Validating Generative AI Scoring of Constructed Responses with Cognitive Diagnosis

Hyunjoo Kim
University of Illinois Urbana-Champaign
hyunjoo5@illinois.edu

Abstract

Generative AI has been investigated as a tool for scoring constructed responses (CRs). Although generative AI can provide both numeric scores and qualitative feedback on written tasks effectively and efficiently, its lack of transparency in output makes it challenging to build strong validity. Validity evidence for outputs from generative AI scoring is evaluated mainly through expert reviews or statistical concordance measures with human raters. As additional validity evidence for CR scores produced by generative AI, particularly for essay-type tasks, this research examines the feasibility of applying the Cognitive Diagnosis (CD) framework in psychometrics. The results of the study indicate that the classification information of CRs and item-parameter estimates from cognitive diagnosis models (CDMs) could provide a new perspective as additional validity evidence for CR scores and feedback from generative AI with less human oversight.

1 Introduction

Constructed responses (CRs), ranging from short sentences to full essays, are widely used to assess understanding of subject-matter knowledge. Unlike multiple-choice questions, which require selecting a given option, CRs ask examinees to generate answers in their own words. This task not only requires applying knowledge but also engaging in critical thinking and constructing logical arguments. As CRs can reveal a deeper level of understanding than simple recall questions, they purportedly provide strong evidence of the skills examinees possess. Accurately scoring CRs and providing personalized, detailed feedback is hence crucial, as it can support more subsequent learning (Gan and Liu, 2021; Kochmar et al., 2020; Parr and H.S., 2010). However, scoring CRs with high-quality feedback from human raters is labor-intensive, and issues such as fatigue and subjectivity are inevitable. These challenges limit the widespread use of CRs, especially in large-scale assessments.

Recent advances in generative AI, such as ChatGPT, have been explored as potential alternatives to address these limitations (Casabianca et al., 2025; Mizumoto and Eguchi, 2023; Yoon et al., 2023). Generative AI offers efficiency and scalability for CR scoring, but concerns about its sensitivity and lack of transparency make building strong validity evidence challenging. To date, most validity evidence for generative AI-based scoring systems has come from expert reviews or statistical concordance measures comparing AI outputs with human ratings or traditional Natural Language Processing (NLP) scores (Bui and Barrot, 2025a; Casabianca et al., 2025; Seßler et al., 2024; Tate et al., 2024).

As an additional source of validity evidence, this research draws on the Cognitive Diagnosis (CD) framework in psychometrics. CD is designed to evaluate examinees' mastery of instructional content and to provide feedback on their strengths and weaknesses in terms of learned and yet-to-be learned cognitive skills. This way, CD offers fine-grained diagnostic information, unlike traditional assessments that typically provide a single numeric score. It has been successfully applied to language assessments, including English writing, demonstrating its value in diagnosing both examinees' skills and the characteristics of test items (Lee and Sawaki, 2009; Mei and Chen, 2022). By bridging the CD framework with generative AI-based CR scoring, this study explores the extent to which CD can increase the interpretability and transparency of generative-AI scoring systems, thereby improving their validation.

Proceedings of the Artificial Intelligence in Measurement and Education Conference (AIME-Con) – Volume 2: Works in Progress, pages 166–177
October 27-29, 2025

2 Background

2.1 Generative AI for CR scoring and its validity evidence

The use of generative AI has the potential to automate CR grading by providing both quantitative scores and personalized feedback. Recent studies have investigated the performance of several generative AI systems for this purpose (Bui and Barrot, 2025b; Pack et al., 2024). However, a key distinction lies in the process by which traditional methods and generative AI produce scores. While human raters and traditional Natural Language Processing (NLP) approaches rely on trained expertise and handcrafted features, generative AI derives scores through processes that are largely opaque. This lack of transparency necessitates strong validity evidence for the use of generative AI in CR scoring.

Because the scoring process of generative AI is often difficult to interpret, one integral approach to validation has been to measure agreement between human raters and generative AI outputs. To date, the most common evaluation metrics have been Spearman rank correlation coefficients (r) and Quadratic Weighted Kappa (QWK) between scores assigned by generative AI and those assigned by human raters. The Spearman rank correlation coefficient measures the monotonic association between ordinal data, and QWK evaluates inter-rater agreement on ordinal scales by assigning heavier penalties for larger disagreements and lighter penalties for smaller ones, with penalties increasing quadratically as rating differences grow. Landis and Koch (1977) provided interpretive benchmarks for QWK: below 0.0 = Poor; 0.00–0.20 = Slight; 0.21–0.40 = Fair; 0.41–0.60 = Moderate; 0.61–0.80 = Substantial; 0.81–1.00 = Almost perfect. Although these metrics reflect the degree of agreement and thus serve as evidence of validity, they are limited. Specifically, they capture only overall agreement without offering insight into the reasoning behind scores or into systematic differences between responses that receive similar scores. Hence, complementary approaches are necessary to provide stronger validity evidence of generative AI's outputs.

2.2 Cognitive diagnosis models and their use in writing assessments

Cognitive Diagnosis (CD) provides fine-grained information about examinees' mastery of specific cognitive skills by describing ability in a knowledge domain as a composite of K (specific) binary skills, called "attributes" (denoted α_k, $k = 1, 2, ..., K$) (Sessoms and Henson, 2018; von Davier and Lee, 2019). These attributes form profiles that define proficiency classes, written as $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_K)'$–mastered (1) or not (0)– with 2^K possible combinations.

Cognitive diagnosis models (CDMs) classify examinees into one of these proficiency classes based on their observed item responses, Y_{ij}, where $Y_{ij} = 1$ if examinee $i = 1, 2, ..., N$ answered item $j = 1, 2, ..., J$, $Y_{ij} = 0$ correctly, and $Y_{ij} = 0$ otherwise. As such, CDMs are restricted latent class models in which the latent constructs are attributes, and the discrete latent classes are examinees' proficiency classes.

Similarly, test items are also characterized by K-dimensional attribute profiles $\mathbf{q}_j$, which specify the attributes required for a correct response. The q_{jk} entries indicate whether the jth item requires the kth attribute (1) or not (0). These $\mathbf{q}$-vectors form a "Q-matrix" of size $J \times K$ (Tatsuoka, 1985).

Using the observed item-response data $\mathbf{Y}_{N \times J}$ and Q-matrix $\mathbf{Q}$, typically provided by domain experts, CDMs estimate each examinee's attribute profile by selecting the most likely class. The functional relation between attribute mastery and the probability of a correct response to an item, $P(Y_j = 1|\boldsymbol{\alpha})$, is modeled according to the chosen CDM, which specifies how attributes interact with items. Various CDMs differ in complexity and assumptions (e.g., compensatory vs non-compensatory; conjunctive vs disjunctive), but they all share the core principle that the probability of success on an item depends on the interaction between an examinee's attribute profile and the item's $\mathbf{q}$-vector (Henson et al., 2009; Williamson, 2024; Dibello et al., 2006; Rupp and Templin, 2008). The selection of a CDM depends on assessment goals.

Several CDMs have been applied to writing assessments, with a focus on English as a Foreign Language (EFL) (Effatpanah et al., 2019; He et al., 2021; Kim, 2011; Shi et al., 2023; Xie, 2016). Unlike typical applications of CD, which use direct item responses (correct/incorrect), writing assessments first require transforming examinees' constructed responses into binary response data. Prior CD studies on writing have commonly used rating checklists for this transformation, a task done by writing experts. Kim (2011) developed an empirically derived, descriptor-based (EDD) checklist of 35 fine-grained descriptors measuring five constructs–content fulfillment, organizational effectiveness, grammatical knowledge, vocabulary use, and mechanics (see Appendix A.2 for the descriptors of EDD). In her study, they had ten ESL (English as a second language) teachers assess 480 TOEFL (Test of English as a Foreign Language) essays using the EDD checklist and converted them into binary item-response data. Then, using the reduced reparameterized unified model (R-RUM) (Hartz, 2002), they diagnosed learners' writing ability and demonstrated that the CD framework could reliably identify strengths and weaknesses. Building upon this work, Xie (2016) applied the same EDD checklist with R-RUM in a different assessment context and found that it provided more diagnostic information than raw scores alone. The checklist was further validated using the additive CDM (ACDM) (de la Torre, 2011) and R-RUM (Effatpanah et al., 2019; Shahsavar, 2019). More recently, He et al. (2021) employed polytomous CDMs to extend binary models, diagnosing English learners' writing ability with a different checklist measuring four constructs. Under this approach, each constructed response is treated as an "examinee," each checklist descriptor as an "item," and each latent construct targeted in the rubric as an "attribute." When a response meets a checklist criterion, $Y_{ij} = 1$; otherwise, $Y_{ij} = 0$. With this mapping, CDMs estimate each constructed response's attribute profile/proficiency class, indicating which rubric-defined attributes are present.

3 Methodology

3.1 Dataset

Ideally, this study would have used essay datasets from previous writing assessment studies, such as the TOEFL independent writing, along with the corresponding binary item-response data coded by expert raters using the empirically derived descriptor-based diagnostic (EDD) checklist and the Q-matrix. However, because those datasets were not publicly accessible, we sought an open essay dataset that closely resembled the formats and tasks used in prior research, so that the EDD checklist and Q-matrix developed by (Kim, 2011) and subsequent studies could still be applied.

As a surrogate, we identified Set 2 from the Kaggle "The Hewlett Foundation: Automated Essay Scoring (AES)" dataset[1]. Set 2 consists of persuasive essays averaging 350 words, written by grade 10 native English speakers. The scoring scale ranged from 2 to 10, with two domains: Writing Applications (1-6) and Language Conventions (1-4). A detailed comparison of the formats between TOEFL Independent Writing and AES Set 2 is presented in Table 1.

The dataset contains 1,800 essays, each scored by two human raters, along with a detailed rubric and exemplar essays for each score. Based on the prior studies, we selected 500 essays from Set 2 for analysis. The specific essay prompt is provided in Appendix A.1.

Table 1: Comparison between TOEFL Independent Writing Task and AES Set 2 Dataset

Feature	TOEFL Ind.	AES Set 2
Task type	Persuasive "agree/disagree"	Persuasive "agree/disagree"
Avg length	300–400 words	350 words
Time limit	30 minutes	Not specified
Population	EFL learners (Primarily ages 21–25)	Grade 10 native English speakers
Score scale	1–5	2–10
# of Raters	2	2
Domains	Single holistic score	Writing Applications (1–6) + Language Conventions (1–4)

3.2 Procedure

We used the auto version of ChatGPT-5 from OpenAI, with the default temperature setting (which controls the variability of responses) for all stages of this study.

First, ChatGPT-5 was customized with the instructions to act as a rater evaluating student essays. The essay scoring guide and rubric file, including exemplar essays from the original AES website, were provided via Retrieval Augmented Generation (RAG). This setup enabled ChatGPT-5 to assign numerical scores and generate qualitative feedback of the selected 500 essays. The scoring prompt was adapted from Casabianca et al. (2025) and tailored to this study (see Appendix A.1). In this setting, generation followed a zero-shot Chain of Thought (CoT) approach, since no worked examples were included. Each essay was scored using a single

[1]https://www.kaggle.com/c/asap-aes/data

prompt that requested both the score and feedback.

The major challenge was the absence of binary item-response data for the AES dataset based on the EDD checklist, which are required inputs for CDMs. Without resources to recruit expert raters to transform essays into item-response data, we relied on ChatGPT-5 for this task. To mitigate concerns about reliability, we used a few-shot Chain of Thought (FsCoT) design. Three essays, drawn from outside the 500-essay sample, were manually rated with the EDD checklist and provided as exemplars in the prompt to ChatGPT-5. The model was then asked to generate a binary vector of size 1×35 for each essay, indicating whether each descriptor was met. This task was conducted in a separate session from the scoring task. The full prompt for this task is included in Appendix A.2.

This process yielded a 500×35 item-response matrix, ready for CDM analysis with the Q-matrix by Kim (2011). The Q-matrix specifies five latent attributes that the EDD checklist aims to measure: Content Fulfillment (CON), Organizational Effectiveness (ORG), Grammatical Knowledge (GRM), Vocabulary Use (VOC), and Mechanics (MCH). Based on the initial Q-matrix by Kim (2011), we empirically validated it using the method suggested by de la Torre and Chiu (2016), as in previous studies, and refined it to reduce misspecification and enhance CDM performance for datasets used in the study.

3.3 Data Analysis

The constructed item-response matrix and the refined Q-matrix were used as inputs for the Reduced Reparameterized Unified Model (R-RUM), which has been widely applied in prior CDM-based writing assessment research. The item response function of the R-RUM is given by:

$$P(Y_{ij} = 1|\boldsymbol{\alpha}_i) = \pi_j^* \prod_{k=1}^{K} r_{jk}^{*q_{jk}(1-\alpha_{ik})},$$

where $0 < \pi_j^* < 1$ represents the probability of a correct response when examinee i has mastered all attributes required by item j, and $0 < r_{jk}^* \leq 1$ is the penalty parameter for not mastering attribute k. π_j^* and r_{jk}^* correspond to item difficulty and item discrimination, respectively (Kim, 2011). Values of $\pi_j^* < 0.6$ suggest overly difficult items, and $r_{jk}^* < 0.5$ indicates that an item discriminates well between mastery and non-mastery on skill k (Roussos et al., 2007).

The validity evidence for the generative AI scoring was evaluated by examining the following:

1. the agreement between ChatGPT scores and human rater scores via Spearman correlation and QWK

2. the consistency between ChatGPT scores & feedback and the attribute mastery profiles estimated by the CDM.

For example, generative AI scoring can be supported as valid if essays grouped into the same proficiency class by CDMs also receive consistent scores from ChatGPT, and if the qualitative feedback aligns with mastered and non-mastered attributes.

We also analyzed the estimated item parameters from R-RUM to enhance the interpretability of ChatGPT's scoring process. This analysis allowed us to examine the extent to which latent attributes (from the Q-matrix) or items (descriptors from the checklist) influenced ChatGPT's scoring. For example, two constructed responses might receive the same overall score from ChatGPT, yet differ in their attribute profiles or in the number of mastered attributes, providing insight into how ChatGPT's scoring reflects specific attributes.

Although generating item-response datasets with expert raters would be more rigorous in practice, our approach demonstrates the feasibility of using CDMs to evaluate generative AI scoring in greater depth and to strengthen validity evidence by moving beyond numerical scores to attribute-level reasoning.

4 Results

As a preliminary check, inter-rater reliability between two human experts was examined using Spearman correlation coefficients and quadratic weighted kappa (QWK). Both indices were 0.82, indicating strong agreement. A t-test comparing the two raters' mean raw scores showed no statistically significant difference (p-value $= 0.55$).

Next, the validity of ChatGPT-5 scores was evaluated against the human ratings. For total scores, the Spearman coefficient was 0.67 and QWK was 0.56, indicating moderate agreement. At the domain level, Domain 1 showed 0.58 (Spearman) and 0.55 (QWK), while Domain 2 was lower–0.56 (Spearman) and 0.26 (QWK). The relatively weak agreement in Domain 2 reflects ChatGPT's tendency to score more strictly on language conventions.

The cross-tabulation of human mean scores and ChatGPT-5 scores shown in Table 2 confirms this pattern. ChatGPT frequently assigned scores about one point lower than the human raters, particularly in the mid-range of the scale.

Table 2: Agreement matrix between human mean total scores (Hu) and ChatGPT-5 (AI) total scores

Hu \AI	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	1	2	0	0	0	0	0	0	0
2.5	0	2	0	0	0	0	0	0	0
3	1	1	2	0	0	0	0	0	0
3.5	1	1	2	0	0	0	0	0	0
4	0	4	3	4	1	0	0	0	0
4.5	0	5	6	2	1	0	0	0	0
5	0	4	5	14	5	0	0	0	0
5.5	0	2	1	8	1	1	0	0	0
6	0	7	10	33	34	10	5	0	1
6.5	0	2	2	12	16	7	4	0	0
7	0	2	4	20	30	15	11	1	0
7.5	0	0	3	3	9	14	12	1	0
8	0	0	0	12	27	28	46	9	3
8.5	0	0	0	0	1	2	11	3	1
9	0	0	0	1	0	0	4	3	0
10	0	0	0	0	0	0	0	0	1

Applying R-RUM, essays were classified into proficiency profiles (Table 3). Fourteen of the 32 possible classes were observed, with the distribution concentrated in higher-mastery profiles. Across all classes, ChatGPT-5 assigned lower average scores than human raters. Profiles with fewer mastered attributes tended to receive lower scores overall, while broader mastery was associated with higher scores from both humans and AI.

Table 3: Summary of human (Hu) and ChatGPT-5 (AI) scores by proficiency class

Class	**Count**	**Hu Mean (SD)**	**AI Mean (SD)**
(00000)	25	4.64 (1.56)	3.96 (1.67)
(01000)	6	6.58 (0.58)	5.33 (0.63)
(00100)	42	5.15 (1.27)	4.05 (0.96)
(00001)	8	6.94 (0.82)	6.38 (1.19)
(11000)	4	7.00 (1.83)	5.75 (1.50)
(10001)	3	8.00 (0.00)	6.33 (1.53)
(01100)	67	6.37 (0.96)	5.30 (0.98)
(00101)	14	5.50 (1.14)	5.14 (0.86)
(11100)	73	6.84 (0.89)	5.60 (1.04)
(11010)	1	7.00 (NA)	6.00 (NA)
(01101)	55	6.74 (1.08)	5.98 (0.83)
(11110)	9	7.06 (1.10)	6.78 (1.39)
(11101)	39	6.90 (0.93)	6.82 (0.97)
(11111)	154	7.64 (0.80)	7.49 (1.20)

Beyond serving as validity evidence for ChatGPT's scores, the estimated proficiency classes of essays also aid in the interpretation of ChatGPT's scoring with respect to mastery status. Table 4 presents three essays' estimated proficiency classes, scores, and ChatGPT feedback. Essays from higher-level classes received feedback acknowledging strengths, while essays from lower-level classes received feedback identifying weaknesses consistent with their attribute profiles. For example, Essay 3213–mastering CON, ORG, GRM, and MCH but not VOC–received relatively positive comments, with vocabulary flagged as the main area for improvement. In contrast, Essays 2995 and 3103, from lower classes, received feedback emphasizing underdeveloped reasoning, weak transitions, and frequent grammar issues.

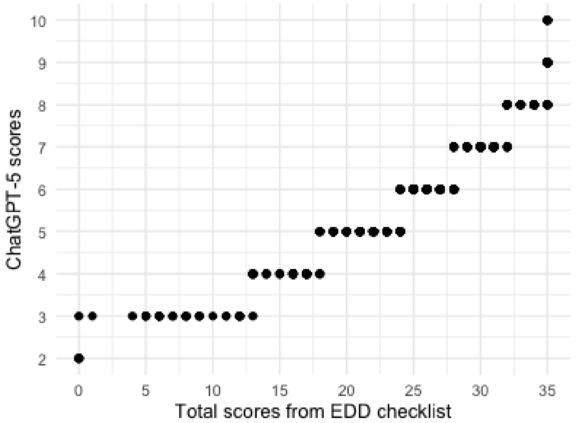

Figure 1: Total EDD checklist score vs. ChatGPT-5 score

Furthermore, even when two essays shared the same AI and/or human score, their estimated proficiency classes show more fine-grained diagnostic information with different combinations of attributes. The high alignment between classification results and ChatGPT feedback strengthens the point that the generative AI's comments are sufficiently valid to guide targeted revisions.

Figure 1 further illustrates the relation between EDD checklist totals (0–35) and ChatGPT-5 scores (2–10). A clear positive trend emerges: as more checklist items were met, ChatGPT assigned higher scores. This convergence with the checklist also supports the construct validity of ChatGPT's scoring.

Drawing on the CD framework provides further insight into factors that may influence ChatGPT's scoring. Based on the R-RUM estimates, the proportion mastering each attribute was: CON$= 0.57$, ORG$= 0.82$, GRM$= 0.91$, VOC$= 0.33$, and MCH$= 0.55$. Thus, many essays satisfied GRM

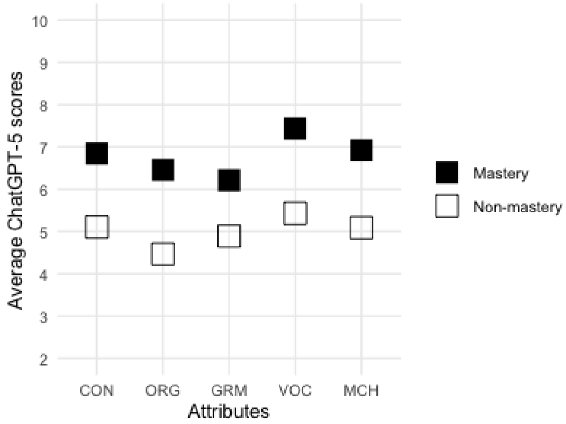

Figure 2: Average ChatGPT-5 scores by mastery status for each attribute

but not VOC (e.g., limited vocabulary variety). As shown in Figure 2, ChatGPT scores were higher when an attribute was mastered, with the largest mastery-non-mastery gap for VOC–indicating that vocabulary mastery was especially influential on ChatGPT-5 scoring.

The checklist-descriptor parameter estimates (Table 5) corroborate these patterns. Descriptors linked to VOC exhibit comparatively strong discrimination, consistent with the larger VOC-related score gap in Figure 2. Based on Table 5, most descriptors showed high π^* values (near 1) in general, indicating that once the relevant attributes were mastered, the probability of meeting the corresponding criteria was very high. Some $\gamma*$ parameters showed little or no penalty, implying that certain descriptors were less diagnostic of underlying mastery.

Overall, ChatGPT-5 scoring agrees reasonably well with human raters, although it tends to be more strict, particularly in certain domains. Together with numerical scores, its narrative feedback corresponds meaningfully with latent attribute (non)mastery, showing where specific strengths and weaknesses lie in each essay.

5 Discussion

This study aimed to adapt psychometric modeling, grounded in cognitive diagnosis theory, to introduce new forms of validity evidence for generative-AI scoring. Specifically, we demonstrated how cognitive diagnostic models and checklist-based item-response data of essays can serve as a framework for investigating ChatGPT-generated scores at both the holistic and attribute levels.

Several limitations should be noted. First, we did not examine the consistency of AI scores across multiple time points. Given that large language models are sensitive to prompt wording and contextual framing, test–retest reliability remains an open question. Future work should estimate intraclass correlation coefficients (ICCs) to evaluate score stability over repeated administrations (Seßler et al., 2024).

Second, our results may be suboptimal because (i) the EDD checklist was not originally designed for the essay samples in this study, and (ii) our artificial item-response dataset has not undergone validation. Access to established item-response datasets from prior studies would provide stronger grounding and allow more robust validation of this approach. Moreover, a formal implementation would require sustained expert involvement–from rubric design and Q-matrix specification to interpreting constructed-response evaluations.

Despite these limitations, the findings are promising. While challenges remain in applying CDMs within AI-automated assessment, the results suggest that such models can enhance the transparency of AI scoring. By linking scores to specific latent attributes, this framework provides an additional source of validity evidence-—helping explain not just what score was assigned, but also why.

In sum, this study demonstrates the feasibility of using psychometric frameworks to support the validation of AI-generated scores. By combining the interpretive strengths of cognitive diagnosis with the efficiency of generative AI, this approach offers a novel pathway toward transparent, evidence-based scoring systems in educational assessment.

Future research should extend this work in several directions. Different CDMs could be compared by model fit across diverse testing contexts. Beyond extended essays, the approach could also be adapted to shorter constructed-response tasks, such as sentence-level prompts. For these tasks, the checklist and Q-matrix would be smaller and easier to operationalize, reducing reliance on expert judgment while still producing meaningful validity evidence.

References

N. M. Bui and J. S. Barrot. 2025a. Chatgpt as an automated essay scoring tool in the writing classrooms:

how it compares with human scoring. *Education and Information Technologies*, 30:2041–2058.

N. M. Bui and J. S. Barrot. 2025b. Using generative artificial intelligence as an automated essay scoring tool: a comparative study. *Innovation in Language Learning and Teaching*, pages 1–16.

J. M. Casabianca, D. F. McCaffrey, M. S. Johnson, N. Alper, and V. Zubenko. 2025. Validity arguments for constructed response scoring using generative artificial intelligence applications. *Preprint*, arXiv:2501.02334.

J. de la Torre. 2011. The generalized dina model framework. *Psychometrika*, 76:179–199.

J. de la Torre and C-Y. Chiu. 2016. A general method of empirical q-matrix validation. *Psychometrika*, 81:253–273.

L.V. Dibello, L.A. Roussos, and W.F. Stout. 2006. Review of cognitively diagnostic assessment and a summary of psychometric models. In C.R. Rao and S. Sinharay, editors, *Handbook of Statistics*, pages 979–1030. Elsevier.

F. Effatpanah, P. Baghaei, and A.A. Boori. 2019. Diagnosing efl learners' writing ability: a diagnostic classification modeling analysis. *Language Testing in Asia volume 9*, 9, 12.

Z. Gan and F. Liu. 2021. Teacher feedback practices, student feedback motivation, and feedback behavior: How are they associated with learning outcomes? *Frontiers in Psychology*, 12:697045.

S. Hartz. 2002. *Bayesian framework for the Unified Model for assessing cognitive abilities: blending theory with practice*. Ph.D. thesis, University of Illinois at Urbana-Champaign.

L. He, Z. Jiang, and S. Min. 2021. Diagnosing writing ability using china's standards of english language ability: Application of cognitive diagnosis models. *Assessing Writing*, 50:100565.

R.A. Henson, J.L. Templin, and J.T. Wilse. 2009. Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika*, 74:191–210.

Y-H. Kim. 2011. Diagnosing eap writing ability using the reduced reparameterized unified model. *Language Testing*, 28(4):509–541.

E. Kochmar, D. Vu, R. Belfer, V. Gupta, I. Serban, and J. Pineau. 2020. Automated personalized feedback improves learning gains in an intelligent tutoring system. *Artificial Intelligence in Education*, 12164:140–146.

J.R. Landis and G.G. Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174.

Y.W. Lee and Y. Sawaki. 2009. Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly*, 6(3):172–189.

H. Mei and H. Chen. 2022. Cognitive diagnosis in language assessment: A thematic review. *RELC Journal*, 55(2):529–537.

A. Mizumoto and M. Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2:100050.

A. Pack, A. Barrett, and H. Escalante. 2024. Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence*, 6:100234.

J.M. Parr and Timperley H.S. 2010. Feedback to writing, assessment for teaching and learning and student progress. *Assessin Writing*, 15:68–85.

L. A. Roussos, L. V. DiBello, W. Stout, S. M. Hartz, R. A. Henson, and J. L. Templin. 2007. *Cognitive Diagnostic Assessment for Education: Theory and Applications*, chapter The fusion model skills diagnosis system. Cambridge University Press.

A.A. Rupp and J. Templin. 2008. Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research and Perspectives*, 6(4):219–262.

K. Seßler, M. Fürstenberg, B. Bühler, and E. Kasneci. 2024. Can ai grade your essays? a comparative analysis of large language models and teacher ratings in multidimensional essay scoring. *Preprint*, arXiv:2411.16337.

J. Sessoms and R. Henson. 2018. Applications of diagnostic classification models: A literature review and critical commentary. *Measurement: Interdisciplinary Research and Perspectives*, 16:1–17.

Z. Shahsavar. 2019. Diagnosing english learners' writing skills: A cognitive diagnostic modeling study. *Cogent Education*, 6(1).

X. Shi, X. Ma, W. Du, and X. Gao. 2023. Diagnosing chinese efl learners' writing ability using polytomous cognitive diagnostic models. *Language Testing*, 41(1):109–134.

T.P. Tate, J. Steiss, D. Bailey, S. Graham, Y. Moon, D. Ritchie, W. Tseng, and M. Warschauer. 2024. Can ai provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

K.K. Tatsuoka. 1985. A probabilistic model for diagnosing misconception in the pattern classification approach. *Journal of Educational and Behavioral Statistics*, 12:55–73.

M. von Davier and Y.S. Lee. 2019. *Handbook of diagnostic classification models*. Springer International Publishing, Cham, CH.

J. Williamson. 2024. Cognitive diagnostic models and how they can be useful. *Cambridge University Press & Assessment*.

Q. Xie. 2016. Diagnosing university students' academic writing in english: Is cognitive diagnostic modelling the way forward? *Educational Psychology*, 37(1):26–47.

S-Y. Yoon, E. Miszoglad, and L.R. Pierce. 2023. Evaluation of chatgpt feedback on ell writers' coherence and cohesion. *Preprint*, arXiv:2310.06505.

A Appendix

A.1 ChatGPT prompt for scoring essays

A student is assigned to an essay question as follows:

"All of us can think of a book that we hope none of our children or any other children have taken off the shelf. But if I have the right to remove that book from the shelf—that work I abhor—then you also have exactly the same right, and so does everyone else. And then we have no books left on the shelf for any of us." (Katherine Paterson Author)

Write a persuasive essay for a newspaper reflecting your views on censorship in libraries. Do you believe that certain materials, such as books, music, movies, magazines, etc., should be removed from the shelves if they are found offensive? Support your position with convincing arguments from your own experience, observations, and/or reading.

In addition to the rubric files in the "File" folder, the following are the specific guidelines for scoring the essay question above. Grade the essay based on all the rubrics provided. Remember that there are two types of scores: Domain 1 (Writing Applications) and Domain 2 (Language Conventions). As mentioned in the rubrics, give a score from 1 to 6 for Domain 1 and a score from 1 to 4 for Domain 2. Then, provide a 20-30 word feedback highlighting strengths and areas for improvement. Return the response in a JSON format of:

{ Score for Domain 1, Score for Domain 2, "Reasons": [{ reasons }] } The reasons should be an array of 3 objects. Each object should be in the structure shown above and described below. For each object in the reasons array, a reason must be provided. This reason should be one of the reasons for giving the score based on the rubric. The reason should not be a full sentence but in 20-30 words and be suitable to be displayed as bullet points to a person with a high school-college-level education, rather than copied directly from the rubric.

This is the end of the explanation. Now, I'll give you an essay from each student.

A.2 ChatGPT prompt for the checklist

You are an essay rater. Using the provided 35-item EDD checklist below, evaluate 10th-grade essays. For each item on the checklist, determine if the essay satisfies the criterion.

The EDD checklist is presented as follows:

1. This essay answers the question.
2. This essay is written clearly enough to be read without having to guess what the writer is trying to say.
3. This essay is concisely written and contains few redundant ideas or linguistic expressions.
4. This essay contains a clear thesis statement.
5. The main arguments of this essay are strong.
6. There are enough supporting ideas and examples in this essay.
7. The supporting ideas and examples in this essay are appropriate and logical.
8. The supporting ideas and examples in this essay are specific and detailed.
9. The ideas are organized into paragraphs and include an introduction, a body, and a conclusion.
10. Each body paragraph has a clear topic sentence tied to supporting sentences.
11. Each paragraph presents one distinct and unified idea.
12. Each paragraph is connected to the rest of the essay.
13. Ideas are developed or expanded well throughout each paragraph.
14. Transition devices are used effectively.
15. This essay demonstrates syntactic variety, including simple, compound, and complex sentence structures.
16. This essay demonstrates an understanding of English word order.
17. This essay contains few sentence fragments.
18. This essay contains few run-on sentences or comma splices.
19. Grammatical or linguistic errors in this essay do not impede comprehension.
20. Verb tenses are used appropriately.
21. There is consistent subject-verb agreement.
22. Singular and plural nouns are used appropriately.
23. Prepositions are used appropriately.

24. Articles are used appropriately.
25. Pronouns agree with referents.
26. Sophisticated or advanced vocabulary is used.
27. A wide range of vocabulary is used.
28. Vocabulary choices are appropriate for conveying the intended meaning.
29. This essay demonstrates facility with appropriate collocations.
30. Word forms (noun, verb, adjective, adverb, etc) are used appropriately.
31. Words are spelled correctly.
32. Punctuation marks are used appropriately.
33. Capital letters are used appropriately.
34. This essay contains appropriate indentation.
35. Appropriate tone and register are used throughout the essay.

Score each criterion with a binary value—1 if the essay meets the criterion or 0 if the essay does not. Your final evaluation should be a binary vector of size (1 x 35).

Let me give some examples of a student's essays and the corresponding evaluations:

Example 1: Certain materials being removed from libraries such as books, music and magazines, shouldn't be removed from the libraries. It gives people a chance to understand how the real world @CAPS2. Having certain materials such as books and music definitly should not be removed, because most books and music can show most people how bad the statement in the book @CAPS2 or how bad the lyrics are in a song, and help that person to avoid that type of thing that the book or song @CAPS2 saying to the reader or listener. People should give every type of music at least a try and not always doubt what they hear about what people say about that type of music. I always hear about people saying how bad the band @PERSON1 A.M. @CAPS2, just because in the lyrics it talks about drugs and how much cursing each song has. Really the band @CAPS2 talking about one mans life and how he turns his life from being a drug addict to having the best life someone could ever live. People always doubted him and never gave his music a chance. Another example would be @PERSON1's book, '@CAPS1 @CAPS2 @CAPS3 @CAPS4' for it talks about drug addicts, homeless people, people who have been born with disfigured arms or even someone who lost there legs, and telling how beautiful each and everyone of them really are. His book taught me a few things and made me think different about people. It doesn't matter how they look or how they talk, no matter what, that person @CAPS2 beautiful. As far as movies and magazines has gone within the last few years, I think that the also shouldn't be taken from libraries. I think @CAPS1 for the same reason of how I feel about the books and music. Of course we see previews of movies and think that they @MONTH1 not be good, but libraries shouldn't keep leave them out. Movies @CAPS2 a great way to learn how to treat others and how to act around other people when you don't know how to act. If you act differently around people that you've never been around before, then you could feel embarassed or maybe even get @CAPS4. Movies can help people learn about the real world by seeing how to do those type of things as we get older. Same goes with the magazines, they also help people see what not to do or to help them understand the consequences of something that shouldn't be done. Knowing what to do from a magazine could possible save your life or perhaps maybe even someone elses life. I don't understand why some libraries would want to banned certain materials to help people understand the things that happen in someone elses life and to help them not make the same mistakes as that person once did.

Evaluation based on the EDD checklist: $(1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1)$

Example 2: Do you believe that certain should be removed i think so be no that yes i think should no person that in chager the book, music, movies, magazines, ect., that be no agure why do i think if you need that please think i no thank you please if i need why do we if know that if i failure the this test i who need to graduate please the children allow to home please yes. Why do we need to be a prafece person please why do we need to do this why write this assgiment because you mean to be the best teaches ever and ever facebook is my password is @PERSON1 @NUM1 that why i need my myspace is the same thingh but different at same time please know that i need to know i really i need to my e-mail address is @EMAIL1 that is my e-mail please work m

Evaluation based on the EDD checklist: $(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0)$

Example 3. Do you think that libraries should remove certain materials off the shelves? People have different oppions, of whats good and whats bad. I have read and seen a lot of books in my life *time. I hear people telling me, 'oh dont read that book its a bad book.'But I ask myself, @CAPS2 do I know it's a bad book when I haven't even given it a chance?' @CAPS1 are some books, music, movies, and magazines out @CAPS1 that are offensive. Yet we still want to read, listen, watch, and look at them. If we tried to remove all the offensive books, from the libraries we wouldn't have anything left on the shelves. Katherine Paterson said, 'If I have the right to remove that book from the shelf that work i abhor- then you also have exactly the same right and so does everyone else. And then we have no books left on the shelf for any of us.' Katherine Paterson makes a great point out of her quote. Why should we have to remove a book if just some people think its offensive? Ask yourself the question again, '@CAPS2 do you know it's a bad book when you haven't even given it a chance?' @CAPS3't judge a book by what you hear. Find out what your own oppion is*

Evaluation based on the EDD checklist: $(1, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1)$

Refer to the above evaluation examples, rate the provided essay in the same form using the EDD checklist. When you evaluate, ignore words starting with an "@" symbol in essays, as it is marked for privacy and copyright issues. Also, one thing to remember regarding the indentation of the essay–which is the 34th criterion of EDD—is that consider tab-spacing between two sentences as the indentation for a new paragraph.

This is the end of the instruction. Now, I'm providing each essay.

Table 4: Essay classifications, human and AI scores, and excerpts of ChatGPT feedback

Essay #	Class	Hu Score	AI Score	AI Feedback
2995	11100	6	6	• Essay clearly takes a stance against censorship with two main arguments, but reasoning is repetitive and lacks deeper development or specific supporting details. • Organization includes introduction, body, and conclusion, but transitions are weak, and the structure is somewhat formulaic with limited expansion of ideas. • Frequent grammar and spelling errors (e.g., shout for should, there selves, ofeneded, intersting) and awkward phrasing disrupt fluency and reduce overall polish.
3103	01000	6	6	• The essay presents a clear opinion in favor of censorship, with supporting reasons such as protecting children from adult content and the influence of media. Personal experience is included, which adds some strength to the argument, though examples are somewhat vague and underdeveloped. • Organization is generally clear with an introduction, supporting paragraphs, and conclusion, but transitions are weak and ideas repeat instead of expanding. Some claims lack detail or evidence, which reduces the persuasiveness. • Language control is limited. There are numerous errors in grammar, spelling, and word choice (e.g., 'boundries,' 'consent,' 'hazardess,' 'drasticly,' 'cencorships,' 'portarying'). Sentence fluency is uneven, and phrasing is sometimes awkward. These issues interfere somewhat with clarity.
3213	11101	8	7	• Essay presents a clear stance against censorship with supporting arguments and examples, though some ideas are repetitive and lack deeper exploration or nuanced development. • Organization includes introduction, body, and conclusion, but transitions are uneven and some paragraphs meander, reducing the overall clarity and impact of progression. • Language shows good control overall, but errors in spelling and grammar ('Weather' vs. 'Whether,' 'twards,' subject-verb agreement) slightly disrupt flow. Vocabulary is basic and could be more varied.

Table 5: Parameter estimates for checklist descriptors

Descriptor	π^*	r^*_{CON}	r^*_{ORG}	r^*_{GRM}	r^*_{VOC}	r^*_{MCH}
D01	0.99	1.00	0.98	0.78		
D02	0.99		0.81	0.65		
D03	0.81				**0.25**	
D04	0.99	0.81	0.59			
D05	0.95	0.88	**0.16**			
D06	0.99	0.82	**0.08**			
D07	0.99		**0.03**			
D08	0.95	**0.17**				**0.37**
D09	0.99		**0.46**			
D10	0.95				**0.07**	
D11	0.93	**0.08**				
D12	0.98		**0.25**			
D13	0.99	**0.11**				
D14	0.96				**0.04**	
D15	0.97	0.80	**0.43**	**0.35**		0.84
D16	0.99		1.00	**0.28**		
D17	0.99			**0.39**		
D18	0.88			**0.05**		**0.32**
D19	0.99		0.98	**0.23**		
D20	0.99		0.98	**0.33**		
D21	0.98			**0.31**		
D22	0.99			**0.22**		
D23	0.96			**0.13**		
D24	0.78			1.00		**0.15**
D25	0.99			**0.30**		0.89
D26	0.31				**0.02**	1.00
D27	0.80				**0.09**	
D28	0.99		0.92	0.58		0.98
D29	0.98		0.79	**0.24**		
D30	0.94		0.71	**0.29**		0.87
D31	0.54	**0.46**	1.00			**0.28**
D32	0.99		1.00			**0.34**
D33	0.95			**0.13**		
D34	0.97		0.83	0.56		0.98
D35	0.99		0.93	0.82		0.99

Notes. $\gamma*$ estimates with strong discriminant power (< 0.5) are highlighted in bold.