Predicting Initial Essay Quality Scores to Increase the Efficiency of Comparative Judgment Assessments

Michiel De Vrindt [1ac] **Anaïs Tack** [1ad] **Renske Bouwer** [2b]
Wim Van Den Noortgate [1ac] **Marije Lesterhuis** [3e]

[1] KU Leuven [a] imec research group itec [b] Institute for Language Sciences
[c] Faculty of Psychology and Educational Sciences [2] Utrecht University [3] UMC Utrecht
[d] Faculty of Arts [e] Center for Research and Development of Health Professions Education

Abstract

Comparative judgment (CJ) is a method that can be used to assess the writing quality of student essays based on repeated pairwise comparisons by multiple assessors. Although the assessment method is known to have high validity and reliability, it can be particularly inefficient, as assessors must make many judgments before the scores become reliable. Prior research has investigated methods to improve the efficiency of CJ, yet these methods introduce additional challenges, notably stemming from the initial lack of information at the start of the assessment, which is known as a cold-start problem. This paper reports on a study in which we predict the initial quality scores of essays to establish a warm start for CJ. To achieve this, we construct informative prior distributions for the quality scores based on the predicted initial quality scores. Through simulation studies, we demonstrate that our approach increases the efficiency of CJ: On average, assessors need to make 30% fewer judgments for each essay to reach an overall reliability level of 0.70.

1 Introduction

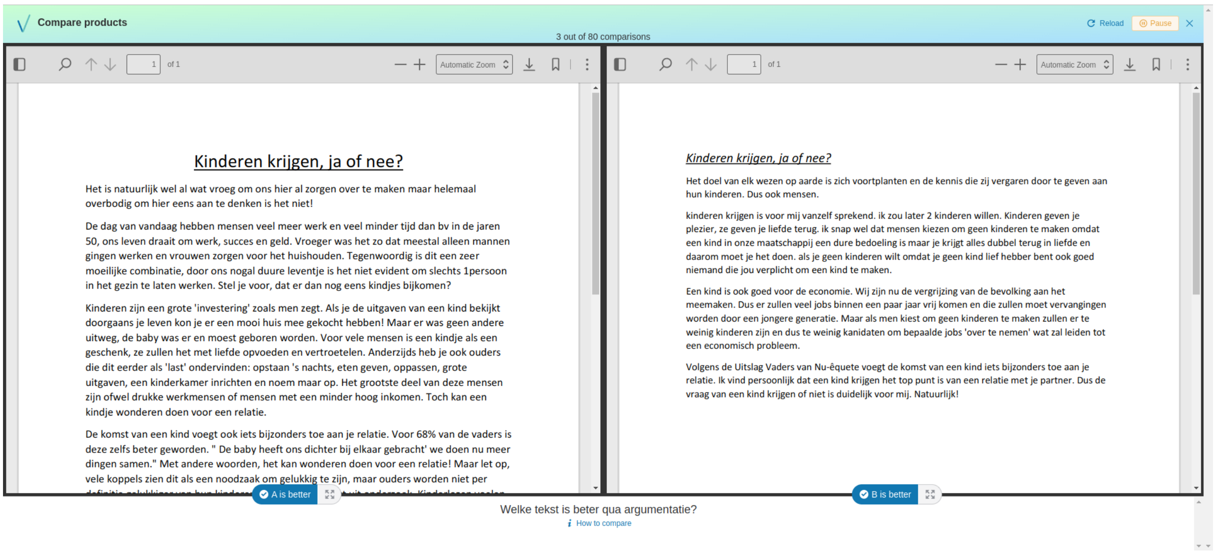
The Comparative Judgment (CJ) method is utilized in diverse educational assessments, and specifically, some educational institutions employ it for the assessment of student essays. As shown in Figure 1, this approach involves presenting two essays in a web-based tool, where assessors compare them to determine the best one. After a sufficient number of judgments, all pairwise comparisons are used to calculate a quality score for each essay. In contrast to rubric marking, CJ provides distinctive advantages. Assessors can apply their expertise and experience flexibly, without strict adherence to rubrics (Bloxham, 2009; Laming, 2003). Additionally, CJ enhances the reliability and validity of scores by incorporating multiple judgments from various assessors (Lesterhuis et al., 2022; Verhavert et al., 2019).

Despite the advantages of CJ, it still requires many judgments from assessors before quality scores become reliable enough, typically requiring between 10 and 14 judgments per essay to achieve a reliability level of 0.70 (Verhavert et al., 2019), rendering the assessment method rather inefficient (McMahon and Jones, 2015). A cause of its inefficiency is that, at the start of the assessment, there is no information about the quality scores, as no judgments have been made yet. In adaptive learning systems, this problem is commonly referred to as cold-start problem (Sun et al., 2022a; Pliakos et al., 2019).

A solution to alleviating this cold-start problem, and subsequently increasing the efficiency of CJ, would be to introduce a 'warm start' in the assessment by automatically predicting initial quality scores for essays. Although the prediction of essay quality has already been extensively explored in automated essay scoring (AES) (see a review by Klebanov and Madnani, 2022), these studies have mostly focused on what could be defined as non-comparative, or absolute (Bouwer et al., 2023), essay scoring, where each essay is scored as a standalone piece without comparison to other essays. To the best of our knowledge, there have been few to no studies that explored the automatic prediction of essay quality scores obtained through CJ assessments.

To address this research gap, we studied the extent to which essay quality scores, resulting from a CJ assessment, can be automatically predicted and used to alleviate the cold start of CJ with the goal of increasing the efficiency of CJ for assessing essay quality. We focused on Dutch essays written for argumentative assignments. Firstly, we conducted a machine learning experiment in which deep learning models were trained on data collected from CJ assessments to predict quality scores of essays. Secondly, we ran simulations where we used the predicted quality scores as initial quality scores to

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 125–136
June 20, 2024 ©2024 Association for Computational Linguistics



Kinderen krijgen, ja of nee?

Het is natuurlijk wel al wat vroeg om ons hier al zorgen over te maken maar helemaal overbodig om hier eens aan te denken is het niet!

De dag van vandaag hebben mensen veel meer werk en veel minder tijd dan bv in de jaren 50, ons leven draait om werk, succes en geld. Vroeger was het zo dat meestal alleen mannen gingen werken en vrouwen zorgen voor het huishouden. Tegenwoordig is dit een zeer moeilijke combinatie, door ons nogal duure leventje is het niet evident om slechts 1persoon in het gezin te laten werken. Stel je voor, dat er dan nog eens kindjes bijkomen?

Kinderen zijn een grote 'investering' zoals men zegt. Als je de uitgaven van een kind bekijkt doorgaans je leven kon je er een mooi huis mee gekocht hebben! Maar er was geen andere uitweg, de baby was er en moest geboren worden. Voor vele mensen is een kindje als een geschenk, ze zullen het met liefde opvoeden en vertroetelen. Anderzijds heb je ook ouders die dit eerder als 'last' ondervinden: opstaan 's nachts, eten geven, oppassen, grote uitgaven, een kinderkamer inrichten en noem maar op. Het grootste deel van deze mensen zijn ofwel drukke werkmensen of mensen met een minder hoog inkomen. Toch kan een kindje wonderen doen voor een relatie.

De komst van een kind voegt ook iets bijzonders toe aan je relatie. Voor 68% van de vaders is deze zelfs beter geworden. " De baby heeft ons dichter bij elkaar gebracht' we doen nu meer dingen samen." Met andere woorden, het kan wonderen doen voor een relatie! Maar let op, vele koppels zien dit als een noodzaak om gelukkig te zijn, maar ouders worden niet per

Kinderen krijgen, ja of nee?

Het doel van elk wezen op aarde is zich voortplanten en de kennis die zij vergaren door te geven aan hun kinderen. Dus ook mensen.

kinderen krijgen is voor mij vanzelf sprekend. ik zou later 2 kinderen willen. Kinderen geven je plezier, ze geven je liefde terug. ik snap wel dat mensen kiezen om geen kinderen te maken omdat een kind in onze maatschappij een dure bedoeling is maar je krijgt alles dubbel terug in liefde en daarom moet je het doen. als je geen kinderen wilt omdat je geen kind lief hebber bent ook goed niemand die jou verplicht om een kind te maken.

Een kind is ook goed voor de economie. Wij zijn nu de vergrijzing van de bevolking aan het meemaken. Dus er zullen veel jobs binnen een paar jaar vrij komen en die zullen moet vervangingen worden door een jongere generatie. Maar als men kiest om geen kinderen te maken zullen er te weinig kinderen zijn en dus te weinig kanidaten om bepaalde jobs 'over te nemen' wat zal leiden tot een economisch probleem.

Volgens de Uitslag Vaders van Nu-êquete voegt de komst van een kind iets bijzonders toe aan je relatie. Ik vind persoonlijk dat een kind krijgen het top punt is van een relatie met je partner. Dus de vraag van een kind krijgen of niet is duidelijk voor mij. Natuurlijk!

Welke tekst is beter qua argumentatie?

Figure 1: Screenshot of the Comproved web application (https://comproved.com), showcasing a comparative judgment assessment. Here, two Dutch essays discussing the topic ‘Having children, yes or no?’ are randomly chosen and presented to an assessor, who determines which essay showed the best argumentation.

alleviate the cold start of CJ. These steps were conducted to answer the following research questions:

1. To what extent can current deep learning models automatically predict essay quality scores that resemble quality scores obtained from CJ assessments?

2. If these predicted scores are used as initial quality scores within CJ, to what extent can we decrease the number of comparative judgments needed to obtain reliable scores?

2 Background

2.1 Comparative Judgment Assessments

Generally, CJ assessments consist of three steps that are repeated. In a first step, a pair of two essays is selected and presented to one of the multiple assessors. In a second step, the assessor is tasked with comparing the two essays and determining which is of a higher quality given the task description of the assignment, that is, the prompt. In a third step, statistical models such as the Bradley-Terry-Luce (BTL) model are used to model the outcomes of all pairwise comparisons on a quality scale (Bradley and Terry, 1952; Luce, 1959).

More formally, BTL model relates $\mathbb{P}(i \succ j)$, that is the probability that essay i is preferred over essay j, to the difference in their estimated quality scores, θ_i and θ_j (see Equation 1), with $i \in \{1, \ldots, n\}$ and $i \neq j$. The smaller the difference, the closer the probability is to 0.50. The outcome of comparing essay i with essay j is denoted by $Z_{ij} \in \{0, 1\}$, where $Z_{ij} = 1$ in case essay i is preferred over essay j, and 0 otherwise. Each quality is a logit value $\theta_i \in \mathbb{R}$ where $\sum_{i=1}^{n} \theta_i = 0$.

$$\mathbb{P}(i \succ j) := \mathbb{P}(Z_{ij} = 1) = \frac{e^{\theta_i - \theta_j}}{1 + e^{\theta_i - \theta_j}} \quad (1)$$

$$Z_{ij} \sim \text{Bernoulli}\left(\mathbb{P}(i \succ j)\right) \quad (2)$$

Different selection rules for CJ (step 1) have been proposed to increase the efficiency of the assessment. These selection rules rely on certain characteristics of essays. Most notably, Pollitt (2012) proposed to select pairs of essays adaptively based on the closest estimated quality scores. The outcomes of these judgments are the most uncertain and, therefore, the most informative for the quality scores in a statistical sense. However, there are two drawbacks to adaptive selection: First, it cannot be used at the start of the assessment, as quality scores are still unknown, and second, during the assessment, adaptive selection can lead to an overly optimistic view of reliability, causing the assessment to end prematurely (Bramley and Vitello, 2019; Crompvoets et al., 2020). Alternatively, pairs of essays can be selected based on the textual information of essays. De Vrindt et al. (2022) proposed to select pairs of essays that are semantically similar during the initial phase of the CJ assessment. They encoded the essay texts as numeric vectors using doc2vec (Le and Mikolov, 2014) and selected the pairs with the highest cosine similarity. However, the efficiency gain they observed was only limited. Therefore, it is of interest to investigate other ways of using textual information of essays to speed up CJ assessments. We focus on the automatic prediction of quality scores based on previously assessed essay texts.

2.2 Automated Essay Scoring

In the field of AES, the automatic prediction of scores has been extensively investigated with as goal to reduce the workload of assessors. This field has experienced significant advances driven by deep learning (Ramesh and Sanampudi, 2022). The proposed deep learning techniques depend on the educational setting in which AES is used. In scenarios where no previously scored essays are available, the prediction relies solely on the essay text itself. This can be achieved, for example, through unsupervised learning (Mim et al., 2019; Wang et al., 2023). In AES research, it is typical to have scored essays on hand. These scored essays help researchers understand the connection between scores and essay content, enabling them to predict essay scores more accurately. This can be achieved through supervised learning (Alikaniotis et al., 2016; Taghipour and Ng, 2016; Yang et al., 2020; Li et al., 2022). For supervised learning, essays that have been scored in the training set can be written for a different assignment than the essays in the test set for which scores are predicted. In such a setting, the prompt for the assignments is often considered to predict the essay scores in addition to the essay texts (Li et al., 2020; Do et al., 2023; Liu et al., 2019).

2.3 Cold-start Problem in Psychometry

The cold-start problem is most commonly termed in the context of recommender systems to denote the difficulty of proposing items to users when the preferences of the users or the characteristics of the items are unknown due to limited user interactions. Using language models, this issue has been addressed by extracting characteristics from item texts (Penha and Hauff, 2020) or by generating user preferences based on the textual description of user historical preferences and items (Wang et al., 2024).

Similarly, in computerized adaptive testing, the cold-start problem persists. These systems select test items so that the difficulty of the item matches the test takers ability, but when responses for items are lacking, inferring item difficulty becomes challenging. Therefore, to calibrate the characteristics of the items, responses for the items need to be collected during a pilot phase. To mitigate the need for extensive piloting, Settles et al. (2020) extracted the linguistic features of test items measuring their difficulty. Alternatively, McCarthy et al. (2021) used pre-trained embeddings of test items to estimate their difficulty and discriminatory power.

The cold-start problem for CJ is similar: quality scores for essays are unknown at the start of an assessment because assessors have not judged them, requiring assessors to make many judgments during the assessment. Analogously to recommender systems and computerized adaptive testing, we address the cold start of CJ by inferring the unknown measures, namely the quality scores, from essay texts.

3 Method

3.1 Data

This study was based on data gathered in a previous study by Lesterhuis et al. (2022). The dataset, described in Table 1, comprised three assignments in which students around the age of 16 wrote argumentative essays in Dutch. The topics for these essays were: (1) having children, (2) organ donation, and (3) stress experienced by students. Students were provided with a prompt detailing the essay topic, the task requirements, and the source texts they were required to integrate in the essay.

Assignment	**Essays**	**Tokens**	**Tokens/Essay**
	N		$M \pm SD$
1. Children	135	42,349	316 (± 93)
2. Organ	136	40,990	304 (± 90)
3. Stress	35	11,286	322 (± 103)

Table 1: Overview of the argumentative writing tasks gathered by Lesterhuis et al. (2022). Tokenization was performed using the Dutch tokenizer from spaCy (Explosion, 2023), which splits the essay texts into meaningful segments.

The essays were assessed by secondary education assessors using a comparative judgment method. Assessors were presented with pairs of randomly selected essays and had to decide which one was better in terms of argumentation, as illustrated in Figure 1. The number of assessors for each assignment and the total of judgments per essay are detailed in Table 2.

Assignment	Judgments/Essay	Assessors
1. Children	18	55
2. Organ	13	52
3. Stress	27	42

Table 2: Overview of the number of comparative judgments made per argumentative writing assignment

To study the predictability of initial essay quality scores and their role in a warm start, it is of course imperative to have quality scores for each essay. For each of the three assignments separately, essay quality scores were derived from the parameters of a Bayesian BTL model with a cold-start condition. These model parameters were estimated based on all comparative judgments within the same assignment. Since these parameters reflect the quality scores estimated at the end of the CJ assessment, we will refer to them as the 'final quality scores' throughout the remainder of this paper. Additional details regarding this cold-start model will be provided in Section 3.5. The distributions of the quality score for each essay within each assignment are shown in Figure 2. Given the large number of comparative judgments per essay (Verhavert et al., 2019) and the diverse panel of assessors responsible for these judgments (van Daal et al., 2016), we can confidently affirm the reliability and validity of these estimated scores.

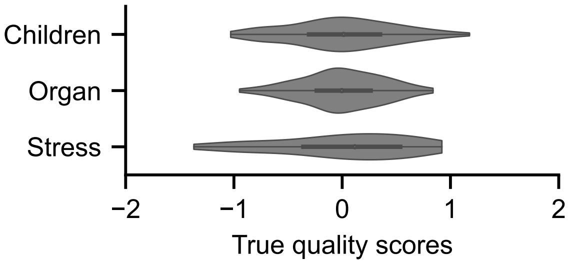


Figure 2: Distributions of final quality scores estimated from a Bayesian BTL model with a cold-start condition

3.2 Models

For predicting essay quality scores, we employed various pre-trained language models and fine-tuned them based on the final quality scores. While alternative feature-based and classical NLP methods exist for this purpose, we focused on fine-tuning transformer models due to their demonstrated superiority in AES research (Uto et al., 2020; Ormerod et al., 2021). We specifically avoided multilingual models, concentrating solely on Dutch models, as prior studies indicate that monolingual models tend to outperform on tasks involving Dutch texts (de Vries et al., 2019; Delobelle et al., 2020). We used three different pre-trained Dutch language models, namely **BERTje** (base, uncased) (de Vries et al., 2019), **RobBERT** (v2) (Delobelle et al., 2022), and **RobBERTje** (non-shuffled) (Delobelle et al., 2021). BERTje is built upon the BERT architecture trained on 12GB of Dutch texts containing 2.4B tokens. RobBERT on the other hand, is based on the RoBERTa architecture, which boosts BERT's efficacy by pre-training in batches on 36GB of Dutch texts containing 6.6B tokens. RoBERTje employs a DistilBERT architecture, derived from RobBERTje, while preserving comparable efficacy with fewer parameters by using knowledge distillation.

We conducted a machine learning experiment with two model configurations: (a) fine-tuning the model solely on the provided essay text as input, and (b) fine-tuning the model on both the essay text and the given prompt as input. The models were imported with the Hugging Face library with a Pytorch backend and implemented to perform a regression task.

More details on the specific computing infrastructure can be found in Appendix A. For the final regression layer, we employed a sigmoid activation function as a way of bounding the scalar values to enhance the training stability. These bounded values functioned as predicted quality scores. Consistent with common practice in essay scoring (Alikaniotis et al., 2016; Yang et al., 2020; Li et al., 2022), all quality scores were min-max normalized before training. These normalized scores, along with the predicted scores, were used to compute the mean squared error, which functioned as the training loss. After training, the predicted scores were reverted to the original scale.

In the second configuration, the assignment prompt was taken into account in addition to the essay texts for the prediction of quality scores. We hypothesized that prompt information would be important for the prediction of quality scores, as the essays in the training set and the test set were written for different assignments. To incorporate this information into the model, we encoded the prompt using the same transformer model as for the essay text (i.e., a shared encoder). Two additional cross-attention layers were added to model the relationship between essays and prompts in both directions. This is similar to the configuration proposed by Liu et al. (2019).

The hyperparameters are given in Appendix B. These were selected based on preliminary results on a held-out set, comprising 15% of essays randomly selected from the training set, which were omitted during training but used for model evaluation.

3.3 Experimental Setup

To evaluate the reliability of the quality scores predicted by the fine-tuned models, we ran a machine learning experiment with the following training and test splits: $\{1, 2\} \to 3$, $\{1, 3\} \to 2$, and $\{2, 3\} \to 1$. In each fold, the three pre-trained models were fine-tuned on essays coming from two assignments (e.g., 1 and 2) and were evaluated on essays coming from the remaining assignment (e.g., 3). We employed this setup to emulate a real-world assessment scenario where we would have an assignment for which we do not have any scores yet (e.g., 3) and for which we need to predict initial quality scores based on scores estimated for other assignments (e.g., 1 and 2).

It is crucial to note that, despite the scores being logit values derived from distinct assignments, there was no complication in joining them within the training set. This was possible because the assignments were very similar, each assessing the quality of argumentative writing.

3.4 Evaluation Metric

Because our objective was to establish the reliability of predicted quality scores, we utilized the **squared Pearson correlation** (Bi, 2003)

$$\rho^2_{\theta^{init},\theta^*} = \frac{\mathrm{Var}_{\theta^*}}{\mathrm{Var}_{\theta^{init}}} \tag{3}$$

to assess the reliability between the predicted initial quality scores θ_i^{init} and final quality scores θ_i^* for $i = \{1, \ldots, n\}$ the essays in the test set. The reliability can be interpreted as the proportion of variance of the predicted initial quality scores that is attributed to the final quality scores. The closer this ratio is to one, the higher the reliability.

3.5 Efficiency Simulation Study

After having fine-tuning and evaluated pre-trained models, we simulated the impact of integrating model predictions as initial quality scores in CJ assessments. For each train-test split, we selected the model and its configuration (i.e., essay text with or without prompt) that exhibited the highest reliability. Subsequently, we conducted simulations to compare CJ assessments under **two conditions**: a warm-start BTL model (our experimental condition, where initial quality scores were predicted using the best model) and a cold-start BTL model (our control condition, where initial quality scores were absent).

While likelihood-based techniques (Hunter, 2004) are typically employed for parameter estimation in the BTL model (Equation 1), we adopted a **Bayesian approach** to simulate CJ assessments with both cold-start and warm-start BTL models. Within this framework, we could establish prior assumptions about the distribution of quality scores. Bayes' theorem allowed us to integrate these priors with judgments in the BTL model, resulting in posterior distributions for all quality scores. Compared to maximum likelihood estimation, Bayesian inference provides more stable estimates and a clearer understanding of the associated uncertainty (Phelan and Whelan, 2017).

3.5.1 Cold-Start Bayesian BTL Model

Under the cold-start condition, we formulated for each quality score a normal prior distribution (Equation 4) having a mean of 0 for all quality scores.

$$\theta_i \sim \text{Normal}\left(0, \sigma_i^2\right) \tag{4}$$

This prior serves to regularize the distribution of quality scores, rendering it weakly informative. The lack of specificity about the essays for which quality scores are estimated characterizes this Bayesian BTL model as having a 'cold start'.

For the variance of each quality score, we specified a normal-truncated prior distribution (Equation 5), which is a common choice for $\sigma_i^2 \in (0, \infty)$.

$$\sigma_i^2 \sim \text{Normal}_{Trunc}\left(\mu_0, \sigma_0^2\right) \tag{5}$$

The parameters of the distribution of σ_i^2 determined the level of uncertainty of the prior quality scores: the larger the location and scale parameters, the greater the prior uncertainty of the quality scores. Based on preliminary results, we chose to fix these parameters for all quality scores: $\mu_0 = 0.5$ and $\sigma_0^2 = 0.1$.

3.5.2 Warm-Start Bayesian BTL Model

Under the warm-start condition, we formulated prior distributions for the quality scores using the predicted quality scores. These priors are deemed informative, as they incorporate information about each essay's quality score.

To construct informative priors, we assumed a normal prior distribution for all quality scores θ_i for $i = \{1, \ldots, n\}$ with as mean their predicted initial quality scores θ_i^{init}.

$$\theta_i \sim \text{Normal}\left(\theta_i^{init}, \sigma_i^2\right) \tag{6}$$

All predicted quality scores were first centered, $\theta_i^{init} - \sum_{i=1}^{n} \theta_i^{init}$, to speed up convergence and encourage $\sum_{i=1}^{n} \theta_i \approx 0$. As in the cold-start condition, prior distributions were specified for the variance of the quality scores, measuring the uncertainty of the estimates (see Equation 5).

3.5.3 Sampling and Simulations

To estimate the posterior distribution of each θ_i and σ_i^2, samples were drawn according to the Hamiltonian Monte Carlo algorithm using Stan (Gelman et al., 2015), with 4 chains of 2000 steps of which 500 were warm-up steps. These were sufficient to reach convergence as diagnosed by a r-hat value of 1 (Vehtari et al., 2021). After convergence, the averages of the posterior distributions were used as point estimates.

To simulate a CJ assessment, we repeatedly estimated θ_i and σ_i^2 using increasingly more judgments; for an example of a simulated CJ assessment, see Appendix C. To account for possible effects of the order of judgments, we shuffled the sequence of judgments twenty times, resulting in twenty simulations of a CJ assessment. We repeated this process for each assessment, employing both a cold and a warm start.

3.5.4 Measuring Efficiency Gain

We assessed the gain in efficiency when introducing a warm start by observing the decrease in the average number of judgments required per essay to achieve a specific reliability level. The reliability of the quality scores was determined by the squared Pearson correlation (ρ^2_{θ,θ^*}) between the final quality scores θ^*, estimated at the end of the assessment, and the quality scores in a Bayesian BTL model estimated at a certain point during the assessment θ.

However, the use of this reliability metric presents a practical challenge. In practice, the reliability cannot be calculated during an assessment, as the final quality scores that would be estimated at the end of the assessment are still unknown. Hence, the reliability has to be approximated based on the estimated quality scores, which can be achieved using the Scale Separation Reliability (SSR$_\theta$). More specifically, the SSR$_\theta$ estimates Var_{θ^*} in Equation 3 by $\text{Var}_\theta - \mathbb{E}_{\sigma^2}$; see Equation 7. For a detailed derivation of the SSR$_\theta$, please refer to Verhavert et al. (2018). Note that we adjusted the reliability of the estimated quality scores to account for the reliability level of the final quality scores; see Appendix D.

$$\text{SSR}_\theta = \frac{\text{Var}_\theta - \mathbb{E}_{\sigma^2}}{\text{Var}_\theta} \to \rho^2_{\theta,\theta^*} \tag{7}$$

4 Results

4.1 Machine Learning Experiment

Table 3 shows the results of the machine learning experiment. The findings indicate that all fine-tuned language models effectively predicted quality scores for a completely new assignment, with correlation coefficients significantly different from zero. Notably, RobBERT consistently exhibited the highest reliability in predicting quality scores, aligning with its superior performance over other Dutch transformer models in diverse tasks (Delobelle et al., 2022).

Furthermore, when integrating both essay and prompt information, the RobBERT model consistently achieved the highest reliability with true quality scores. This observation aligns with previous AES research, emphasizing the predictive accuracy of essay scores across various prompts (Li et al., 2020; Do et al., 2023). As a result of these findings, we opted for the RobBERT model incorporating additional prompt information to predict initial quality scores in the simulation study.

It is crucial to note, however, that despite achieving high reliability, the fact that the reliability levels did not surpass 0.70 underscores the importance of assessor judgments to further improve the reliability of essay quality scores.

4.2 Simulation of CJ Assessments

The simulation study results, shown in Figure 3, highlight the comparison between CJ assessments under warm-start and cold-start conditions. The outcomes indicate that adopting a warm-start approach proved more efficient in terms of the number of judgments per essay needed to achieve a reliability level of at least 0.70.

In both Assignment 1 (Figure 3.c) and Assignment 3 (Figure 3.a), the desired reliability was

Fold	Essay Texts			+ Prompt Information		
	BERTje	**RobBERT**	**RobBERTje**	**BERTje**	**RobBERT**	**RobBERTje**
$\{1,2\} \to 3$	0.56	0.61	0.54	0.60	**0.63**	0.52
$\{1,3\} \to 2$	0.51	0.55	0.43	0.50	**0.59**	0.45
$\{2,3\} \to 1$	0.43	0.56	0.16	0.42	**0.57**	0.17
Average	0.50	0.57	0.38	0.52	**0.59**	0.37

Table 3: Squared Pearson correlations computed on the test set, comparing final quality scores and scores predicted by fine-tuned models, utilizing either only the essay texts or the prompts as well. Maximum scores are boldfaced.

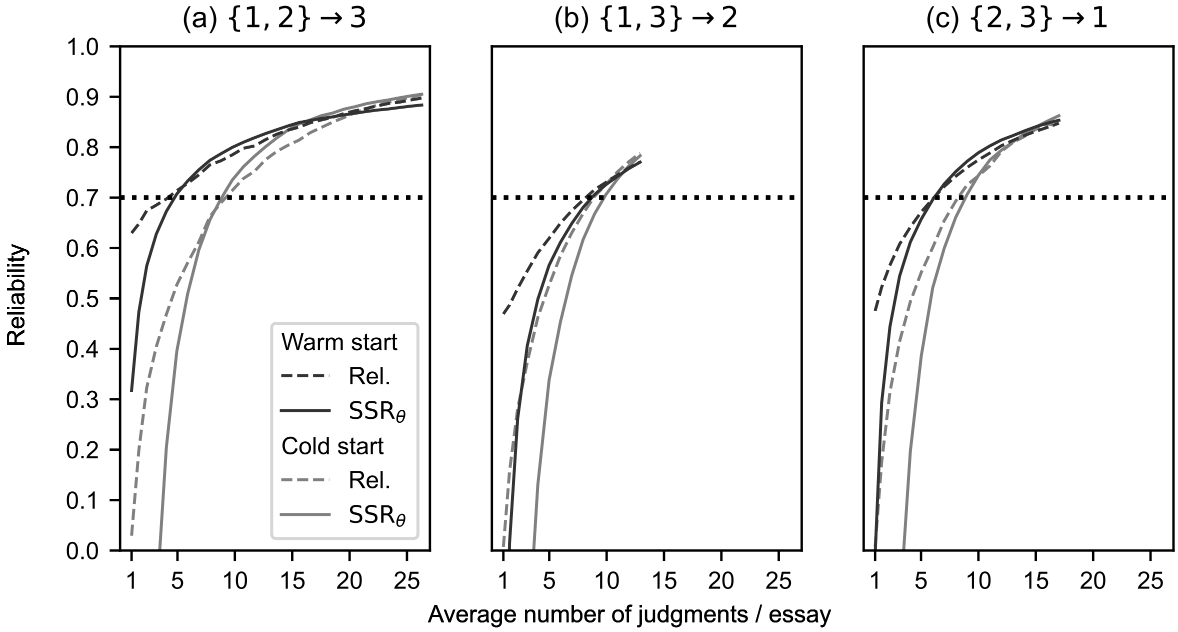


Figure 3: Results of simulated CJ assessments with a warm and a cold start. The average reliability and the average SSR$_\theta$ of the estimated quality scores are given in function of the average number of comparisons made per essay. These scores are averaged over 20 different orders of comparative judgments used to simulate an assessment.

reached with fewer than six judgments per essay. Conversely, employing a cold-start method required more than nine judgments per essay to attain an equivalent reliability level. Consequently, the warm-start approach resulted in efficiency gains of 35% and 41%, respectively. For Assignment 2 (Figure 3.b), a reliability of 0.70 required less than nine judgments per essay, while with a cold start at least ten judgments per essay were needed, which corresponds to an efficiency gain of 15%.

When exceeding ten judgments per essay, the disparity in reliability between warm and cold starts decreased across all assignments. This can be attributed to the diminishing impact of prior distributions on posterior distributions as the number of judgments increases. Additionally, for assignments 2 and 3, the reliability with a warm start begins to slightly trail behind that of the cold-start condition after ten judgments per essay. We posit that this observed difference may be associated with the choice to estimate final quality scores using a Bayesian BTL model with a cold start.

In practical scenarios, reliability is not accessible during assessments, making accurate measurement with the SSR$_\theta$ crucial. As shown in Figure 3, the SSR$_\theta$ demonstrated a faster approximation of reliability when employing a warm start compared to a cold start. Specifically, the SSR$_\theta$ reached the 0.70 reliability level for all assignments under a warm start. In contrast, the SSR$_\theta$ approached reliability at levels of 0.75 for Assignment 2 and 0.80 for Assignments 1 and 3 under a cold start.

To examine the impact of warm-starting assessments on *individual* quality scores, we compared the progression of quality score rankings. For illustration purposes, we show the results of one simulated assessment for Assignment 3. Figure 4 demonstrates that adopting a warm start led to qual-

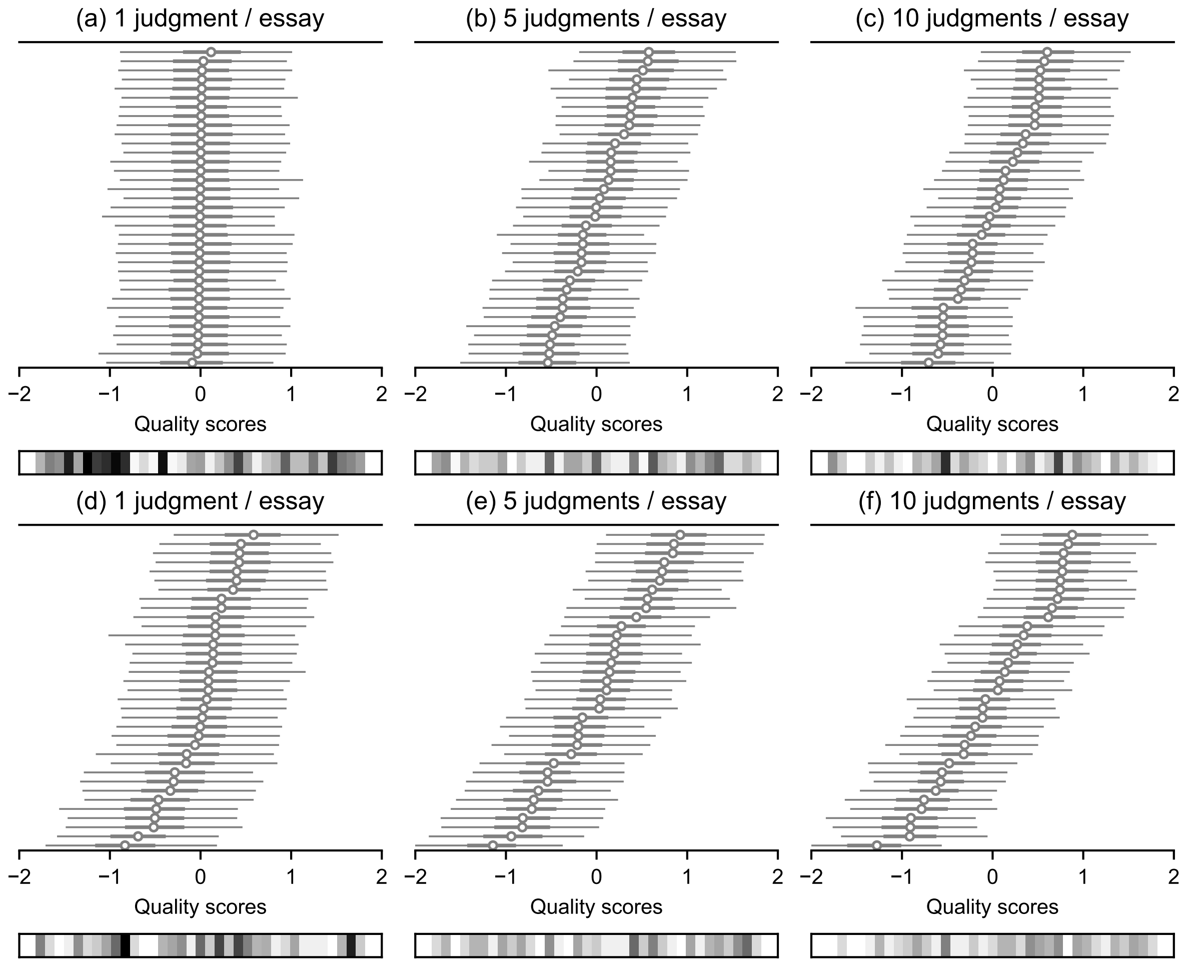


Figure 4: Forest plots of quality scores with 94%-high density intervals estimated at different stages of the CJ assessment of Assignment 3, with a cold-start condition (plots a–c above) and a warm-start condition (plots d–f below). The bar plots at the bottom show ranking accuracy based on the absolute differences in rank order of estimated and final quality scores, with darker shades indicating more incorrect rankings of estimated quality scores.

ity scores being more spread out, yielding a fairly accurate ranking at the start of the assessment. In contrast, quality scores under the cold-start condition clustered around the mean value, resulting in less precise rankings. This highlights the efficacy of informative priors in the warm-start condition in discerning between quality scores. Even after ten judgments per essay, the warm-start approach displayed a wider range of quality scores and a better ranking compared to the cold-start method.

5 Discussion

Our findings underscore the ability of current deep learning models, particularly transformer models, to predict initial quality scores that provide valuable information on the argumentative writing quality of essays. Furthermore, incorporating the assignment prompts for fine-tuning enhances the reliability of predicted quality scores, which aligns with prior research in AES (Li et al., 2020; Do et al., 2023; Sun et al., 2022b). We posit that prompt information is especially important for the prediction of initial quality score, since, in this study, the essays in the training set were written for different assignments than the essays in the test set.

When warm-starting CJ assessments with these predicted initial quality scores, the necessary number of comparative judgments to obtain reliable quality scores decreases significantly. This suggests that less effort from assessors is required while upholding high levels of reliability of the quality scores. Furthermore, our approach to increase the efficiency of CJ avoids any undesirable effects with respect to the reliability measures, which have been noted when employing an adaptive selection rule (Bramley, 2015; Bramley and

Vitello, 2019; Crompvoets et al., 2020). Additionally, our method demonstrates a more substantial improvement in efficiency compared to the approach of De Vrindt et al. (2022), who devised a more efficient selection rule based on similarities in essay texts.

6 Conclusion

We successfully improved the efficiency of CJ assessments by introducing a warm start for the estimation of the quality scores. This involved predicting essay quality scores, which were then used to form informative prior distributions within a Bayesian BTL model. Through an extensive simulation study, we demonstrated that our approach led to a reduction, ranging between 15% and 41%, in the number of comparative judgments needed to reach a reliability of 0.70 and produced more accurate rankings of essays at the start of an assessment. Furthermore, our findings indicate that these efficiency gains can be measured in practical settings, as the SSR$_\theta$ approximates the reliability well.

7 Limitations

To fine-tune the transformer models for the prediction of quality scores, we devised a training set combining the quality scores from different CJ assessments. This was feasible, as the quality scores measured the same quality of argumentative writing. However, if the essays were written in different text genres, such as informative writing, combining the quality scores would become non-trivial, since they measure a different kind of writing quality. Therefore, we recommend that before combining quality scores, they are first calibrated on a fixed scale using, for example, the method of Fair Averages (Linacre, 1989). Furthermore, differences in the genre of essays in train and test could make predicting the initial scores more difficult, causing lower reliability.

In this study, we assumed that the quality scores of essays written for other assignments were available to train a deep learning model for score prediction. However, settings may arise where these quality scores are unavailable, particularly in educational contexts where privacy concerns may prevent the inclusion of students' essays in a training set. In such cases, alternative methods for predicting scores must be explored. One approach is to train a deep learning model on publicly available AES datasets, such as the Automated Student Assessment Prize (ASAP) dataset published by the Hewlett Foundation (Hamner et al., 2012). However, it should be noted that these essays are written in English, prompting the need to evaluate how well a model trained on these can predict scores for Dutch essays. Alternatively, in case no essay scores are available for training, unsupervised learning approaches for AES could be considered (Ridley et al., 2020; Zhang and Litman, 2021).

To simulate the CJ assessments, we chose to repeatedly shuffle the order of judgments (see Appendix C). However, this approach may not reflect a realistic CJ assessment process, as, typically, pairs of essays for judgment are selected in such a way that each essay is compared (close to) the same number of times. For example, if an essay is compared 9 times and the others 10, that essay is selected and paired with a randomly selected essay. Based on preliminary results, we observed that our choice to repeatedly shuffle judgments has a negligible impact on the reliability results, as outlined in this study.

The current study reports an increase in reliability at the start of the assessment, but after more judgments have been made, the difference in reliability between a cold and a warm start became minimal (see Figure 3). For future research, we recommend exploring methods that use essay texts for the selection of pairs in a way that increases the reliability toward the end of an assessment, while avoiding the perverse effects that adaptive selection rules introduce (Bramley, 2015; Bramley and Vitello, 2019; van Daal et al., 2017).

Acknowledgements

This research was supported by a Baekeland mandate (HBC.2022.0164) granted by the Flanders Innovation & Entrepreneurship (VLAIO) to the first author, in cooperation with the company Comproved.

References

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

Jian Bi. 2003. Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies*, 18(1):61–76.

Sue Bloxham. 2009. Marking and moderation in the uk: false assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, 34(2):209–220.

Renske Bouwer, Elke Van Steendam, and Marije Lesterhuis. 2023. Guidelines for the validation of writing assessment in intervention studies. In Fien De Smedt, Renske Bouwer, Teresa Limpo, and Steve Graham, editors, *Conceptualizing, Designing, Implementing, and Evaluating Writing Interventions*, volume 40, pages 199–223. Brill.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom Bramley. 2015. Investigating the reliability of adaptive comparative judgment. Technical report, University of Cambridge.

Tom Bramley and Sylvia Vitello. 2019. The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(1):43–58.

Elise Anne Victoire Crompvoets, Anton A Béguin, and Klaas Sijtsma. 2020. Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, 45(3):316–338.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. *arXiv preprint arXiv:1912.09582*.

Michiel De Vrindt, Wim Van den Noortgate, and Dries Debeer. 2022. Text mining to alleviate the cold-start problem of adaptive comparative judgments. *Frontiers in Education*, 7:132–147.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2021. Robbertje: A distilled dutch bert model. *Computational Linguistics in the Netherlands Journal*, 11:125–140.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2022. Robbert-2022: Updating a dutch language model to account for evolving language use. *arXiv preprint arXiv:2211.08192*.

Heejin Do, Yunsu Kim, and Gary Geunbae Lee. 2023. Prompt- and trait relation-aware cross-prompt essay trait scoring. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1538–1551, Toronto, Canada. Association for Computational Linguistics.

Explosion. 2023. Available trained pipelines for dutch: nl_core_news_sm. `https://spacy.io/models/nl#nl_core_news_sm` [Accessed on April 3, 2024)].

Andrew Gelman, Daniel Lee, and Jiqiang Guo. 2015. Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, 40(5):530–543.

Ben Hamner, Jaison Morgan, Mark Shermis lynnvandev, and Tom Vander Ark. 2012. The hewlett foundation: Automated essay scoring.

David R Hunter. 2004. MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1):384–406.

Beata Beigman Klebanov and Nitin Madnani. 2022. *Automated Essay Scoring*. Number 52 in Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael.

Donald Laming. 2003. *Human judgment: The eye of the beholder*. Cengage Learning, London, United Kingdom.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*.

Marije Lesterhuis, Renske Bouwer, Tine van Daal, Vincent Donche, and Sven De Maeyer. 2022. Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education*, 7:122–131.

Xia Li, Minping Chen, and Jian-Yun Nie. 2020. Sednn: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems*, 210:106491.

Xia Li, Huali Yang, Shengze Hu, Jing Geng, Keke Lin, and Yuhai Li. 2022. Enhanced hybrid neural network for automated essay scoring. *Expert Systems*, 39(10):e13068.

John Michael Linacre. 1989. *Many-faceted Rasch measurement*. Ph.D. thesis, The University of Chicago.

Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. 2019. Automatic short answer grading via multiway attention networks. *arXiv preprint arXiv:1909.10166*.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*.

R. Duncan Luce. 1959. On the possible psychophysical laws. *Psychological Review*, 66(2):81–95.

Arya D. McCarthy, Kevin P. Yancey, Geoffrey T. LaFlair, Jesse Egbert, Manqian Liao, and Burr Settles. 2021. Jump-starting item parameters for adaptive language tests. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 883–899, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Suzanne McMahon and Ian Jones. 2015. A comparative judgement approach to teacher assessment. *Assessment in Education: Principles, Policy & Practice*, 22(3):368–389.

Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. 2019. Unsupervised learning of discourse-aware text representation for essay scoring. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 378–385, Florence, Italy. Association for Computational Linguistics.

Christopher M Ormerod, Akanksha Malhotra, and Amir Jafari. 2021. Automated essay scoring using efficient transformer-based language models. *arXiv preprint arXiv:2102.13136*.

Gustavo Penha and Claudia Hauff. 2020. What does bert know about books, movies and music? probing bert for conversational recommendation. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 388–397, New York, NY, USA. Association for Computing Machinery.

Gabriel C Phelan and John T Whelan. 2017. Hierarchical bayesian bradley-terry for applications in major league baseball. *arXiv preprint arXiv:1712.05879*.

Konstantinos Pliakos, Seang-Hwane Joo, Jung Yeon Park, Frederik Cornillie, Celine Vens, and Wim Van den Noortgate. 2019. Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers & Education*, 137:91–103.

Alastair Pollitt. 2012. The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 19(3):281–300.

Dadi Ramesh and Suresh Kumar Sanampudi. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3):2495–2527.

Robert Ridley, Liang He, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2020. Prompt agnostic essay scorer: A domain generalization approach to cross-prompt automated essay scoring. *arXiv preprint arXiv:2008.01441*.

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263.

Geng Sun, Wei Wei, Tingru Cui, Dongming Xu, Shiping Chen, Alex Shvonski, Li Li, Jun Shen, and Soheila Garshasbi. 2022a. Adapting new learners and new resources to micro open learning via online computation. *IEEE Transactions on Computational Social Systems*, 9(6):1807–1819.

Jingbo Sun, Tianbao Song, Jihua Song, and Weiming Peng. 2022b. Improving automated essay scoring by prompt prediction and matching. *Entropy*, 24(9):1–15.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Masaki Uto, Yikuan Xie, and Maomi Ueno. 2020. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6077–6088, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tine van Daal, Marije Lesterhuis, Liesje Coertjens, Vincent Donche, and Sven De Maeyer. 2016. Validity of comparative judgement to assess academic writing: examining implications of its holistic character and building on a shared consensus. *Assessment in Education Principles Policy and Practice*, 26:59–74.

Tine van Daal, Marije Lesterhuis, Liesje Coertjens, Marie-Thérèse van de Kamp, Vincent Donche, and Sven De Maeyer. 2017. The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. *Frontiers in Education*, 2:1–13.

Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. 2021. Rank-normalization, folding, and localization: An improved r^ for assessing convergence of mcmc (with discussion). *Bayesian Analysis*, 16(2).

San Verhavert, Renske Bouwer, Vincent Donche, and Sven De Maeyer. 2019. A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, 26(5):541–562.

San Verhavert, Sven De Maeyer, Vincent Donche, and Liesje Coertjens. 2018. Scale separation reliability: what does it mean in the context of comparative judgment? *Applied Psychological Measurement*, 42(6):428–445.

Cong Wang, Zhiwei Jiang, Yafeng Yin, Zifeng Cheng, Shiping Ge, and Qing Gu. 2023. Aggregating multiple heuristic signals as supervision for unsupervised automated essay scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13999–14013, Toronto, Canada. Association for Computational Linguistics.

Jianling Wang, Haokai Lu, James Caverlee, Ed Chi, and Minmin Chen. 2024. Large language models as data augmenters for cold-start item recommendation. *arXiv preprint arXiv:2402.11724*.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Haoran Zhang and Diane Litman. 2021. Essay Quality Signals as Weak Supervision for Source-based Essay Scoring. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 85–96, Online. Association for Computational Linguistics.

A Computing Information

We implemented both transformer models for quality score prediction using Pytorch 2.1.0, Hugging Face 4.32.1, and Python 3.9.12. We conducted the experiments on a system running Ubuntu 22.04.2.

B Hyperparameters

The AdamW optimizer was used (Loshchilov and Hutter, 2017), with a polynomial learning rate scheduler and a starting learning rate of $1e-5$. The warm-up ratio was set at 10% of the steps, with a batch size of 5. The weight decay was set to 0.09. Furthermore, a 5% dropout was used to prevent overfitting. The transformer models were fine-tuned for 40 epochs with the possibility of early stopping based on the evaluation metric measured on the held-out set.

C Example of Simulated CJ Assessment

For the CJ assessment of Assignment 3, 27 judgments were made for each essay, as detailed in Table 2. This means that each essay was involved in 27 pairwise comparisons. Given that there are 35 essays part of the assessment, assessors had to make $35 \times 27/2 \approx 473$ judgments in total. To simulate the CJ assessment of Assignment 3, all θ_i and σ_i^2 parameters in a Bayesian BTL model were iteratively estimated using 1 to 473 judgments. Following each estimation, the SSR$_\theta$ and reliability were computed. Recognizing that the order of judgments selected could influence the estimates and reliability levels, we shuffled the sequence of judgments twenty times and repeated the procedure mentioned above.

D Adjusting the Reliability Measure

In studies on the reliability of CJ, the ‘true quality scores’ are obtaining using a all-play-all design (Bramley, 2015; Crompvoets et al., 2020), where every pairwise combination essays has been judged. Since the data in this study were not gathered using an all-play-all, we assume that the final quality scores are, in fact, the true scores. However, these final quality scores possess their own level of reliability, as given by the SSR of the estimated quality scores at the end of a CJ assessment: SSR$_{\theta*}$. To account for this, we adjusted the reliability of the estimated quality scores, $\rho^2_{\theta,\theta*}$, by multiplying it by SSR$_{\theta*}$. Consequently, SSR$_\theta$ converges to SSR$_{\theta*}$, when the estimated quality scores align with the final quality scores at the end of the assessment (i.e., when $\rho^2_{\theta,\theta*} \approx 1$).