Computers and Education: Artificial Intelligence 8 (2025) 100414

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence

# Opportunities of natural language processing for comparative judgment assessment of essays

Michiel De Vrindt [a,b,*], Anaïs Tack [b,c], Wim Van den Noortgate [a,b], Marije Lesterhuis [d], Renske Bouwer [e]

[a] Faculty of Psychology and Educational Sciences, KU Leuven, Etienne Sabbelaan 53, 8500, Kortrijk, Belgium
[b] itec, an imec research group at KU Leuven, Etienne Sabbelaan 51, 8500, Kortrijk, Belgium
[c] Faculty of Arts, KU Leuven, Etienne Sabbelaan 53, 8500, Kortrijk, Belgium
[d] Center for Research and Development of Health Professions Education, UMC Utrecht, Etienne Sabbelaan 53, 8500, Utrecht, the Netherlands
[e] Institute for Language Sciences, Utrecht University, Trans 10, 3512 JK, Utrecht, the Netherlands

ARTICLE INFO

*Keywords:*
Comparative judgment
Natural language processing
Hybrid human-AI
Automated essay scoring
Partial-automation

ABSTRACT

Comparative judgment (CJ) is an assessment method commonly used for assessing essay quality, where assessors compare pairs of essays and judge which essays are superior in quality. A psychometric model is used to convert judgments into quality scores. Although CJ yields reliable and valid scores, its widespread implementation in educational practice is hindered by its inefficiency and limited feedback capabilities. This conceptual study explores how Natural Language Processing (NLP) can address these limitations, drawing upon existing NLP techniques and the very limited research on their integration within CJ. More specifically, we argue that, at the start of the assessment, initial essay quality scores could be predicted from essay texts using NLP, mitigating the cold-start problem of CJ. During the CJ assessment, selection rules could be constructed using NLP to efficiently increase the reliability of the scores while supporting assessors by not letting them make too difficult comparisons. After the CJ assessment, NLP could automate feedback, helping to better understand how assessors arrived at their judgments and explaining the scores to assessees (students). To support future research, we overview appropriate methods based on existing research and highlight important considerations for each opportunity. Ultimately, we contend that integrating NLP into CJ can significantly improve the efficiency and transparency of the assessment method, all while preserving the crucial role of human assessors in evaluating writing quality.

## 1. Introduction

Comparative judgment (CJ) (Thurstone, 1927a, 1927b), also known as paired comparison or comparative assessment, is an assessment method used to assess and rank a set of objects based on their relative traits or qualities. In the field of educational measurement, the method has often been applied to reliably assess the writing quality of student-written essays by comparing them in pairs (van Daal et al., 2016; Steedle & Ferrara, 2016; Baniya et al., 2019). The assessment of machine-generated essays using CJ has not been studied to date. Beyond measuring the quality of essays, CJ has also been applied to a variety of other educational assessments, including conceptual understanding (Jones et al., 2019), problem-solving skills in mathematics (Jones & Inglis, 2015), geography (Pollitt & Whitehouse, 2012), design portfolios (Newhouse, 2014), formative assessments (Potter et al., 2017; Bartholomew et al., 2019), and comparison of assessment standards between examination boards (Bramley, 2007; D'Arcy, 1997).

The CJ assessment process proceeds as follows. First, a pair of two essays is selected and presented to an assessor. Then, the assessor is tasked with comparing the two essays and determining which has the higher quality. Based on their judgment, the quality scores of all essays are estimated. This is an iterative process which involves a group of assessors, each judging a series of (different) pairs of essays. Psychometric models such as the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959) are used to relate the quality scores of essays to the assessors' judgments. More specifically, the BTL model links the probability of one essay being preferred over another in a pair to the difference in their quality scores. The higher the quality score of the first essay with respect to that of the second essay in a pair, the higher the probability that the first essay wins that comparison. In practice, the simplest

* Corresponding author at: Faculty of Psychology and Educational Sciences, KU Leuven, Etienne Sabbelaan 53, 8500, Kortrijk, Belgium.
*E-mail address:* michiel.devrindt@kuleuven.be (M. De Vrindt).

https://doi.org/10.1016/j.caeai.2025.100414
Received 22 November 2024; Received in revised form 31 March 2025; Accepted 5 May 2025
Available online 13 May 2025

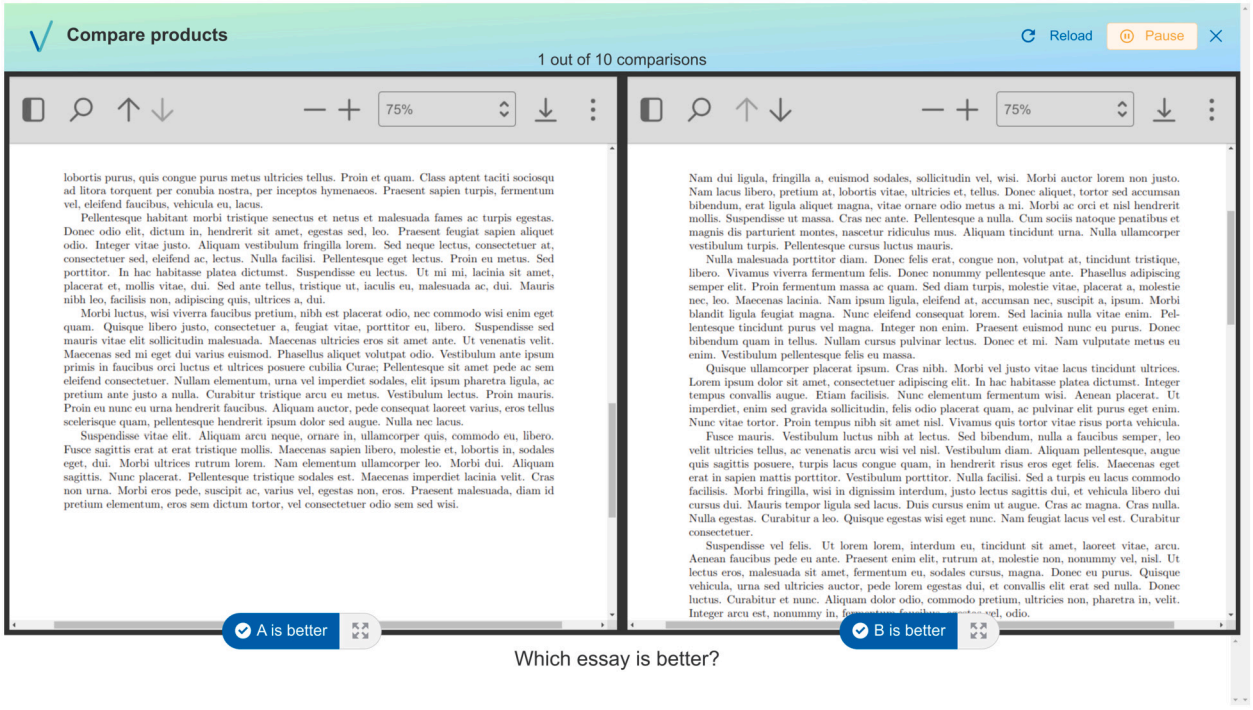


**Fig. 1.** Screenshot of the Comproved web application (https://comproved.com) used for conducting CJ assessments online. The assessor is repeatedly presented with a pair of essays and asked to select the one with superior writing quality.

specification of the BTL model is often adopted, where ties are not allowed and dependencies between judgments are disregarded (Bramley, 2007). The CJ assessment concludes once sufficient comparisons have been made, typically when each essay has been judged enough to yield reliable quality scores. Practically, this occurs after each essay has been involved in multiple comparisons, often around 10 to 14, resulting in quality scores of acceptable reliability of 0.7 (Verhavert et al., 2019). An example of a web platform where assessors have to compare pairs of essays is shown in Fig. 1.

CJ offers several advantages over rubric-based grading, where essays are evaluated according to specific criteria or dimensions relevant to the assessment's learning objectives. Firstly, CJ assessments are generally less tedious for assessors, as they are not required to evaluate how well essays adhere to detailed rubrics (Bloxham, 2009). Instead, assessors can use their expertise and experience when comparing essays (Lesterhuis et al., 2022), which is a more natural approach to assessment (Laming, 2003). Secondly, CJ assessments yield valid scores that reflect a consensus among assessors on what the assessed quality comprises (Bouwer et al., 2023; Wheadon et al., 2019; Jones & Inglis, 2015), even when assessors have varying conceptualizations of the assessed quality (van Daal et al., 2016). Finally, CJ assessments tend to produce reliable scores because they aggregate numerous judgments made by multiple assessors (Heldsinger & Humphry, 2010; Verhavert et al., 2019; Bouwer et al., 2023).

Despite the advantages of CJ in terms of validity and reliability, the assessment method still faces practical challenges at different stages, impeding the efficiency and transparency of the assessment. These practical challenges, along with goals, are depicted in Fig. 2.

At the start of a CJ assessment, the quality scores of essays are still unknown because no judgments have been made yet. As the quality scores are initially unknown, assessors have to make numerous judgments throughout the assessment before the quality scores become reliable, making CJ a rather inefficient assessment method. This issue, stemming from the lack of initial information on the quality scores of essays, is referred to as a 'cold-start' problem (De Vrindt et al., 2022, 2024).

During a CJ assessment, pairs of essays are selected according to a selection rule. However, existing selection rules are either inefficient or cause unintended effects. A random selection rule is inefficient (Verhavert et al., 2019), as it can select essays that are not very informative to compare, such as a high-quality and a low-quality essay. As a result, assessors have to make many judgments before the quality scores are reliable. Bramley et al. (1998) denoted the inefficiency of CJ as the most salient difficulty of the assessment method in practice. To address this inefficiency issue, Pollitt (2012b) proposed to adaptively select pairs of essays based on minimal differences in the current estimates of the quality scores, as these are statistically the most informative to judge. However, adaptive selection rules tend to systematically overestimate the reliability of the estimated quality scores (Bramley, 2015; Bramley & Vitello, 2019; Crompvoets et al., 2020), leading to a biased view of reliability. Furthermore, pairing essays of similar quality makes it more difficult to judge them (Gijsen et al., 2021; van Daal et al., 2017), requiring more time from assessors and thus decreasing the efficiency of the assessment.

After the CJ assessment is completed, there is a lack of feedback opportunities for both assessors and assessees. This issue arises because the scores are derived from holistic judgments (Steedle & Ferrara, 2016; Kelly et al., 2022), which lacks the transparency found when using detailed rubric marking (Jonsson, 2014; Mortier et al., 2015). The lack of transparency in quality scores also complicates the evaluation of the validity of the judgments made by assessors (Kelly et al., 2022). Although assessors could provide feedback comments when making judgments, doing so extensively would be time-consuming and reduce assessment efficiency. Furthermore, writing numerous comments to individual essays can lead assessors to adopt a more analytical approach (Verhavert et al., 2019), which conflicts with the holistic nature of CJ assessments (van Daal et al., 2016).

To tackle the challenges of CJ while maintaining its validity and reliability, we argue that natural language processing (NLP) could be used at different stages of the CJ assessment. NLP involves various computational techniques designed to automatically analyze human languages, both written and spoken (Chowdhary, 2020). A key objective in language analysis is the development of techniques to extract and model

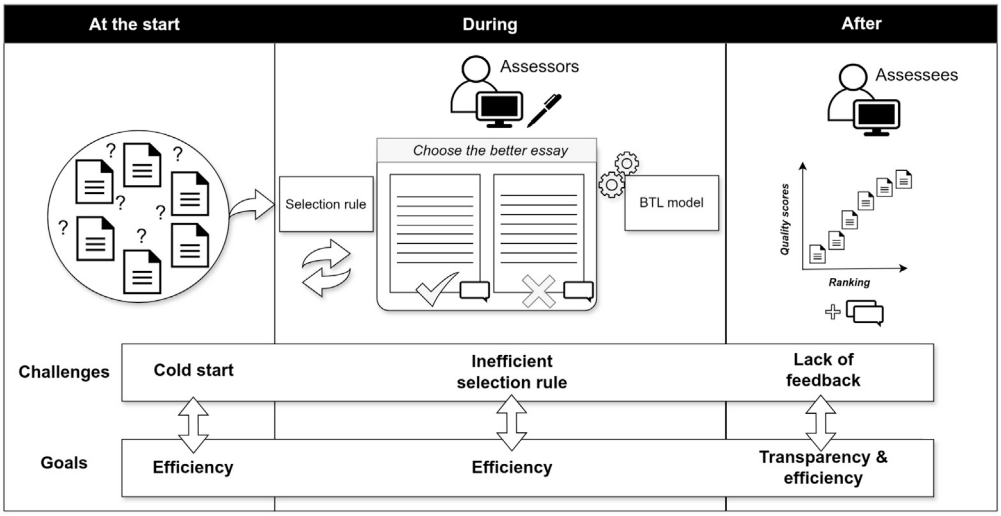


**Fig. 2.** Diagram depicted CJ assessment as it is currently commonly conducted with challenges faced at different stages of the assessment along with assessment goals.

the syntax and semantics of language. Because NLP enables the extraction and modeling of textual information from essay texts, it could be leveraged to enhance the efficiency and transparency of CJ assessments when assessing essays. Although NLP has demonstrated success in automated essay scoring (AES) research (Taghipour & Ng, 2016; Dong et al., 2017; Xue et al., 2021; Wang et al., 2022), its integration into CJ assessments has received little attention. Therefore, in this conceptual study, we explore the potential of NLP to address key practical challenges of CJ assessments based on existing NLP methods and the little empirical research available on their integration for CJ assessments. More specifically, this study identifies key opportunities for integrating NLP at the start, during, and after the CJ assessment process and outlines crucial considerations for future research. Additionally, we discuss the nature of automation introduced by integrating NLP into CJ, focusing on the dynamic between assessors and the'AI system' (i.e., the automation introduced with NLP).

## 2. NLP at the start of the assessment

At the start of the assessment process, there is a lack of information about the quality of the essays, as assessors are yet to make judgments. Hence, many judgments are required before the estimates of quality scores are reliable. De Vrindt et al. (2022) found this problem to be similar to the 'cold-start' problem encountered in adaptive learning systems (Sun et al., 2022a; Pliakos et al., 2019), where the characteristics of test items are unknown and need to be calibrated accordingly, as well as in recommender systems (Schein et al., 2002), where historical user interactions with items are needed for the calibration of item characteristics. To mitigate the need for extensive calibration of adaptive learning systems and recommender systems, linguistic features of test items can be automatically extracted with NLP and used to infer unknown characteristics of items (Settles et al., 2020; McCarthy et al., 2021; Penha & Hauff, 2020).

Similar to mitigating the cold start in adaptive learning systems and recommender systems, we argue that the cold start of CJ can be mitigated by predicting the unknown quality scores from essay texts using NLP. More specifically, these predicted quality scores could function as initial quality scores of essays at the start of the assessment. Then, during the assessment, the initial quality scores could be further refined based on the judgments made by assessors in the BTL model (see Section 3). The specific NLP methods and the reliability of these initial scores are contingent on the available data to predict the initial quality scores. Hence, for predicting initial scores, we can distinguish between two educational settings: Setting A, where no assessed essays are available, and Setting B, where assessed essays from previous assessments are available. Both settings and their opportunities are illustrated in Fig. 3.

### 2.1. Alleviating the cold start based on inherent linguistic features of essays

When no prior assessments are available, quality scores could be predicted directly from the essay texts themselves. This is what we denoted as Setting A. In AES research, weakly supervised learning techniques are commonly used to predict essay scores when no assessed essays are available (Zhang & Litman, 2021; Wang et al., 2023; Mim et al., 2019; Song et al., 2020). These techniques rely on inferring weak signals from data and then using them for supervised learning. To infer the quality of essays from their texts, weak signals of essay quality could be obtained by measuring the linguistic features of essay texts, such as the number of grammatical errors, usage of words and pronouns, style, organization, and relevance to the assignment prompt (Ke & Ng, 2019; Zesch et al., 2015). Essays scoring high or low on these features are likely to be of high or low quality. Based on these weak signals for essay quality, a machine learning model could be trained to predict essay scores as a regression task. Neural networks are commonly applied for supervised learning on these weak signals (Zhang & Litman, 2021; Mim et al., 2019). However, Wang et al. (2023) proposed an alternative approach by formulating score prediction as a ranking problem. In this method, the ranking of essays is first predicted for various linguistic features, which are then aggregated to determine final essay scores. A recent study by Mizumoto and Eguchi (2023) used a large language model to predict essay scores through zero-shot prompting. The model was instructed to generate scores based on a description of the rubrics. They then combined the model's output with a set of linguistic features related to complexity and cohesion to train a regression model to predict the essay scores.

Following a weakly-supervised learning approach for AES, initial quality scores could be predicted at the start of a CJ assessment without requiring scores of assessed essays. However, note that machine learning models that only rely on essay texts generally produce poorer score predictions compared to settings where human assessments are available for training (Wang et al., 2023), highlighting the importance of using human assessments for training automated scoring systems.

### 2.2. Alleviating the cold start based on available assessed essays

In a more practical assessment setting, previously assessed essays that were part of other assessments could be available. This is what we denoted as Setting B. These assessments could be essays written for the

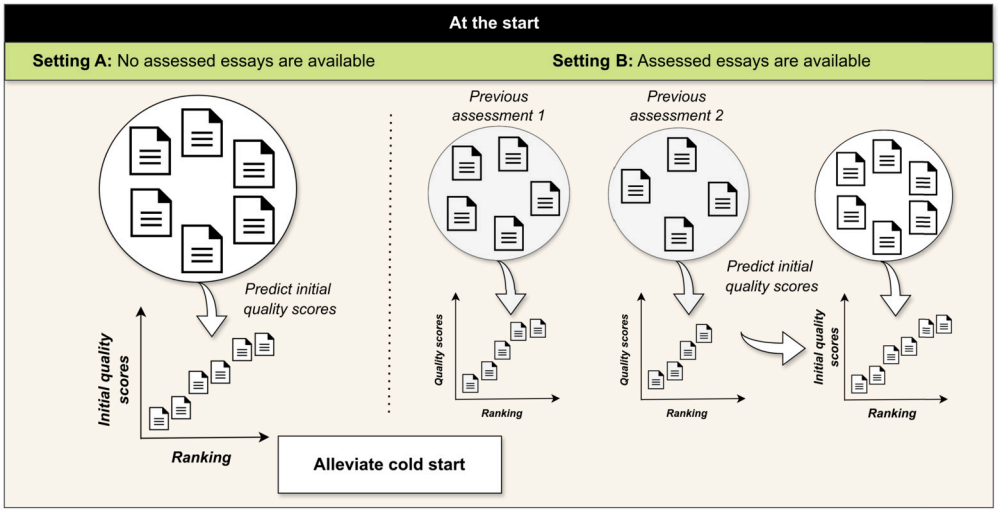


**Fig. 3.** Diagram depicting the opportunities of NLP to predict initial quality scores from essay texts at the start of the CJ assessment, thus alleviating its cold start. Opportunities are outlined for two settings: Setting A, where no assessed essays are available, in which case initial quality scores can be predicted based on inherent linguistic features of the essays, and Setting B, where previously assessed essays from other assessments are available, allowing initial quality scores to be predicted based on the quality scores of those assessed essays.

same assignment but at a different time, for instance, when an exam is retaken or when essays are written for a different assignment. In this setting, quality scores do not need to be inferred from the inherent linguistic features of essays; instead, they could be predicted based on the relationship between available essay texts and their quality scores. This setting is most commonly assumed for AES (Ramesh & Sanampudi, 2022).

Traditionally, machine learning models are used that rely on linguistic features of essay texts to predict scores (Phandi et al., 2015; Zesch et al., 2015). However, in recent years, there has been a shift toward adopting deep-learning models for AES (Ramesh & Sanampudi, 2022; Uto, 2021). Rather than using specific linguistic features of essays, deep-learning models typically use transfer learning, which maps essay texts to numerical representations by leveraging knowledge acquired from pre-training on large volumes of corpora. The numerical representations, often termed as 'embeddings' (Pilehvar & Camacho-Collados, 2020), can be further fine-tuned for various tasks, such as predicting essay scores (Uto, 2021). Fine-tuned deep-learning models, particularly transformers (Vaswani et al., 2017), predict essay scores that show a high agreement with the scores given by human assessors by adding a regression layer to the final layer of the model (Taghipour & Ng, 2016; Dong et al., 2017; Xue et al., 2021; Wang et al., 2022). Nonetheless, deep-learning models come with certain drawbacks (Yaneva & von Davier, 2023); for example, training or fine-tuning deep-learning models require significant computational resources and deep-learning models operate as black-box systems, making their predictions not directly interpretable, while feature-based essay-scoring models are interpretable. For an overview of deep-learning architectures for AES, see Ramesh and Sanampudi (2022).

In a recent empirical study, De Vrindt et al. (2024) showed the potential of mitigating the cold start of CJ by predicting initial quality scores based on quality scores of assessed essays written for other assignments. They addressed the cold start of CJ by predicting initial quality scores and using them as prior belief in a Bayesian BTL model. First, they fine-tuned a transformer model with a regression layer on quality scores of essays from available CJ assessments to predict initial quality scores of essays at the start of a new assessment. Then, they used the predicted initial quality scores to construct informative prior distributions in a Bayesian BTL model, which are continuously updated during an assessment based on assessors' judgments. Through a simulation study, De Vrindt et al. (2024) showed that their approach improved the efficiency, as the number of judgments needed to reach a reliability of 0.70 dropped from 15% to 41%. Even though the scope of the study was limited, they showed promising results to alleviate the cold-start CJ for Setting B.

### 2.3. Considerations

For future research on predicting initial quality scores to alleviate the cold start of CJ, it is important to recognize that CJ uses different evaluation metrics compared to AES. Typically in AES research, the agreement between predicted scores and 'true scores' of essays, which are the scores given by (multiple) assessors, is evaluated with the quadratic weighted Kappa (Cohen, 1960; Ke & Ng, 2019). The goal is to maximize the overlap between the predicted and true scores as discrete values. Instead of focusing on agreement, in CJ assessments, the reliability of the quality scores is evaluated to indicate the consistency among assessors' judgments (Wheadon et al., 2019). The higher the reliability, the fewer judgments assessors need to make. Following classical test theory, reliability can be measured using the squared Pearson correlation between the estimated quality scores and the true quality scores, representing the proportion of variance of quality qualities explained by the true quality scores (Kim, 2012; Brennan, 2010). In CJ, the true scores are assumed to be the scores that accurately reflect the quality of essays. Although impracticable, these true quality scores could be obtained by letting assessors compare all essays with each other, using an all-play-all design (Bramley, 2015; Bramley & Vitello, 2019; Crompvoets et al., 2020).

A potential challenge in this prediction task arises when the assignment of the essays used for training differs from the assignment of the essays for which quality scores are predicted. To account for such differences, assignment information, such as prompt or source material, could be added as contextual information to essay texts when training essay-scoring models (Do et al., 2023; Sun et al., 2022b; Li et al., 2020). However, if the essays were written in different text genres, combining the quality scores in a training dataset could be inappropriate since they measure different kinds of writing quality and represent different scales. To accommodate for this, the scores of each assessment could be first calibrated on a scale with a fixed range using, for example, the method of Fair Averages (Linacre, 1989). Another challenge may arise when longer-form essays are involved since many transform-based models have token length limitations. To train on longer-form essays, they could be split into smaller chunks, sometimes with overlapping sections, to retain the context of the entire essay (Dong et al., 2023). This chunking strategy ensures that models can effectively capture global and local

features of longer texts without losing critical information for predicting essay scores. Considering this contextual information of the assessment, such as the assignment, scale of scores, and type of essays, is essential to ensure that predicted initial scores are as reliable and valid as possible.

## 3. NLP during the CJ assessment

During the CJ assessment process, assessors repeatedly judge pairs of essays and choose which essay is better. Subsequently, the quality scores are estimated based on the judgments of assessors (see Fig. 2). These quality scores are logit values in a BTL model and are estimated based on the number of times essays won comparisons and the essays with which they were compared. Generally, the reliability of the estimated quality scores increases as the number of comparisons increases (Verhavert et al., 2019).

As mentioned above, the reliability of the estimated quality scores is calculated relative to the true quality scores (Kim, 2012; Verhavert et al., 2018). However, in practice, the true quality scores of essays are unknown during an assessment. Therefore, the reliability cannot be directly calculated. Instead, in practical assessments, the scale separation reliability (SSR) is calculated to approximate the reliability of the estimated quality scores (Brennan, 2010; Bi, 2003). The SSR only depends on the spread and uncertainty of the estimated quality scores. For a derivation and interpretations of the SSR as a reliability measure, refer to Verhavert et al. (2018).

The algorithm or rule that determines which pairs of essays are selected affects how fast the SSR increases. The faster the SSR increases, the fewer judgments the assessors need to make and the more efficient the assessment becomes. Different selection rules exhibit different levels of efficiency. When pairs of essays are randomly selected, assessors are required to make many judgments: between 10 and 14 judgments per essay for an SSR of 0.70 and between 26 and 37 judgments per essay for an SSR of 0.90 (Verhavert et al., 2019). In contrast, when using adaptive selection based on the closest estimated quality scores, the number of judgments that assessors need to make can be reduced to 11 judgments per essay for an SSR of 0.9 (Pollitt, 2012b). However, adaptive selection has two major limitations during the CJ assessment: firstly, it leads to selecting pairs that are difficult for assessors to compare (Gijsen et al., 2021; van Daal et al., 2017), and secondly, it causes the SSR to artificially inflate with respect to the actual reliability as given by the squared Pearson correlation coefficient (Bramley, 2015; Bramley & Vitello, 2019; Crompvoets et al., 2020).

When using NLP on essay texts, a selection rule is not limited solely to using the estimated quality scores. There is the opportunity to leverage information from essay texts in a selection rule for CJ. We contend that by integrating the essay texts, pairs of essays can be selected more efficiently in two respects (see Fig. 4): first, by selecting the set of pairs of essays that are not overly time-consuming to judge, as they are not too difficult to compare, and second, by selecting pairs of essays from this set to minimize the number of judgments needed to obtain reliable quality scores without artificially inflating the SSR.

### 3.1. Avoiding the selection of pairs difficult to judge

Not all pairs of essays pose the same difficulty for assessors to judge. Specifically, judging pairs of essays with a similar quality score, selected with an adaptive selection rule, is difficult for assessors (van Daal et al., 2017). In an eye-tracking study by Gijsen et al. (2021), they found that assessors perceive it to be more difficult to judge pairs of essays with small differences in qualities, leading to more inconsistent judgments. Hence, the more difficult it is to judge pairs of essays, the more time it takes for assessors, and thus, the more inefficient the CJ assessment becomes. Besides time inefficiency, the increased difficulty also leads to more inconsistency in the judgments assessors make, which has a negative impact on the reliability of the quality scores.

To avoid the selection of pairs that are difficult to judge, heuristic rules have been proposed that prevent the selection of pairs of essays with the closest estimated quality scores. For instance, Rangel-Smith and Lynch (2018) selected pairs of essays with quality scores using the BTL model of at least a difference between 1.50 and 2.50, depending on the variance of the quality scores. Similarly, Pollitt (2012a) only selected pairs of essays where the probability of winning, as given by the BTL model, is close to 0.33 or 0.67. Although smaller differences in estimated quality scores coincide with more difficult judgments (van Daal et al., 2017), this may not always be the case. As van Daal (2020) argued, similarity in quality only partially explains the difficulty of judging. The relationship between the similarity in quality scores and judging difficulty could be more nuanced. A pair of essays could very well be of different quality levels but still be difficult to compare, such as when comparing a well-structured essay with many spelling mistakes with a poorly structured essay with few spelling mistakes. However, the impact of the content or features of essays in a pair on the difficulty of judging them has not been explored. By using NLP, pairs of essays that are difficult to judge could be identified based on their texts and subsequently be avoided in a selection rule.

More specifically, the similarity between essay texts could be used to devise a selection rule that avoids difficult judgments. Then, a set of pairs of essays that are not difficult to judge could be selected, as shown in Fig. 4. In the context of AES, scores from assessors need to be collected for training machine learning models. To facilitate easier scoring for assessors, a common approach is to cluster essays based on their similarity so that assessors can assess a batch of essays simultaneously and identify common errors (Basu et al., 2013; Brooks et al., 2014; Weegar & Idestam-Almquist, 2023). Nevertheless, techniques following this approach were mainly developed to quickly grade short-form essays, which are generally easier to compare than regular or long-form essays (van Daal, 2020). As essays can contain a lot of information, aligning them and differentiating between them can be difficult for assessors when comparing them pairwise. To avoid difficult judgments, a selection rule could help focus the attention of assessors. When using NLP, essay pairs could be selected that are similar with respect to unimportant aspects of text quality, such as language conventions and usage (Lesterhuis et al., 2022). By matching unimportant aspects of text quality of essays, assessors could focus more on differences in critical higher-order aspects, such as organization, argumentation, or source integration, which are more decisive when assessing them with CJ (Lesterhuis et al., 2022; Lesterhuis, 2018). This approach could help assessors to better structurally align essays when comparing them, which could reduce their cognitive load (Posten & Mussweiler, 2017; Medin et al., 1995), making comparing them less difficult. For instance, essays of widely different lengths would not be selected to be compared as they are difficult to structurally align with each other. This approach of aligning essays based on unimportant features is likely most beneficial for regular or long-form essays, as they contain more information, which makes comparing them easier. However, even for short-form essays, this could also result in less difficult judgments for assessors, though the impact could be smaller as they are already easier to align.

In AES research using feature-based models, sets of linguistic features are often employed to measure aspects of text quality, such as readability (Zesch et al., 2015), argument strength (Ghosh et al., 2016), prompt relevance (Beigman Klebanov et al., 2016), discourse features (Somasundaran et al., 2014), and language complexity (Latifi & Gierl, 2021). Using the features, similarity metrics could be calculated to measure whether essay pairs are dissimilar with respect to higher-order aspects, which are most important when judging, and more similar with respect to lower-order aspects, which are less unimportant. However, the distinction between important and unimportant aspects of text quality is not always clear and could depend on the assignment or the quality of the essays that are compared. For instance, assessors take into account lower-order aspects more so when comparing lower-quality essays than when comparing higher-quality essays (Humphry & Heldsinger, 2019).

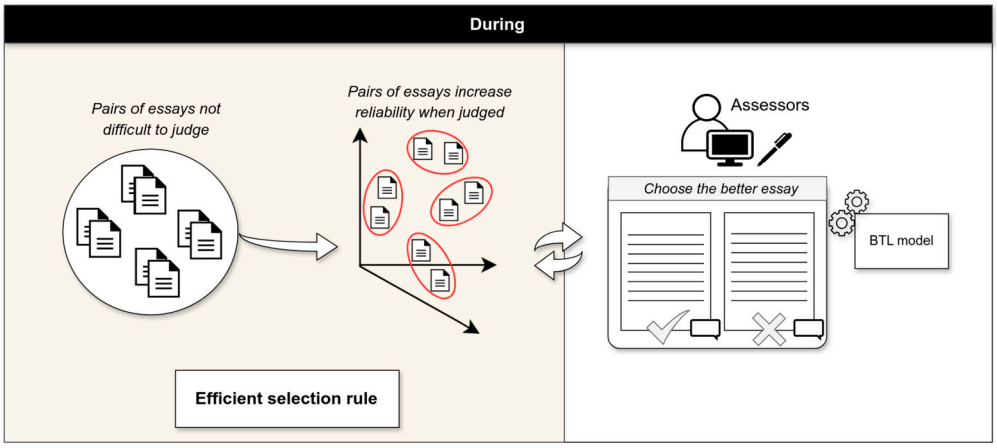


**Fig. 4.** Diagram depicting the opportunities of NLP during the CJ assessment to select essay pairs based on essay texts, aiming to enhance the efficiency of the assessment. To achieve this, first, a subset of essay pairs that are not difficult to judge could be selected. Then, from the subset of essay pairs, pairs could be selected that increase the reliability when judged without inflating the SSR.

However, in general, higher-order features are most important when judging (Lesterhuis et al., 2022).

### 3.2. Increasing the reliability of CJ without inflating the SSR

Besides the difficulty of judging essays, the number of judgments assessors have to make determines the efficiency of a CJ assessment. CJ assessments typically require numerous judgments before the estimated quality scores are reliable (Verhavert et al., 2019). Proposals to reduce the number of judgments with adaptive selection are often ineffective, as they artificially increase reliability when it is measured using the SSR (Bramley, 2015; Bramley & Vitello, 2019; Crompvoets et al., 2020), and increase the difficulty of judging (van Daal et al., 2017), as mentioned previously. A cause for the inflation of the SSR with adaptive selection is its sensitivity to inconsistent judgments, especially at the start of the assessment (Bramley, 2015; Bramley & Vitello, 2019). When assessors make inconsistent judgments at the start of the assessment, the essays are incorrectly placed on the scale. When pairs are selected adaptively based on the most similar estimated quality scores, the judgment of these pairs seems informative as the SSR increases. However, in reality, the true quality scores of the essays are not as similar as their estimates indicate, and the actual reliability does not increase as much as the SSR. As a result, when using adaptive selection, SSR gives an overly optimistic view of the reliability of the estimated quality scores.

Several solutions have been proposed to mitigate the inflation of the SSR caused by adaptive selection. For instance, Bramley (2015) proposed first selecting a smaller subset of essays as a calibration set, where all possible comparisons of essays are judged. Subsequently, adaptive selection is used to pair the remaining essays with those in the calibration set. Verhavert et al. (2022) tested this approach empirically, affirming that this approach does not inflate the SSR. Nevertheless, we contend that such an adaptive selection rule is often inefficient since constructing a reliable calibration set requires many judgments (McGrane et al., 2018). However, this approach could be efficient when a set of assessed essays for the same assignment are readily available, and new essays have to be placed on the same scale (Verhavert et al., 2022), for instance, when an exam has to be retaken later. Moreover, Rangel-Smith and Lynch (2018) proposed an adaptive selection rule that pairs essays with estimated quality scores that differ by at least 1.50. They observed that introducing this constraint for adaptive selection helps mitigate the inflation of the SSR. Nevertheless, we argue that this selection rule is not highly efficient as it prevents the selection of pairs that are very informative to compare, given that pairs of essays with similar estimated quality scores are never selected.

De Vrindt et al. (2022) incorporated information from essay texts into a selection rule using NLP. They selected initial pairs of essays with the most similar texts based on the cosine similarity of their doc2vec embeddings (Le & Mikolov, 2014). Subsequent pairs were selected adaptively based on the estimated quality scores. Although they were able to prevent the SSR from inflating, the efficiency gain with respect to random selection was limited. As of the time of writing this article, NLP can be used to prevent inflation of the SSR, but improving the assessment's efficiency remains a challenge.

There is the opportunity to use the essay texts to devise a more efficient selection rule with NLP. More specifically, the information in essay texts can be used in a selection rule for the exploration of the exploitation of the quality scores of essays. The strategies of exploration and exploitation can be seen as a trade-off that needs to be made to optimize an objective efficiently, in this case, maximizing the reliability efficiently through the selection of pairs of essays. This trade-off is central to decision-making processes (Berger-Tal et al., 2014) and active learning (Mosqueira-Rey et al., 2023). In the context of CJ, we discuss strategies to use essay text for exploration and exploitation in a selection rule to increase the reliability of the quality scores in as few judgments as possible.

A selection rule for CJ could exploit the uncertainty in predicted initial quality scores (see Section 2). A common approach used in AES to minimize the number of scores assessors need to give to train a scoring model involves selecting essays for which the model's highest uncertainty, as these are the most informative to score (Horbach & Palmer, 2016; Dronen et al., 2015; Hastings et al., 2018). Hastings et al. (2018) achieved comparable accuracy by training the model on only 30% of the original training data. In the same way, for CJ, a selection rule could be constructed to select pairs of essays with the most similar predicted initial quality scores (refer to Section 2), as these are the most uncertain. The predicted initial quality scores could be more reliable than the estimated quality scores, particularly in the early stages of the assessment when few or no judgments have been made yet.

Besides using predicted initial quality scores, a selection rule could also use the representations of essays, either through their linguistic features (see Section 2.1) or their (fine-tuned) embeddings (see Section 2.2). The embeddings of essays could be especially informative, as the geometric relationships between embeddings represent semantic and syntactic relationships (Liu et al., 2017). For instance, pairs of essays could be selected with the most similar representations of essay texts, as they are similar with respect to relevant aspects of quality.

To explore the quality of essays based on their texts, selection rules that ensure diversity in selecting pairs of essays could be constructed. Firoozi et al. (2023) and Dronen et al. (2015) selected essays that were dissimilar from previously assessed ones to train essay-scoring models, as similar essays provide less information for the model. Similarly, for

CJ, a selection rule could select pairs of essays that are dissimilar from previously assessed pairs, as judging these helps explore the quality of essays. This exploration strategy could be combined with an exploitation strategy by selecting essays with the closest embeddings while being dissimilar from the embeddings of previously assessed essay pairs. However, as mentioned previously, if the essays in a pair are too similar, judging them may be difficult for assessors. Therefore, it could be preferable to first select essays that are not too difficult to judge before selecting pairs of essays that increase the reliability efficiently (see Fig. 4).

### 3.3. Considerations

To construct an efficient selection rule for CJ, it has to avoid selecting pairs that are difficult for assessors to judge. Additionally, when these pairs are compared, the reliability has to increase without inflating the SSR. The importance of both factors within a selection rule could depend on various assessment conditions, such as the spread of quality scores or the level of expertise of assessors (Verhavert et al., 2019). Neglecting these assessment conditions could render the selection rules ineffective in practice. For example, if assessors are experts, avoiding difficult pairs to judge could be trivial since assessors are more able to judge difficult pairs correctly (Gijsen et al., 2021). Alternatively, if assessors are novices, avoiding the selection of difficult pairs could be very important. Furthermore, when the spread of the quality scores of the essays is small, it is generally more difficult for assessors to discriminate between the essays according to their qualities (van Daal et al., 2017). On the contrary, when the spread in the quality scores of essays is large, discriminating between essays is generally easier for assessors. Before starting the assessment, such assessment conditions could be estimated by experts, such as lead assessors or instructors, given that they have knowledge of the difficulty of the assessment and the expertise of assessors.

## 4. NLP after the CJ assessment

When the assessors have completed a sufficient number of judgments, and the estimated quality scores are reliable, the CJ assessment typically concludes. After the assessment, feedback can be provided to the assessees and assessors. Assesses receive feedback comments written by assessors when making judgments, as well as the rank order of the other assessed essays (Coenen et al., 2018). Assessors, on the other hand, receive feedback in the form of assessment statistics, such as misfit information for individual essays or assessors, average response duration, and SSR of the final quality scores, i.e., the quality scores estimated at the end of the assessment. This feedback serves to help assessors evaluate the reliability and validity of the conducted assessment. Van Gasse et al. (2017) found that these types of feedback for CJ are generally accepted by assessors and assessees, which is the most important precursor for the feedback being effective (Anseel & Lievens, 2009).

Despite the acceptance of existing feedback for CJ, they have several practical limitations. Firstly, writing feedback comments when making judgments is time-consuming for assessors, making CJ assessments less efficient. Furthermore, letting assessors write feedback comments could lead to them making more analytical judgments rather than holistic (Verhavert et al., 2019), which could harm the validity of the assessment. Secondly, the final scores obtained with CJ could be uninformative to assessees as the judgments from which the scores are derived are non-transparent regarding the assessors' decision-making processes (Steedle & Ferrara, 2016; Kelly et al., 2022; Mortier et al., 2015). Furthermore, the lack of transparency makes it challenging for assessors to evaluate the validity of quality scores. To gain insight into how assessors formed their judgments, assessors can write decision statements about how they judged the essays (Lesterhuis et al., 2022; van Daal et al., 2016; Landrieu et al., 2022). However, this is impractical as it requires more effort from the assessors and could potentially encourage more analytical judgments instead of holistic ones.

### 4.1. Automated feedback opportunities for CJ

To improve the lack of feedback opportunities for CJ and enhance its transparency, essay texts could be used to automate feedback for CJ. Based on the literature on automated feedback for AES, we discern two opportunities for extending the feedback opportunities for CJ. Both of these opportunities allow assessees to obtain feedback by comparing feedback for their essays with that for higher-quality essays.

Firstly, the final quality scores can be explained based on aspects of text quality displayed in essays (see Fig. 5). To automate feedback for AES, Kumar and Boulanger (2020) explained the predicted essay scores based on a set of higher- and lower-order linguistic features measuring relevant aspects of text quality, such as organization, style, and mechanics. To achieve this, they compared different SHapley Additive Explanations (SHAP) models which measure the marginal contributions of the aspects for each individual predicted score. In a subsequent study, Kumar and Boulanger (2021) performed feature selection for each aspect of text quality, making the explanations more accurate. Similarly, to automate feedback after CJ assessments, explanations of final quality could be provided by scoring essays on aspects of text quality measured with linguistic features. Especially, scores on higher-order aspects for text quality, such as argumentation and organization, could be informative, as they typically carry more weight when making pairwise comparisons (Lesterhuis et al., 2022; Lesterhuis, 2018). However, lower-order aspects could still be of importance, as for judgments of lower-quality essays, assessors still focus on those aspects (Humphry & Heldsinger, 2019). In practice, it is advised not to show the scores on different aspects of text quality on their own. Instead, assessors should compare their scores on these aspects to those of a better essay while also comparing the essay texts themselves. This could be done by displaying a chart with scores on aspects of quality for two essays. By letting students directly compare feedback, they better internalize the goals needed to improve their performance. Following the model of internal feedback proposed by Nicol (2021), feedback by making comparisons enables higher-order thinking of recipients (Rittle-Johnson & Star, 2011) and ultimately enhances feedback effectiveness. As a result of automating feedback that is more effective, we anticipate that assessors will not need to write as many feedback comments during assessments. This could save them time during the assessment and prevent them from making too analytical judgments. In addition, explanations of the final quality scores based on aspects of text quality could provide assessors and assessees with insight into the aspects of the assessed writing quality without requiring decision statements from them. However, the scores on linguistic features may not entirely explain the quality scores, as writing quality is a complex construct (Sadler, 1989). Therefore, an additional form of automated feedback should be provided as well.

Secondly, the final quality scores could be explained by highlighting the most important differences between their essay text and a better one. Showing exemplars of higher quality could be more instructive feedback than comments or scores on aspects of text quality (Sadler, 2014), as it is a more feed-forward form of feedback. Previously, Parekh et al. (2020) automatically highlighted the most salient text segments for predicting essay scores using the integrated gradient method. As CJ assessments consist of comparing essays in pairs, there is the opportunity to highlight the text segments in the pairs of essays in which they differ the most with respect to their quality (see Fig. 5). When highlighting text segments, the assessees can more easily compare differences in quality between two essay texts. In practice, the highlighted text segments could be shown together with scores on aspects of text quality for two essays. For instance, if an essay is found lacking in organization compared to a higher-quality essay, the feedback report would explicitly highlight where and how the organization differs, offering clear guidance for improvement. As a result, this feedback opportunity could make

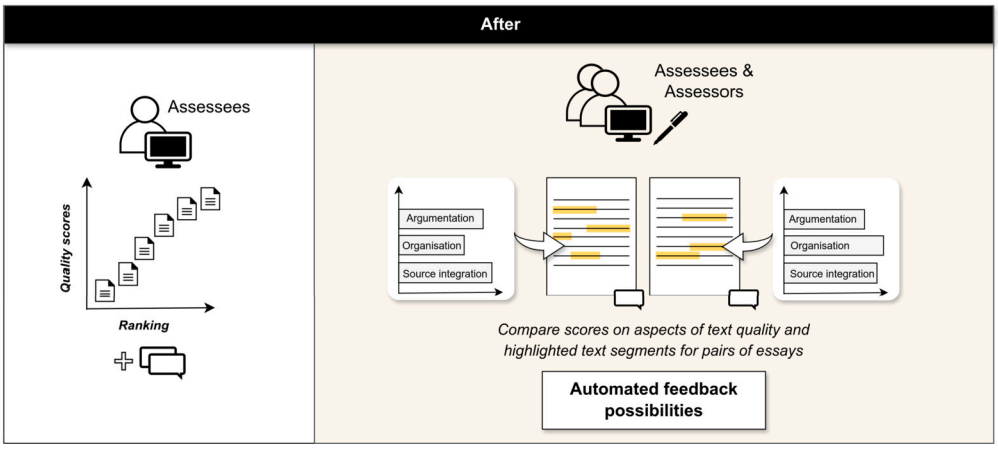


**Fig. 5.** Diagram depicting the opportunities of NLP after the CJ assessment to automate feedback based on essay texts, aiming to provide additional feedback for CJ. The automated feedback could include scores on aspects of text quality and highlighted differences between essay texts.

the judgments more transparent regarding assessors' decision-making and also prevent assessors from writing as many feedback comments during the assessment.

The explanations for the predicted scores, whether based on aspects of text quality (Boulanger & Kumar, 2020; Kumar & Boulanger, 2020, 2021) or highlighted text segments in essay texts (Parekh et al., 2020) could be obtained using explainable AI models. Such models produce explanations for each individual prediction of machine learning models (Letzgus et al., 2022), making them generally more suitable to explain predictions than statistical regression models. The most popular explainable AI models are SHAP (Lundberg & Lee, 2017) and LIME (Local Interpretable Model-agnostic Explanations) (Ribeiro et al., 2016), which explain the predictions of a machine learning model by evaluating the change in predictions when removing or perturbing input features for the model. Both SHAP and LIME are versatile, as they are applicable regardless of the essay scoring model for which they generate the explanations. However, since quality scores are interpreted based on different types of inputs, including aspects of text quality and individual texts or tokens, different essay-scoring models must be trained. A feature-based model is necessary to explain quality scores in terms of specific aspects of text quality (see Section 2.1), while a transformer-based model is required to interpret quality scores based on essay texts (see Section 2.2).

### 4.2. Considerations

For automated feedback after a CJ assessment to be effective, the feedback needs to be accepted by assessors and assessees (Anseel & Lievens, 2009). To measure the level of acceptance of feedback, assessors and assessees can be surveyed with respect to the level of trust they have in the system, as well as how accurate and comprehensible they perceive automated feedback to be (Conijn et al., 2023; Wilson et al., 2021). Furthermore, to ensure the effectiveness of automated feedback, we argue that it should align with existing feedback, serving a complementary role. Discrepancies between automated and human-written feedback, for example, when scores on specific features or the highlighted text segments in essays contradict the feedback comments assessors wrote, should be avoided. Achieving alignment in feedback could be challenging, as Dikli and Bleyle (2014) noted that AES systems show large discrepancies between human-written and automated feedback.

To evaluate the transparency introduced by feedback, we propose assessing the effectiveness of automated feedback. Instead of focusing on a social-constructivist model or the model of Hattie and Timperley (2007), we suggest using the broader model proposed by Winstone and Nash (2023). This model conceptualizes effective feedback across multiple dimensions, including the desire to receive feedback, attention to feedback, appraisal of feedback, elaboration on feedback, revisitation of feedback, and its integration into learning. In practice, these aspects of feedback effectiveness could be evaluated using a mixed-methods approach grounded in psychological science, incorporating diverse data sources such as log data, survey responses, and interviews to provide a comprehensive understanding of these aspects of feedback effectiveness.

## 5. Discussion

Assessing writing quality with CJ is regarded as both valid and reliable (van Daal et al., 2016; Verhavert et al., 2019; Wheadon et al., 2019; Jones & Inglis, 2015). However, its broader application for assessing writing quality is hindered by several practical limitations present during different stages of the assessment process. At the start of an assessment, there is no information about the quality scores of essays, as no judgments have been made yet, resulting in a cold start. During CJ assessments, assessors must make numerous judgments before the scores become reliable. Although adaptive selection rules have been proposed to address this problem, they result in marginal efficiency gains (Rangel-Smith & Lynch, 2018; Pollitt, 2012a; De Vrindt et al., 2022) or cause unintended consequences, such as selecting pairs of essays that are difficult for assessors to judge (van Daal et al., 2017; Gijsen et al., 2021) or inflating the reliability of the estimated quality scores (Bramley, 2015; Bramley & Vitello, 2019; Crompvoets et al., 2020). When the assessment is completed, there is a lack of feedback opportunities for assessees and assessors. Providing additional feedback opportunities is crucial due to the lack of transparency of the final quality scores as they are derived from pairwise judgments (Kelly et al., 2022; Steedle & Ferrara, 2016).

Based on relevant literature on NLP, we identified opportunities of using NLP at different stages of the CJ assessment. This includes appropriate methods that could be used, the overarching objectives pursued, and the key considerations to take into account. We recapitulate the discussed opportunities of NLP in Table 1. To alleviate the cold start problem at the start of the CJ assessment, we considered literature on AES. Our approach involved predicting initial quality scores based on the essay texts, both in settings where no scores of essays from other assessments are available and when they are. During the CJ assessment, we proposed to first select pairs of essays with similar unimportant features, as judging them could be easier for assessors. Then, from the pairs, exploration and exploitation strategies could be used to devise selection rules that increase the reliability of the estimated quality scores without inflating the SSR. After the assessment process, we refer to explainable AI models, such as SHAP models, to provide additional feedback opportunities for CJ, thereby enhancing the transparency of the final quality scores.

**Table 1**
Overview of opportunities of NLP for different stages of the CJ assessment, including methods that can be used with reference to prior research, the overarching objectives pursued, and key considerations to take into account.

| Stage in CJ assessment | Opportunity of NLP | Method | Prior research | Objective | Considerations |
|---|---|---|---|---|---|
| At the start (see Section 2) | Alleviating the cold start | Predicting initial quality scores based on inherent linguistic features of essays[1] or based on sets of assessed essays[2] | (Zhang & Litman, 2021; Wang et al., 2023; Mim et al., 2019; Song et al., 2020)[1] (De Vrindt et al., 2024; Dong et al., 2017; Xue et al., 2021; Wang et al., 2022)[2] | Decreasing the number of judgments assessors have to make | Ability to predict quality scores of essays written for different assignments and for essays of different lengths |
| During the assessment (see Section 3) | Selecting pairs that are not difficult to judge | Selecting essays with similar unimportant aspects of text quality to focus assessors' attention on important aspects | (Basu et al., 2013; Brooks et al., 2014; Weegar & Idestam-Almquist, 2023) | Decreasing the time needed to make judgments | Efficiency gain could be subject to assessment conditions |
| | Selecting pairs of essays that increase the reliability of scores without inflating the SSR | Selecting pairs based on essay texts using exploration and exploitation strategies | (Horbach & Palmer, 2016; Dronen et al., 2015; Hastings et al., 2018; Firoozi et al., 2023) | Decreases the required number of judgments for obtaining reliable scores | Efficiency gain could be subject to assessment conditions |
| After the assessment (see Section 4) | Providing additional feedback opportunities | Explaining the final quality scores based on linguistic features of essays and highlighted differences between pairs of essay texts | (Kumar & Boulanger, 2020, 2021; Parekh et al., 2020) | Improving transparency of assessment & Reducing effort required from assessors to provide feedback | Potential discrepancy between human-written and automated feedback |

[1] Opportunities of NLP in Setting A where no assessed essays are available.
[2] Opportunities of NLP in Setting B where assessed essays are available.

### 5.1. Relationship of assessors with AI system

When developing any AI systems to support educational assessment, it is important to consider the relationship that assessors have with the system (Molenaar, 2022). Instead of aiming to replace assessors with these systems, the prevailing view has shifted more toward supporting teachers and students. This point of view, also known as the augmentation perspective (Mavrikis et al., 2021), emphasizes that these systems should have the ability to support teaching or learning activities by leveraging human strengths while mitigating their weaknesses (Akata et al., 2020). The opportunities of NLP we put forward to improve CJ assessments align with this viewpoint, as we seek to harness the human ability to assess the writing quality of essays by letting them make judgments while minimizing their efforts through the use of NLP. The assessors' judgments are still important to enforce the reliability and validity of the scores.

The degree of automation in the assessment can be further defined based on the six levels of automation model proposed by Molenaar (2022). We view applying NLP for CJ as partial automation, where assessors and the system have distinct responsibilities: the system alleviates the cold start, efficiently allocates pairs of essays, and automates feedback, while the assessor makes judgments and writes feedback comments, just as in existing CJ assessments. By embracing partial automation, assessors retain a significant level of control over assessments, as they ultimately determine the quality scores of essays by making judgments. This allows for more control over the scores and a lesser degree of automation than is typically the case with AES systems, where the systems function as a replacement of one of the multiple assessors (Yan & Bridgeman, 2020).

Moreover, by choosing partial automation, we mitigate the risk of the essay model overfitting on specific writing genres or types of assignments in the training set. The initial predictions of quality scoring serve only as a starting point and are iteratively updated during the assessments. As more judgments are made by assessors, the impact of the initial quality scores decreases. As explained in Section 2.2, De Vrindt et al. (2024) constructed informative priors for the quality scores based on the predicted initial quality scores, which are then updated using the judgments made by assessors. A possible consequence of this process of updating the initial quality scores based on judgments is that it helps to correct systematic biases of the predictive model, while biases from assessors are also identifiable as misfits (Pollitt, 2012b). Importantly, the assessors are not reliant on the predicted quality scores when making their judgments, as they do not have access to them. However, these scores can still support the assessors during the assessment by reducing the total number of judgments required and preventing the selection of pairs that are difficult to judge.

From a privacy perspective, the essay-scoring models do not have access to any personally identifiable information to make predictions, safeguarding students' privacy. This approach extends to the selection rule as well, which operates solely based on the content of the essays, rather than any personal data. Assessors also do not have knowledge of the initial predicted quality scores used to alleviate the cold start. This approach extends to the selection rule as well; assessors have no knowledge of which essay belongs to whom and what initial quality scores were predicted. Additionally, the automation of feedback with explainable NLP techniques could enhance the transparency of the assessment. Both assessors and assessees can see how the final quality scores are derived by examining the scores on relevant aspects of text quality and observing highlighted differences between essay texts. This level of transparency not only clarifies the quality construct being measured but also provides a basis for ensuring the fairness of the assessment process.

### 5.2. Practical implications

It is important to recognize that NLP can serve goals beyond improving the efficiency and transparency of CJ. AI applications in education often focus on enhancing personalized learning experiences (AlShaikh & Hewahi, 2021). In the context of CJ, NLP could be used to tailor feedback based on assessees' prior performances or individual characteristics. For instance, a pair selection rule could also be personalized to assessors to enhance their learning by making comparisons as in formative assessments (Bartholomew et al., 2019). In formative CJ assessments, it is less important to achieve high efficiency by minimizing the number of judgments and the time needed to make judgments. Instead, the goal is to help assessors learn by making judgments of essay pairs. In line with variation theory by Guo et al. (2012), a selection rule could be developed using NLP that selects pairs of essays that differ in features not yet

mastered by an assessor while being similar in features an assessor has already mastered.

Additionally, there is an opportunity to enhance the large-scale assessments of essay writing, especially in contexts where there are vast numbers of students and assessors involved. For instance, CJ is already used for large-scale assessments in the UK, involving over 50,000 students and more than 12,000 assessors (Wheadon et al., 2019). Wheadon et al. (2019) concluded that CJ is highly suitable for high-stakes assessments as it produces reliable and valid scores. Integrating NLP into large-scale assessments with CJ could increase the efficiency of the assessment, and therefore, reduce costs. Specifically, as discussed in this study, NLP could alleviate the cold-start problem and help develop a more efficient selection, which reduces the number of judgments assessors have to make while avoiding difficult comparisons. Furthermore, for recurring assessments, scores from previous years could serve as valuable training data for essay-scoring models and result in more reliable predicted initial quality scores. This, in turn, would make conducting the large-scale assessments with CJ more efficient.

## 6. Conclusion

In this conceptual study, we explored multiple opportunities of NLP to mitigate key challenges present at different stages of CJ assessments, with the aim of broadening the use of CJ for assessing writing quality of essays. A conceptual paper is necessary, given the little available research in this field. We argued that NLP can be used at the start of an assessment to alleviate the cold start by predicting the quality based on essay texts. During the assessment process, we discussed how NLP could be used to select pairs of essays based on essay texts, ensuring that the pair is not overly difficult to judge and, when judged, increases the reliability without inflating the SSR. Consequently, we anticipate that the assessment process will become more efficient than current CJ assessments. After the assessment, we argued that NLP could be used to explain quality scores by displaying scores on the most relevant essay features and highlighting important differences between pairs of essay texts. By automating feedback with NLP, we anticipate that more effective feedback will be provided to assessors and assessees than currently available, thereby enhancing the transparency of the scores. Ultimately, the opportunities of NLP to improve CJ are designed to support assessors when conducting CJ assessments. At the same time, they retain control over the assessment as they still have to make judgments on the quality of essays. We presented a framework of opportunities for integrating NLP into CJ assessments, intended to guide future empirical research. The validation of these NLP applications remains a subject for further investigation.

## CRediT authorship contribution statement

**Michiel De Vrindt:** Writing – review & editing, Writing – original draft, Visualization, Conceptualization. **Anaïs Tack:** Writing – review & editing, Supervision, Conceptualization. **Wim Van den Noortgate:** Writing – review & editing, Supervision, Conceptualization. **Marije Lesterhuis:** Writing – review & editing, Supervision, Conceptualization. **Renske Bouwer:** Writing – review & editing, Supervision, Conceptualization.

## Statements on open data and ethics

As no data was collected for this study, no participants were involved.

## Declaration of competing interest

This work was supported by a grant gifted by Flanders Innovation & Entrepreneurship Foundation (VLAIO) (HBC.2022.0164) to Michiel De Vrindt, in collaboration with the company Comproved (D-Pac BV) and KU Leuven. Marije Lesterhuis is co-founder of Comproved (D-Pac BV). Anaïs Tack, Wim Van den Noortgate, and Renske Bouwer declare they have no competing interests.

## References

Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, *53*, 18–28. https://doi.org/10.1109/MC.2020.2996587.

AlShaikh, F., & Hewahi, N. (2021). AI and machine learning techniques in the development of intelligent tutoring system: A review. In *2021 international conference on innovation and intelligence for informatics, computing, and technologies (3ICT)* (pp. 403–410). IEEE. https://doi.org/10.1109/3ICT53449.2021.9582029.

Anseel, F., & Lievens, F. (2009). The mediating role of feedback acceptance in the relationship between feedback and attitudinal and performance outcomes. *International Journal of Selection and Assessment*, *17*, 362–376. https://doi.org/10.1111/j.1468-2389.2009.00479.x.

Baniya, S., Mentzer, N., Bartholomew, S. R., Chesley, A., Moon, C., & Sherman, D. (2019). Using adaptive comparative judgment in writing assessment. *The Journal of Technology Studies*, *45*, 24–35. https://doi.org/10.21061/jots.v45i1.a.3.

Bartholomew, S. R., Strimel, G. J., & Yoshikawa, E. (2019). Using adaptive comparative judgment for student formative feedback and learning during a middle school design project. *International Journal of Technology and Design Education*, *29*, 363–385. https://doi.org/10.1007/s10798-018-9442-7.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402. https://doi.org/10.1162/tacl_a_00236.

Beigman Klebanov, B., Flor, M., & Gyawali, B. (2016). Topicality-based indices for essay scoring. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 63–72). Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0507.

Berger-Tal, O., Nathan, J., Meron, E., & Saltz, D. (2014). The exploration-exploitation dilemma: A multidisciplinary framework. *PLoS ONE*, *9*, 1–8. https://doi.org/10.1371/journal.pone.0095693.

Bi, J. (2003). Agreement and reliability assessments for performance of sensory descriptive panel. *Journal of Sensory Studies*, *18*, 61–76. https://doi.org/10.1111/j.1745-459X.2003.tb00373.x.

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment Evaluation Higher Education*, *34*, 209–220. https://doi.org/10.1080/02602930801955978.

Boulanger, D., & Kumar, V. (2020). SHAPed automated essay scoring: Explaining writing features' contributions to English writing organization. In V. Kumar, & C. Troussas (Eds.), *Intelligent tutoring systems* (pp. 68–78). Cham: Springer International Publishing.

Bouwer, R., Lesterhuis, M., De Smedt, F., Van Keer, H., & De Maeyer, S. (2023). Comparative approaches to the assessment of writing: Reliability and validity of benchmark rating and comparative judgement. *Journal of Writing Research*, *15*, 497–518. https://doi.org/10.17239/jowr-2024.15.03.03.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, *39*, 324–345.

Bramley, T. (2007). Paired comparison methods. In P. Newton, J. Baird, H. Goldstein, H. Patrick, & P. Tymms (Eds.), *Techniques for monitoring the comparability of examination standards. Qualifications and curriculum* (pp. 246–300). London, London, United Kingdom: Authority.

Bramley, T. (2015). *Investigating the reliability of adaptive comparative judgment*. Technical report, University of Cambridge. https://www.cambridgeassessment.org.uk/Images/232694-investigating-the-reliability-of-adaptive-comparative-judgment.pdf.

Bramley, T., Bell, J. F., & Pollitt, A. (1998). Assessing changes in standards over time using thurstone paired comparisons. *Education Research and Perspectives*, *25*, 14.

Bramley, T., & Vitello, S. (2019). The effect of adaptivity on the reliability coefficient in adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*, 43–58. https://doi.org/10.1080/0969594X.2017.1418734.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*, 1–21. https://doi.org/10.1080/08957347.2011.532417.

Brooks, M., Basu, S., Jacobs, C., & Vanderwende, L. (2014). Divide and correct: Using clusters to grade short answers at scale. In *Proceedings of the first ACM conference on learning @ scale conference* (pp. 89–98). https://doi.org/10.1145/2556325.2566243.

Chowdhary, K. R. (2020). Natural language processing. In *Fundamentals of artificial intelligence* (pp. 603–649). New Delhi: Springer India.

Coenen, T., Coertjens, L., Vlerick, P., Lesterhuis, M., Mortier, A. V., Donche, V., Ballon, P., & De Maeyer, S. (2018). An information system design theory for the comparative judgement of competences. *European Journal of Information Systems*, *27*, 248–261. https://doi.org/10.1080/0960085X.2018.1445461.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Conijn, R., Kahr, P., & Snijders, C. (2023). The effects of explanations in automated essay scoring systems on student trust and motivation. *Journal of Learning Analytics*, *10*, 37–53. https://doi.org/10.18608/jla.2023.7801.

Crompvoets, E. A. V., Béguin, A. A., & Sijtsma, K. (2020). Adaptive pairwise comparison for educational measurement. *Journal of Educational and Behavioral Statistics*, *45*, 316–338. https://doi.org/10.3102/1076998619890589.

van Daal, T. (2020). *Making a choice is not easy?! Unravelling the task difficulty of comparative judgement to assess student work*. Ph.D. thesis, University of Antwerp.

van Daal, T., Lesterhuis, M., Coertjens, L., Donche, V., & De Maeyer, S. (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 & Practice*, *26*, 59–74. https://doi.org/10.1080/0969594X.2016.1253542.

van Daal, T., Lesterhuis, M., Coertjens, L., van de Kamp, M. T., Donche, V., & De Maeyer, S. (2017). The complexity of assessing student work using comparative judgment: The moderating role of decision accuracy. *Frontiers in Education*, *2*, 1–13. https://doi.org/10.3389/feduc.2017.00044.

De Vrindt, M., Van den Noortgate, W., & Debeer, D. (2022). Text mining to alleviate the cold-start problem of adaptive comparative judgments. *Frontiers in Education*, *7*, 132–147. https://doi.org/10.3389/feduc.2022.854378.

De Vrindt, M., Tack, A., Bouwer, R., Van Den Noortgate, W., & Lesterhuis, M. (2024). Predicting initial essay quality scores to increase the efficiency of comparative judgment assessments. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th workshop on innovative use of NLP for building educational applications (BEA 2024)* (pp. 125–136). Association for Computational Linguistics.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing*, *22*, 1–17. https://doi.org/10.1016/j.asw.2014.03.006.

Do, H., Kim, Y., & Lee, G. G. (2023). Prompt- and trait relation-aware cross-prompt essay trait scoring. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: ACL 2023* (pp. 1538–1551). Association for Computational Linguistics.

Dong, F., Zhang, Y., & Yang, J. (2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In R. Levy, & L. Specia (Eds.), *Proceedings of the 21st conference on computational natural language learning (CoNLL 2017)* (pp. 153–162). Association for Computational Linguistics.

Dong, Z., & Tang, T., Li, L., Zhao, W. X. A survey on long text modeling with transformers. https://doi.org/10.48550/arXiv.2302.14502 (2023).

Dronen, N., Foltz, P. W., & Habermehl, K. (2015). Effective sampling for large-scale automated writing evaluation systems. In *Proceedings of the second (2015) ACM conference on learning @ scale* (pp. 3–10). New York, NY, USA: Association for Computing Machinery.

D'Arcy, J. (1997). *Comparability studies between modular and non-modular syllabuses in gce advanced level biology, English literature and mathematics in the 1996 summer examinations*. Standing Committee on Research on behalf of the Joint Forum for the GCSE and GCE.

Firoozi, T., Mohammadi, H., & Gierl, M. J. (2023). Using active learning methods to strategically select essays for automated scoring. *Educational Measurement, Issues and Practice*, *42*, 34–43. https://doi.org/10.1111/emip.12537.

Ghosh, D., Khanam, A., Han, Y., & Muresan, S. (2016). Coarse-grained argumentation features for scoring persuasive essays. In K. Erk, & N. A. Smith (Eds.), *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)* (pp. 549–554). Berlin, Germany: Association for Computational Linguistics.

Gijsen, M., van Daal, T., Lesterhuis, M., Gijbels, D., & De Maeyer, S. (2021). The complexity of comparative judgments in assessing argumentative writing: An eye tracking study. *Frontiers in Education*, *5*, 1–11. https://doi.org/10.3389/feduc.2020.582800.

Guo, J. P., Pang, M. F., Yang, L. Y., & Ding, Y. (2012). Learning from comparing multiple examples: On the dilemma of "similar" or "different". *Educational Psychology Review*, *24*, 251–269. https://doi.org/10.1007/s10648-012-9192-0.

Hastings, P., Hughes, S., & Britt, M. A. (2018). Active learning for improving machine learning of student explanatory essays. In *International conference on artificial intelligence in education* (pp. 140–153). Springer.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81–112. https://doi.org/10.3102/003465430298487.

Heldsinger, S., & Humphry, S. (2010). Using the method of pairwise comparison to obtain reliable teacher assessments. *The Australian Educational Researcher*, *37*, 1–19. https://doi.org/10.1007/BF03216919.

Horbach, A., & Palmer, A. (2016). Investigating active learning for short-answer scoring. In J. Tetreault, J. Burstein, C. Leacock, & H. Yannakoudakis (Eds.), *Proceedings of the 11th workshop on innovative use of NLP for building educational applications* (pp. 301–311). Association for Computational Linguistics.

Humphry, S., & Heldsinger, S. (2019). Raters' perceptions of assessment criteria relevance. *Assessing Writing*, *41*, 1–13. https://doi.org/10.1016/j.asw.2019.04.002.

Jones, I., Bisson, M., Gilmore, C., & Inglis, M. (2019). Measuring conceptual understanding in randomised controlled trials: Can comparative judgement help? *British Educational Research Journal*, *45*, 662–680. https://doi.org/10.1002/berj.3519.

Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics*, *89*, 337–355. https://doi.org/10.1007/s10649-015-9607-1.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment Evaluation Higher Education*, *39*, 840–852. https://doi.org/10.1080/02602938.2013.875117.

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 6300–6308). International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2019/879.

Kelly, K., Richardson, M., & Isaacs, T. (2022). Critiquing the rationales for using comparative judgement: A call for clarity. *Assessment in Education: Principles, Policy & Practice*, *29*, 1–15. https://doi.org/10.1080/0969594X.2022.2147901.

Kim, S. (2012). A note on the reliability coefficients for item response model-based ability estimates. *Psychometrika*, *77*, 153–162. https://doi.org/10.1007/S11336-011-9238-0.

Kumar, V., & Boulanger, D. (2020). Explainable automated essay scoring: Deep learning really has pedagogical value. *Frontiers in Education*, *5*, Article 572367. https://doi.org/10.3389/feduc.2020.572367.

Kumar, V., & Boulanger, D. (2021). Automated essay scoring and the deep learning black box: How are rubric scores determined? *International Journal of Artificial Intelligence in Education*, *31*, 538–584. https://doi.org/10.1007/s40593-020-00211-5.

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

Landrieu, Y., De Smedt, F., Van Keer, H., & De Wever, B. (2022). Assessing the quality of argumentative texts: Examining the general agreement between different rating procedures and exploring inferences of (dis) agreement cases. *Frontiers in Education*, *7*, 106–121. https://doi.org/10.3389/feduc.2022.784261.

Latifi, S., & Gierl, M. (2021). Automated scoring of junior and senior high essays using coh-metrix features: Implications for large-scale language testing. *Language Testing*, *38*, 62–85. https://doi.org/10.1177/0265532220929918.

Le, Q., & Mikolov, T. (2014). Distributed representations of sentences and documents. In E. P. Xing, & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning* (pp. 1188–1196). Bejing, China: PMLR.

Lesterhuis, M. (2018). When teachers compare argumentative texts: Decisions informed by multiple complex aspects of text quality. *L1-Educational Studies in Language and Literature*, *18*, 1–22. https://doi.org/10.17239/L1ESLL-2018.18.01.02.

Lesterhuis, M., Bouwer, R., van Daal, T., Donche, V., & De Maeyer, S. (2022). Validity of comparative judgment scores: How assessors evaluate aspects of text quality when comparing argumentative texts. *Frontiers in Education*, *7*, 122–131. https://doi.org/10.3389/feduc.2022.823895.

Letzgus, S., Wagner, P., Lederer, J., Samek, W., Müller, K. R., & Montavon, G. (2022). Toward explainable artificial intelligence for regression models: A methodological perspective. *IEEE Signal Processing Magazine*, *39*, 40–58. https://doi.org/10.1109/MSP.2022.3153277.

Li, X., Chen, M., & Nie, J. Y. (2020). Sednn: Shared and enhanced deep neural network model for cross-prompt automated essay scoring. *Knowledge-Based Systems*, *210*, Article 106491. https://doi.org/10.1016/j.knosys.2020.106491.

Linacre, J. M. (1989). *Many-faceted Rasch measurement*. Ph.D. thesis, The University of Chicago.

Liu, S., Bremer, P. T., Thiagarajan, J. J., Srikumar, V., Wang, B., Livnat, Y., & Pascucci, V. (2017). Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, *24*, 553–562.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*, 81–95. https://doi.org/10.1037/h0043178.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 4768–4777). Curran Associates, Inc.

Mavrikis, M., Cukurova, M., Di Mitri, D., Schneider, J., & Drachsler, H. (2021). A short history, emerging challenges and co-operation structures for artificial intelligence in education. *Bildung und Erziehung*, *74*, 249–263. https://doi.org/10.13109/buer.2021.74.3.249.

McCarthy, A. D., Yancey, K. P., LaFlair, G. T., Egbert, J., Liao, M., & Settles, B. (2021). Jump-starting item parameters for adaptive language tests. In M. F. Moens, X. Huang, L. Specia, Yih, & S. W. t (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 883–899). Association for Computational Linguistics.

McGrane, J. A., Humphry, S. M., & Heldsinger, S. (2018). Applying a thurstonian, two-stage method in the standardized assessment of writing. *Applied Measurement in Education*, *31*, 297–311. https://doi.org/10.1080/08957347.2018.1495216.

Medin, D. L., Goldstone, R. L., & Markman, A. B. (1995). Comparison and choice: Relations between similarity processes and decision processes. *Psychonomic Bulletin & Review*, *2*, 1–19.

Mim, F. S., Inoue, N., Reisert, P., Ouchi, H., & Inui, K. (2019). Unsupervised learning of discourse-aware text representation for essay scoring. In F. Alva-Manchego, E. Choi, & D. Khashabi (Eds.), *Proceedings of the 57th annual meeting of the association for computational linguistics: Student research workshop* (pp. 378–385). Association for Computational Linguistics.

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*, Article 100050. https://doi.org/10.1016/j.rmal.2023.100050.

Molenaar, I. (2022). Towards hybrid human-ai learning technologies. *European Journal of Education*, *57*, 632–645. https://doi.org/10.1111/ejed.12527.

Mortier, A. V., Lesterhuis, M., Vlerick, P., & De Maeyer, S. (2015). Comparative judgment within online assessment: Exploring students feedback reactions. In E. Ras, &

D. Joosten-ten Brinke (Eds.), *Computer assisted assessment. Research into E-assessment* (pp. 69–79). Cham: Springer International Publishing.

Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., & Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, *56*, 3005–3054. https://doi.org/10.1007/s10462-022-10246-w.

Newhouse, C. P. (2014). Using digital representations of practical production work for summative assessment. *Assessment in Education: Principles, Policy & Practice*, *21*, 205–220. https://doi.org/10.1080/0969594X.2013.868341.

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment Evaluation Higher Education*, *46*, 756–778. https://doi.org/10.1080/02602938.2020.1823314.

Parekh, S., Singla, Y. K., Chen, C., Li, J. J., & Shah, R. R. (2020). My teacher thinks the world is flat! Interpreting automatic essay scoring mechanism. https://doi.org/10.48550/arXiv.2012.13872. arXiv:2012.13872.

Penha, G., & Hauff, C. (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* (pp. 388–397). New York, NY, USA: Association for Computing Machinery.

Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. In L. Màrquez, C. Callison-Burch, & J. Su (Eds.), *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 431–439). Lisbon, Portugal: Association for Computational Linguistics.

Pilehvar, M. T., & Camacho-Collados, J. (2020). *Embeddings in natural language processing: Theory and advances in vector representations of meaning*. Morgan & Claypool Publishers.

Pliakos, K., Joo, S. H., Park, J. Y., Cornillie, F., Vens, C., & Van den Noortgate, W. (2019). Integrating machine learning into item response theory for addressing the cold start problem in adaptive learning systems. *Computers and Education*, *137*, 91–103. https://doi.org/10.1016/j.compedu.2019.04.009.

Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education*, *22*, 157–170. https://doi.org/10.1007/s10798-011-9189-x.

Pollitt, A. (2012b). The method of adaptive comparative judgement. Assessment in education: Principles. *Policy & Practice*, *19*, 281–300. https://doi.org/10.1080/0969594X.2012.665354.

Pollitt, A., & Whitehouse, C. (2012). Using adaptive comparative judgement to obtain a highly reliable rank order in summative assessment. Technical report, Manchester, UK: Assessment and Qualifications Alliance.

Posten, A. C., & Mussweiler, T. (2017). That certain something! Focusing on similarities reduces judgmental uncertainty. *Cognition*, *165*, 121–125. https://doi.org/10.1016/j.cognition.2017.05.010.

Potter, T., Englund, L., Charbonneau, J., MacLean, M., Newell, J., & Roll, I. (2017). Compair: A new online tool using adaptive comparative judgement to support learning with peer feedback. *Teaching & Learning Inquiry*, *5*, 89. https://doi.org/10.20343/teachlearninqu.5.2.8.

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review*, *55*, 2495–2527. https://doi.org/10.1007/s10462-021-10068-2.

Rangel-Smith, C., & Lynch, D. (2018). Addressing the issue of bias in the measurement of reliability in the method of adaptive comparative judgment. In *36th pupils' atitudes towards technology conference* (pp. 378–387).

Ribeiro, M., Singh, S., & Guestrin, C. (2016). "why should I trust you¿": Explaining the predictions of any classifier. In J. DeNero, M. Finlayson, & S. Reddy (Eds.), *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Demonstrations* (pp. 97–101). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3020.

Rittle-Johnson, B., & Star, J. R. (2011). The power of comparison in learning and instruction: Learning outcomes supported by different types of comparisons. *Psychology of Learning and Motivation*, *55*, 199–225. https://doi.org/10.1016/B978-0-12-387691-1.00007-7.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119–144.

Sadler, D. R. (2014). Ah!… so that's 'quality'. In P. Schwartz, & G. Webb (Eds.), *Assessment: Case studies, experience and practice from higher education* (pp. 130–136). London: Kogan Page.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). Methods and metrics for cold-start recommendations. In *Proceedings of the 25th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 253–260). New York, NY, USA: Association for Computing Machinery.

Settles, B., LaFlair, G. T., & Hagiwara, M. (2020). Machine learning–driven language assessment. In M. Johnson, B. Roark, & A. Nenkova (Eds.), *Transactions of the association for computational linguistics* (pp. 247–263). Cambridge, MA: MIT Press.

Somasundaran, S., Burstein, J., & Chodorow, M. (2014). Lexical chaining for measuring discourse coherence quality in test-taker essays. In J. Tsujii, & J. Hajic (Eds.), *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers* (pp. 950–961). Dublin, Ireland: Dublin City University and Association for Computational Linguistics.

Song, W., Zhang, K., Fu, R., Liu, L., Liu, T., & Cheng, M. (2020). Multi-stage pre-training for automated Chinese essay scoring. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), online* (pp. 6723–6733). Association for Computational Linguistics.

Steedle, J. T., & Ferrara, S. (2016). Evaluating comparative judgment as an approach to essay scoring. *Applied Measurement in Education*, *29*, 211–223. https://doi.org/10.1080/08957347.2016.1171769.

Sun, G., Wei, W., Cui, T., Xu, D., Chen, S., Shvonski, A., Li, L., Shen, J., & Garshasbi, S. (2022a). Adapting new learners and new resources to micro open learning via online computation. *IEEE Transactions on Computational Social Systems*, *9*, 1807–1819. https://doi.org/10.1109/TCSS.2022.3210406.

Sun, J., Song, T., Song, J., & Peng, W. (2022b). Improving automated essay scoring by prompt prediction and matching. *Entropy*, *24*, 1–15. https://doi.org/10.3390/e24091206.

Taghipour, K., & Ng, H. T. (2016). A neural approach to automated essay scoring. In J. Su, K. Duh, & X. Carreras (Eds.), *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 1882–1891). Association for Computational Linguistics.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review*, *34*, 273–286. https://doi.org/10.1037/h0070288.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology*, *21*, 384.

Uto, M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, *48*, 459–484. https://doi.org/10.1007/s41237-021-00142-y.

Van Gasse, R., Mortier, A., Goossens, M., Vanhoof, J., Van Petegem, P., Vlerick, P., & De Maeyer, S. (2017). Feedback opportunities of comparative judgement: An overview of possible features and acceptance at different user levels. In D. Joosten-ten Brinke, & M. Laanpere (Eds.), *Technology enhanced assessment* (pp. 23–38). Cham: Springer International Publishing.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice*, *26*, 541–562. https://doi.org/10.1080/0969594X.2019.1602027.

Verhavert, S., De Maeyer, S., Donche, V., & Coertjens, L. (2018). Scale separation reliability: What does it mean in the context of comparative judgment? *Applied Psychological Measurement*, *42*, 428–445. https://doi.org/10.1177/0146621617748321.

Verhavert, S., Furlong, A., & Bouwer, R. (2022). The accuracy and efficiency of a reference-based adaptive selection algorithm for comparative judgment. *Frontiers in Education*, *6*, Article 785919. https://doi.org/10.3389/feduc.2021.785919.

Wang, C., Jiang, Z., Yin, Y., Cheng, Z., Ge, S., & Gu, Q. (2023). Aggregating multiple heuristic signals as supervision for unsupervised automated essay scoring. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 13999–14013). Toronto, Canada: Association for Computational Linguistics.

Wang, Y., Wang, C., Li, R., & Lin, H. (2022). On the use of bert for automated essay scoring: Joint learning of multi-scale essay representation. In M. Carpuat, M. C. de Marneffe, & I. V. Meza Ruiz (Eds.), *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 3416–3425). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.249.

Weegar, R., & Idestam-Almquist, P. (2023). Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*, *1–27doi*. https://doi.org/10.1007/s40593-022-00322-1.

Wheadon, C., Barmby, P., Christodoulou, D., & Henderson, B. (2019). A comparative judgement approach to the large-scale assessment of primary writing in England. *Assessment in Education: Principles, Policy & Practice*, *27*, 1–19. https://doi.org/10.1080/0969594X.2019.1700212.

Wilson, J., Huang, Y., Palermo, C., Beard, G., & MacArthur, C. A. (2021). Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of mi write. *International Journal of Artificial Intelligence in Education*, *31*, 234–276. https://doi.org/10.1007/s40593-020-00236-w.

Winstone, N. E., & Nash, R. A. (2023). Toward a cohesive psychological science of effective feedback. *Educational Psychologist*, *58*, 111–129. https://doi.org/10.1080/00461520.2023.2224444.

Xue, J., Tang, X., & Zheng, L. (2021). A hierarchical bert-based transfer learning approach for multi-dimensional essay scoring. *IEEE Access*, *9*, 125403–125415. https://doi.org/10.1109/ACCESS.2021.3110683.

Yan, D., & Bridgeman, B. (2020). Validation of automated scoring systems. In D. Yan, A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring* (pp. 297–318). Chapman and Hall/CRC.

Yaneva, V., & von Davier, M. (2023). Psychometric considerations when using deep learning for automated scoring. In *Advancing natural language processing in educational assessment* (1 ed.) (pp. 15–30). New York, United States: Routledge.

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the tenth workshop on innovative use of NLP for building educational applications* (pp. 224–232). Denver, Colorado: Association for Computational Linguistics.

Zhang, H., & Litman, D. (2021). Essay quality signals as weak supervision for source-based essay scoring. In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, & T. Zesch (Eds.), *Proceedings of the 16th workshop on innovative use of NLP for building educational applications, online* (pp. 85–96). Association for Computational Linguistics.