we do not reject close fit. Collectively, these tests support the construct validity of the four measures of learner language proficiency.

In addition to good model fit, we utilize two additional guidelines to evaluate the specified models. First, a latent construct should explain at least half of the variance in observed variables (Fornell and Larcker, 1981). The one-factor solution has an average variance extracted of 0.787, indicating that 78.7% of the variance in observed variables is explained by the latent proficiency score. Second, we found that all standardized factor loadings exceeded 0.7, revealing that the latent factor explains at least half of the variance in each observed variable. Additionally, the finetuned DeBERTa models had small error terms (.07), indicating that nearly all the variance in these scores is explained by the latent construct of proficiency. Standardized factor loadings and error variances are illustrated in Figure 1.

Collectively, these results support the construct validity of LLM-based language proficiency scores in the ELLIPSE dataset, with both human-produced and LLM-produced scores measuring (with error) the same latent L2 proficiency construct.

Two-Factor Model. An alternative method of assessing construct validity is to pose the question as a counterfactual. We constructed a model under the assumption that there are two distinct latent constructs, one for human-assessed proficiency and one for LLM-assessed proficiency. We then attempted to find evidence that these two constructs are divergent. Failure to demonstrate divergence would indicate that the latent constructs are not sufficiently distinct, leading to a preference to model them as a single latent construct.

The covariance between human-assessed proficiency and LLM-assessed proficiency is 0.902, 90% CI = [0.857, 0.947]. The variance of both

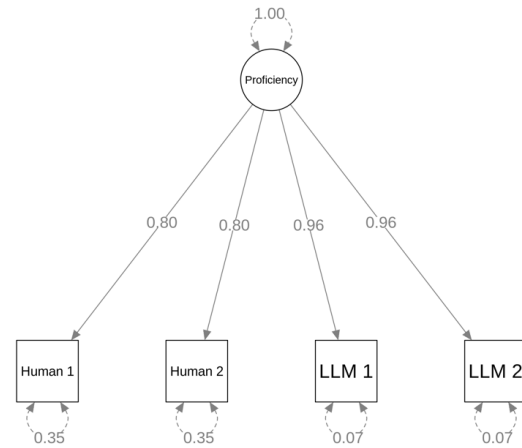


Figure 1: One-factor solution for human-generated and model-generated scores of language proficiency.

latent constructs was fixed to 1.0, so the covariance between them is a simple correlation. While a correlation of 0.9 would be considered strong by any standards, it is also significantly different than 1.0. In other words, the 90% confidence interval does not include perfect unity, so there is evidence of distinctness between human-assessed and LLM-assessed proficiency.

The goodness of fit statistics for the two-factor solution were uniformly superior to those of the one-factor solution, as shown in Table 2. We observe CFI = 1.000, SRMR = 0.013, and RMSEA = 0.0 (90% CI = [0.000, 0.041]), which indicates excellent model fit. The model passes the test of close fit, with a p -value of 0.985 (failure to reject the null hypothesis that RMSEA is $\leq$ 0.05). While both models exhibit good fit, the two-factor solution is better able to explain the observed data. This is consistent with the inter-construct correlation. Adding an additional latent factor would not improve model fit if the additional factor were perfectly correlated with an existing factor in the model. While there is substantial overlap in the latent constructs measured by each method, results from the CFA model indicate that these constructs are not the same.

To facilitate comparisons with previous work, we calculated the true-score correlation between each rating method using the reliability estimates provided by ANOVA. The observed correlation of the mean scores for each rating method was

Rating Method		τ	95% CI	QWK	95% CI
Rater 1	Rater 2	0.53	[0.50, 0.56]	0.60	[0.56, 0.63]
Rater 1	DeBERTa 1	0.61	[0.58, 0.64]	0.66	[0.63, 0.70]
Rater 1	DeBERTa 2	0.60	[0.57, 0.63]	0.65	[0.61, 0.68]
Rater 2	DeBERTa 1	0.57	[0.53, 0.60]	0.63	[0.60, 0.66]
Rater 2	DeBERTa 2	0.55	[0.51, 0.58]	0.61	[0.57, 0.64]
DeBERTa 1	DeBERTa 2	0.85	[0.83, 0.88]	0.88	[0.86, 0.90]

Table 1: Kendall’s τ correlations and QWK between human rater scores and DeBERTa scores of language proficiency on the test set. All agreement statistics are significant at $\alpha = 0.05$.

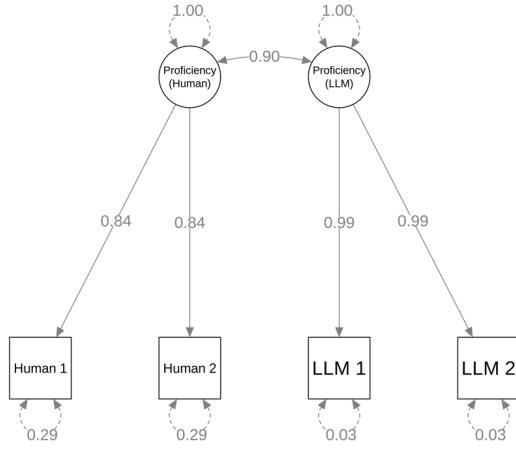


Figure 2: Two-factor solution where LLM-assessed proficiency and human-assessed proficiency are distinct constructs.

Model	CFI	RMSEA [90% CI]	SRMR
1-factor	0.998	.042 [.018, .068]	0.032
2-factor	1.000	.000 [.000, .041]	0.013

Table 2: Model fit statistics for a one-factor and a two-factor solution.

$r(1,324) = 0.738, p < 0.01$. The human-human reliability was 0.595 (95% CI = [0.559, 0.629]) and the model-model reliability was 0.878 (95% CI = [0.865, 0.890]). The true-score correlation was 1.0, which suggests that the two rating methods measure the same true score.¹ This contrasts with the CFA results, which provided evidence of small divergence in the constructs assessed by each rating method.

¹In the language of classical test theory, the expected rating for a given writing sample is the same whether rendered by human raters or finetuned LLMs.

6 Discussion

We trained two LLM-based models to automatically assess L2 learners’ writing samples. We first tested the reliability of the scores produced by these models by comparing correlations across raters and finetuned models. Results indicated that the LLM-based models were at least as reliable as human raters. The model-human correlations were not significantly different from the human-human correlation, which is well-aligned with previous work using automated scoring methods (Mayfield and Black, 2020).

However, the model-model agreement was significantly stronger than other correlations, indicating that the two models had high interrater reliability and thus less random error than human raters. This is unsurprising because the LLMs were finetuned on random collections of ratings from the same dataset and would learn roughly the same information from their respective partitions of the data. Additionally, LLM ratings are deterministic given the model parameters and input text, so all variance in the ratings is necessarily text-dependent, whereas human raters may experience distractions or mental states that influence their rating. In practice, the lower random error of the LLMs can be controlled for by averaging additional raters and does not pose the same threat to construct representation as does systematic error.

Our CFA analyses showed that the LLMs produced proficiency ratings that were well-aligned, but not perfectly aligned, with human ratings. The two-factor model fit better than the unidimensional model, and the correlation between the measured constructs was significantly different from 1.0, indicating that the two rating methods represent marginally divergent constructs. One possible source of this divergence is that the LLMs weight discourse-level features such as cohesion and text

organization differently than human raters (Morris et al., 2025). If the human ratings of proficiency are taken to be perfectly valid measures of language proficiency, then it can be said that there is systematic error in the LLM rating method. Even so, this evidence is not particularly strong, as Cheung et al. (2023) note “high correlations (e.g., 0.9) that are statistically significantly less than 1.0 are difficult to defend as indicating two distinct constructs” (p. 17). We argue that these results constitute weak statistical evidence for a two-construct model but should nonetheless serve to caution test developers against substituting human rater processes with LLM-based scoring procedures.

For comparison, we also calculated the true-score correlation, which was 1.0, somewhat higher than Attali and Burstein’s (2004) report of 0.93 for an automatic rating system. The CFA approach is a more stringent test of construct overlap because it decomposes variance into construct-relevant variance and measurement error before estimating the correlation between latent constructs, whereas the true-score correlation is a correction applied to the observed correlation. Our overall results indicate good, but imperfect, construct overlap between human-generated and LLM-generated scores.

An important qualification is that these results, in themselves, are insufficient for a validity argument for a high-stakes test. Kane (2013) argues that a comprehensive validity argument should evaluate proposed interpretations and uses of test scores within the specific assessment context, including the populations being tested, content structure, and rating procedures. Our study specifically investigates the rating procedure by examining how finetuned LLM-based assessments compare to human rating procedures within the ELLIPSE dataset. Our CFA results provide strong evidence for what Kane categorizes as the scoring inference: that both human and LLM-based scoring procedures can appropriately measure the same construct in this context. Future validation work would need to address additional inferences in Kane’s framework, including whether scores generalize beyond specific testing occasions and whether they predict real-world language ability.

Limitations

There are several limitations to the generalizability of these findings. First, the observed interrater reliability ($\tau = 0.53$) is lower than would be expected

in a high-stakes scenario, suggesting that raters may not have been consistently evaluating all the important aspects of language proficiency that we intended to measure. Second, the ELLIPSE dataset was collected from students enrolled in American schools, and we might expect different patterns of proficiency in other learner populations with different first language backgrounds or who are learning English as a foreign language. Finally, our finetuning approach involved truncating all samples to a fixed length, so the full text sample was not always provided to the finetuned model. We hope that future studies will reproduce our findings on additional datasets and with alternative finetuning approaches.

References

- Yigal Attali. 2007. [Construct validity of E-rater® in scoring TOEFL® essays](#). *ETS Research Report Series*, 2007(1).
- Yigal Attali and Jill Burstein. 2004. [Automated essay scoring with E-rater® V.2.0](#). *ETS Research Report Series*, 2004(2).
- Richard P. Bagozzi. 1981. [Evaluating structural equation models with unobservable variables and measurement error: A comment](#). *Journal of Marketing Research*, 18(3):375–381.
- Khaled Barkaoui. 2011. [Effects of marking method and rater experience on ESL essay scores and rater performance](#). *Assessment in Education: Principles, Policy & Practice*, 18(3):279–293.
- Lukas Biewald. 2020. [Experiment tracking with Weights and Biases](#).
- Niels Blunch. 2008. *Introduction to Structural Equation Modelling Using SPSS and AMOS*. SAGE Publications, Ltd.
- Timothy A. Brown. 2015. *Confirmatory Factor Analysis for Applied Research*, 2nd edition. Guilford Publications.
- Jill Burstein. 2024. [The Duolingo English Test responsible AI standards](#). Technical Report DRR-25-05, Duolingo.
- Robby Cardwell, Ben Naismith, Geoffrey LaFlair, and Steven Nydick. 2023. [Duolingo English Test: Technical manual](#). Technical report, Duolingo.
- Gordon W. Cheung, Helena D. Cooper-Thomas, Rebecca S. Lau, and Lin C. Wang. 2023. [Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations](#). *Asia Pacific Journal of Management*.

- Scott A. Crossley and Langdon Holmes. 2022. [Assessing receptive vocabulary using state-of-the-art natural language processing techniques](#). *Journal of Second Language Studies*, 6(1):1–28.
- Scott A. Crossley, Yong Tian, Patrick Baffour, Aigner Franklin, Yoon Kim, Wesley Morris, Matthew Benner, Aron Picou, and Ulrich Boser. 2024. [The English language learner insight, proficiency and skills evaluation \(ELLIPSE\) corpus](#). *International Journal of Learner Corpus Research*, 9(2):249–270.
- Paul Deane. 2013. [On the relation between automated essay scoring and modern views of the writing construct](#). *Assessing Writing*, 18(1):7–24.
- Carlos G. Forero, Alberto Maydeu-Olivares, and David Gallardo-Pujol. 2009. [Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation](#). *Structural Equation Modeling: A Multidisciplinary Journal*, 16(4):625–641.
- Claes Fornell and David F. Larcker. 1981. [Evaluating structural equation models with unobservable variables and measurement error](#). *Journal of Marketing Research*, 18(1):39–50.
- Lisa Gaudette and Nathalie Japkowicz. 2009. [Evaluation methods for ordinal classification](#). In *Advances in Artificial Intelligence*, pages 207–210. Springer.
- Joseph F. Hair. 2014. *Multivariate Data Analysis*, 7th edition. Pearson.
- Jiangang Hao, Alina A. von Davier, Victoria Yaneva, Susan Lottridge, Matthias von Davier, and Deborah J. Harris. 2024. [Transforming assessment: The impacts and implications of large language models and generative AI](#). *Educational Measurement: Issues and Practice*, 43(2):16–29.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). *arXiv preprint*.
- Pamela Hidalgo, Edynn Sato, and Miho Suzuki. 2022. *Pearson TELL: New technologies for assessing language learners*. Technical report, Pearson.
- Alex Housen and Folkert Kuiken. 2009. [Complexity, accuracy, and fluency in second language acquisition](#). *Applied Linguistics*, 30(4):461–473.
- Alex Housen, Folkert Kuiken, and Ineke Vedder. 2012. [Definitions, measurement and research: Complexity, accuracy and fluency](#). In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, pages 1–20. John Benjamins Publishing Company.
- Li-tze Hu and Peter M. Bentler. 1999. [Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives](#). *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1):1–55.
- David M. Hunter, Roger M. Jones, and Bikkar S. Randhawa. 1996. [The use of holistic versus analytic scoring for large-scale assessment of writing](#). *Canadian Journal of Program Evaluation*, 11(2):61–86.
- Lennart Jølle and Gustaf B. Skar. 2020. [“digging for gold” or “sticking to the criteria”: Teachers’ rationales when serving as professional raters](#). *Scandinavian Journal of Educational Research*, 64(2):283–299.
- Michael T. Kane. 2013. [Validating the interpretations and uses of test scores](#). *Journal of Educational Measurement*, 50(1):1–73.
- Diane Larsen-Freeman. 1978. [An ESL index of development](#). *TESOL Quarterly*, 12(4):439–448.
- Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2024. [On measurement validity and language models: Increasing validity and decreasing bias with instructions](#). *Communication Methods and Measures*.
- Cheng-Hsien Li. 2016. [Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares](#). *Behavior Research Methods*, 48(3):936–949.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Tom Lumley and Tim F. McNamara. 1993. [Rater characteristics and rater bias: Implications for training](#). Technical report.
- Elijah Mayfield and Alan W. Black. 2020. [Should you fine-tune BERT for automated essay scoring?](#) In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162.
- Wesley Morris, Scott A. Crossley, Langdon Holmes, and Amber Trumbore. 2023. [Using transformer language models to validate peer-assigned essay scores in massive open online courses \(MOOCs\)](#). In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 315–323.
- Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott Crossley. 2024. [Automated scoring of constructed response items in math assessment using large language models](#). *International Journal of Artificial Intelligence in Education*.
- Wesley Morris, Langdon Holmes, Joon Suh Choi, and Scott A. Crossley. 2025. [Uncovering differential sensitivity toward linguistic features of cohesion in large language models](#). In *Proceedings of the Twenty-Sixth International Conference on Artificial Intelligence in Education (AIED)*, Palermo, Italy.
- Les Perelman. 2020. [The BABEL generator and E-rater: 21st century writing constructs and automated essay scoring \(AES\)](#). *Journal of Writing Assessment*, 13(1).

- Chaitanya Ramineni, Catherine S. Trapani, David M. Williamson, Tim Davey, and Brent Bridgeman. 2012. [Evaluation of the E-rater® scoring engine for the TOEFL® independent and integrated prompts](#). *ETS Research Report Series*, 2012(1):i–51.
- Yves Rosseel. 2012. [lavaan: An R package for structural equation modeling](#). *Journal of Statistical Software*, 48(2):1–36.
- Dexin Shi and Alberto Maydeu-Olivares. 2020. [The effect of estimation methods on SEM fit indices](#). *Educational and Psychological Measurement*, 80(3):421–445.
- Stefanie A. Wind and Mark E. Peterson. 2018. [A systematic review of methods for evaluating rating quality in language assessment](#). *Language Testing*, 35(2):161–192.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.
- Yan Xia and Yanyun Yang. 2019. [RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods](#). *Behavior Research Methods*, 51(1):409–428.
- Mina Yee, Arjun Roy, Martin Perdue, Cristobal Cuevas, Keenan Quigley, Alexander Bell, Aakash Rungta, and Shigeru Miyagawa. 2023. [AI-assisted analysis of content, structure, and sentiment in MOOC discussion forums](#). *Frontiers in Education*, 8.

A Hyperparameter Optimization

Following He et al. (2023), we treated batch size, learning rate, warmup steps, number of training epochs, and dropout rate as the hyperparameters most likely to affect finetuning performance. We searched learning rates of 5e-6, 8e-6, 9e-6, and 1e-5; warmup steps of 50, 100, 500, and 1,000; 2, 3, and 4 training epochs; and dropout rates of 0.0, 0.15, and 0.30. Because the space of configurations is large, we used the Bayesian optimizer provided by Weights & Biases (Biewald, 2020) to search it over 42 training runs, selecting each new configuration based on the development-set performance of previous runs. The best-performing configuration had a learning rate of 1e-5, a dropout rate of 0.3, 100 warmup steps, and 2 training epochs, yielding a mean squared error of 0.318 on the development

set. No early stopping was used; the model from the final epoch of each run was evaluated.

The final configuration is reported in Table 3. Batch size, optimizer, and weight decay were left at their default values in the transformers library (Wolf et al., 2020). The batch size of 16 was set by the memory available on our GPU (an NVIDIA RTX A6000 with 48 GB of VRAM). The AdamW optimizer has been shown to generalize better to unseen data than stochastic gradient descent (Loshchilov and Hutter, 2019), and its weight decay of 0.01 matches the value used by He et al. (2023) during DeBERTa’s pretraining. We report these values, even though they were not optimized, to support reproducibility.

Parameter	Value	Source
Batch size	16	A priori
Loss function	Mean squared error	A priori
Optimizer	AdamW	A priori
Weight decay	0.01	A priori
Learning rate	1e-5	Optimized
Warmup steps	100	Optimized
Dropout	0.3	Optimized
Epochs	2	Optimized

Table 3: Final hyperparameter configuration for finetuning DeBERTa-v3-large on ELLIPSE data.