Employing a Hierarchical Rater Models for Automated Scoring: Scope Review on the Application in Educational Assessment

Akif AVCU [1]

[1] avcuakif@gmail.com, orcid.org/ 0000-0003-1977-7592, Marmara University, Türkiye

To Cite: Avcu, A. (2025). Employing a hierarchical rater models for automated scoring: Scope review on the application in educational assessment. *Malaysian Online Journal of Educational Technology, 13*(2), 37-47. http://dx.doi.org/10.52380/mojet.2025.13.2.569

ABSTRACT

This scope-review presents the milestones of how Hierarchical Rater Models (HRMs) become operable to used in automated essay scoring (AES) to improve instructional evaluation. Although essay evaluations—a useful instrument for evaluating higher-order cognitive abilities—have always depended on human raters, concerns regarding rater bias, inconsistency, and scalability have motivated the development of automated systems. By modeling rater behaviors, task complexity, and interaction effects, HRMs handle these issues and offer a strong base to minimize biases and increase dependability. Advances in machine learning and natural language processing (NLP) have helped HRMs to be included into AES systems. Leveraging HRMs to rectify rater biases and guarantee fairness, these systems include language analysis, semantic evaluation, and contextual understanding. This review also includes how Signal Detection Theory (SDT) is being included into HRMs to improve their capacity to assess rater sensitivity and bias and provide understandable results. By including HRMs into AES, feedback quality is improved as well as score accuracy, hence facilitating more exact formative evaluations become possible. Still present, though, are difficulties like computing complexity, dataset availability, and algorithmic bias. The paper emphasizes the possibilities of HRMs in creating fair, high-quality, scalable evaluation systems and supports ongoing research to improve these approaches for various educational environments.

Keywords: *Hierarchical rater model,; automated scoring; essays; Scope review*

Article History:
Received: 4 August 2024
Received in revised form: 14 Oct.. 2024
Accepted: 30 November 2024
Article type: Review Article

INTRODUCTION

Transition of essay evaluation from human raters to automated systems

The assessment of student learning is critically important in the rapidly evolving educational settings. The qualith of evaluation for learning outcomes had been done via tests and examinations and can be enhanced by diverse question formats and grading techniques. Question types can be formulated to include several formats, ranging from basic multiple-choice questions to those requiringhigher-order cognitive skills, such as essays and other kinds of complex tasks. Traditional evaluation methods, such multiple-choice and true-false examinations, prioritize recaling and recognizing but sometimes cannot catch deeper cognitive skills (Simkin & Kuechler, 2005). On the other hand, essay examinations are especially effective in assessing a broader spectrum of competencies, such as critical thinking skills and reasoning. Essay tests facilitate

students' engagement in higher-order cognitive processes. Such evaluations require the application, analysis, synthesis, and appraisal levels of cognitive skills (Bloom, 1956; Anderson et al., 2001).

Essay questions necessitate that students synthesize information from diverse sources, draw connections among concepts, and exhibit a thorough comprehension of the subject (Biggs & Tang, 2011). According to Ramsden (1992), when students interact with knowledge meaningfully rather than simply memorizing it, a higher order knowledge occur when they take exams with essays. In their inspirational study on learning methodologies, Marton and Säljö (1976) points to the significance of assignments, such as essays, in fostering high engagement by students. Essay assessments require students to explain their concepts clearly, to organize their thoughts and to persuasively write. Essays could become a real challenge to their capacity to convey complicated ideas (Elbow, 1998; Hyland, 2003). Unlike standardized assessments, essay examinations allow for the constructing of personal viewpoints. Students can interpret questions based on their distinct experiences and views, so it increase their motivation and engagement in educational environments (Entwistle & Ramsden, 1983; Andrade & Du, 2007). Moreover, open-ended problems foster creativity, as noted by Amabile (1996), further differentiating essay from other evaluation forms.

Despite the important benefits of essays, they also present certain kinds of problems on practical side. Essay grading are costly and may introduce subjective biases causing from human raters (Brown & Knight, 1994). These concerns can be handled by implementing standardized rubrics, rigid grading criteria, and strictly training of evaluators (Sadler, 2009). The assessment of essays by human raters in educational settings has a long history and integrates professional judgment with educational goals. Essay scoring, especially in educational assessment, demand a careful balance between objectivity and the assessment of latent attributes such as creativity, reasoning, and rhetorical efficacy. The usage of human evaluators has been a crucial element of this procedure for many centuries. Despite the concerns for inter-rater reliability and the possibility of bias, human evaluators are integral part for assessing of writing that machines cannot interpret with the necessary nuance, including tone, audience awareness, and argument coherence (Weigle, 2002; Barkaoui, 2010).

The assessment of subjective activities, including essay composition, is an area where human evaluators has been deemed essential. Inconsistencies in evaluator judgments, potentially stemming from biases, leniency, severity, and variations in task complexity, might deterrioate the reliability and validity of evaluations (Engelhard, 2013). Hierarchical Rater Models (HRMs) have been created to tackle these difficulties by mathematically modeling rater behavior by focusing on the multilevel nature of data.

Poe and Elliot (2019) states that teacher-based scoring in educational environments is a notably valid method due to teachers' valid comprehension of the instructional context and the developmental needs of their students. This contextual comprehension enables a better assessment of student development and achievement. Nonetheless, the study suggests that the validity of this method may be undermined by leniency bias, wherein educators may exaggerate scores to represent effort rather than assessing based objective criteria. Human raters often utilize holistic or analytic scoring rubrics for essay evaluation. Holistic scoring assigns a single score representing the overall impression s/he got from the essay, while analytic scoring is done by focusing on specific elements, such as content, organization, of the essay and language (Weigle, 2002). Holistic scoring is advantageous for its efficiency, although it is criticized for its comparatively lower diagnostic accuracy. Conversely, analytic scoring provides detailed feedback but is much more arduous (Attali & Burstein, 2006).

The most salient benefit of human scoring is its ability to provide information for instructional decisions and improve student learning. Research demonstrates that instructor input given throughout the evaluation process provides formative advantages. Shepard (2000) asserts that teacher-scored essays often yields diagnostic feedback that direct students in improving their essay composing skills, while making the scoring process both pedagogical. Furthermore, the application of analytic scoring rubrics in educational settings has been shown to aid instructors in identifying specific areas for student deficiency or improvement, such as essay formulation or paragraph congruity (Brookhart, 2013). Research indicates that the effectiveness of such feedback depends on its clarity and association with the specified educational objectives (Hattie & Timperley, 2007).

Ensuring that raters are thoroughly taught in the consistent application of rubrics is essential for guaranteeing the dependability of the scoring process. Training sessions sometimes include norming activities, when raters assess benchmark writings and discuss differences to align their readings of the rubric (Lumley, 2005). Additionally, calibration and monitoring during scoring sessions are utilized to ensure score uniformity. Without training or precautionary steps during the evaluation process, the quality of the assessment is likely to deteriorate.

Maintaining uniformity among raters is a persistent difficulty in essay evaluation. Research indicates that even trained raters may exhibit variability in their interpretations of scoring rubrics, especially when assessing complex elements like argumentation and style (Breland et al., 1999). Agreement quantification is often accomplished using inter-rater reliability coefficients (e.g., Cohen's kappa, intra-class correlation). Nonetheless, achieving adequate levels continues to be a formidable challenge in several circumstances. Research has shown a frequent disparity among evaluators in their judgments of advanced writing competencies, such as reasoning and organization, relative to superficial attributes like grammar and mechanics (Weigle, 2002). Barkaoui (2010) found that less experienced raters exhibited greater variability in their grading, while more experienced raters showed increased consistency. These findings underscore the imperative of rater training and calibration sessions to guarantee uniform interpretation of scoring criteria, especially in educational contexts where educators may exhibit disparate levels of assessment proficiency.

Şata and Karakaya (2022) examined how pre training of raters affect the errors made by raters when they evaluate essays. The study utilized a pretest-posttest control group design and 45 raters participated to the study and incorporated to rater training The findings showed that the experimental group, which underwent training, conducted more valid and reliable assessments, suggesting that rater training successfully minimize the errors associated with severity, leniency, central tendency, and the halo effect.

Research indicates that people grading exams may have biases. These biases can come from the test-takers' characteristics like gender, race, or even handwriting quality (Meadows & Billington, 2005). When graders have to evaluate many essays in big exams, they get tired, which makes grading less accurate and consistent (Schoonen, 2005). Tired graders often give average scores or ignore the scoring guidelines. Also, having more graders helps ensure scores are reliable. To get trustworthy scores, especially for subjective parts like writing style or readability, it's important to use a larger number of graders (Schoonen, 2005). More graders usually lead to fairer ratings. Balancing costs, practicality, and dependable grading is crucial (Lumley, 2005). Human grading can be costly and complicated. Large exams like the SAT and GRE involve high expenses and require hundreds of graders (Powers et al., 2001).

Because there are problems with how humans score essays, a new method uses computers for scoring, known as automated scoring. The rise in online courses and many standardized tests has increased the need for human scoring abilities. Automated Essay Scoring (AES) is a good alternative that helps fix the problems and mistakes often seen in human scoring (Williamson et al., 2012). Over the years, AES systems have improved a lot. They started with basic rule-based methods and have now advanced to complex deep learning models, thanks to better technology. These technological improvements make scoring essays better and help students improve their writing skills (Shermis & Burstein, 2013; Wilson & Czik, 2016). Computerized scoring is especially useful for multiple-choice and short-answer questions as it reduces human errors and makes grades more consistent (Bennett, 2018). Advances in natural language processing also allow AES systems to judge complex written statements more effectively (Shermis & Burstein, 2013). These technologies save time for teachers, letting them focus more on creating better teaching methods and giving personalized feedback.

Statistical Models Employed in AES

In the initial versions of the ATS, responses were assessed based on established linguistic and syntactic criteria. For example, Project Essay Grade (PEG), created by Ellis Page in the 1960s, utilized statistical connections between textual features and human-assigned evaluations (Page, 1966). Despite being groundbreaking for their time, rule-based systems were limited by their rigidity and lack of generalizability across many response categories. With the emergence of machine learning, ATS systems evolved from rule-

based frameworks to data-driven models. The implementation of algorithms like support vector machines (SVMs) and decision trees facilitated a more dynamic examination of text features, enhancing accuracy and adaptability (Shermis & Hamner, 2012). These algorithms, trained on extensive datasets, had the ability to emulate human grading patterns with significant accuracy. Recent improvements in natural language processing (NLP), especially the advent of transformer models, have significantly enhanced the capabilities of AES (Kenton & Toutanova, 2019). These models exhibit remarkable contextual understanding, making them ideal for complex tasks like essay evaluation, where coherence, argumentation, and semantic nuances are critical.

Several AES systems utilize NLP approaches, allowing machines to comprehend and assess text-based responses. Key NLP approaches encompass the essential process of feature extraction, which is vital for any text analysis. The recognition of language characteristics, encompassing grammar, syntax, and vocabulary. Semantic analysis is the evaluation of the meaning and context of a specific text. The assessment of coherence, relevance, and content quality is performed using analytical instruments like Latent Semantic Analysis (LSA) (Landauer et al., 1998). NLP-based systems, exemplified as ETS's e-rater, have been extensively utilized in standardized testing, hence showcasing the scalability and dependability of ATS in large-scale evaluations (Attali, 2015).

Statistical models for automatic essay scoring use machine learning and natural language processing to evaluate essay quality. These models work by comparing essays to scoring guides or examples rated by experts. The main methods include regression models, classification models, and combinations of the two. These use both statistics and language analysis. Some models focus on using different types of regression, which are clear ways to score essays automatically. The process first extracts features like how many words there are, how difficult the sentences are, and the variety of vocabulary used. Then it uses regression to predict scores. Logistic regression can also sort essays into different score levels based on a specific grading system. Latent Semantic Analysis (LSA) is a method that finds patterns by looking at how often words appear together in a large set of texts. It checks how similar an essay is to pre-evaluated reference texts. LSA was crucial in early systems for automatic essay scoring, such as the Intelligent Essay Assessor developed in 1998 by Landauer and colleagues.

Bayesian models help in understanding the probabilities of different characteristics in essays, especially when the data is unclear or limited. They show how aspects like grammar and coherence are connected in the essays (Shermis & Burstein, 2013). Support Vector Machines (SVMs) are another method used to classify essays into different score categories using supervised learning. They work by finding the best way to separate essays with different scores through analyzing many features at once (Attali & Burstein, 2006). Random forests and decision trees are models that use ensemble learning, which means using many decision trees together to decide on a score. Random forests can often handle complex relationships in essay characteristics better than simpler models (Rudner & Liang, 2002). Lastly, deep learning methods, which are not only statistical, can find complicated patterns in text. This helps improve tasks that need a strong understanding of the context (Taghipour & Ng, 2016).

Aim of the Study

This paper aims to provide a concise history of essay scoring evolution and to present and discuss a recent study that integrates hierarchical rater models into automated essay scoring systems.

RESEARCH METHOD

This brief scope review synthesizes the literature on the evolution of HRMs and their adaption in automated essay scoring (AES) systems. A thorough examination of academic databases, such as Scopus, Web of Science, and Google Scholar, was performed to locate pertinent peer-reviewed publications, conference proceedings, and technical reports. The search approach entailed the using of keywords such as "hierarchical rater models," "automated essay scoring," "rater effects," and "essay assessment." The search was not confined to works published within any time frame, consistent with the research objectives. The evaluation encompassed only studies that were major milestones in the development and contributes to the

new uses of HRMs. Moreover, the research by Fink and his friends (2024) was included into this review as it represents latest usage of HRM in the field of educational assessment an it is the motivating article for the author to start to this review. All in all, compatiple with this purpose, seven article were presented here for the review (backgroud of HRMs, development of HRMs, integration of HRM with Bayesian Theory, development of HRMs for using with polytomous data, evolving of HRM for longititunal data, the integration of HRMs with Signal Detection Theory, and, finally, the use of HRM for automated testing).

FINDINGS

This section begins with an analysis of the traditional application of HRMs and milestones in the model's development over the years. The incorporation of HRM into the accounting enterprise system (AES) will be discussed.

The emergence of data-driven approaches in education has highlighted the need for the creation of strong statistical models to evaluate performance and guide decision-making. Hierarchical Rater Models (HRMs), which include both rater effects and examinee factors, are particularly advantageous for improving educational evaluations. By addressing rater biases and inconsistencies, these models promote a more refined comprehension of performance data, thus improving the reliability and equity of educational evaluations (Patz et al., 2002; Wind et al., 2017).

HRMs consist of three primary components. The primary element is rater impacts, which include severity, leniency, and inconsistency. These factors facilitate the identification of systematic biases and the evaluation of rater dependability (Engelhard, 2013). The second element is the complexity of the task. Due to the inherent heterogeneity of the tasks, it is essential to make changes to provide a fair and impartial comparison (Wilson & Adams, 1995). The interaction effects constitute the last element. This component facilitates the modeling of both evaluators and tasks. Consequently, impartial estimations can be performed (Myford & Wolfe, 2003).

The HRM framework has long been used in education. For example, it plays an important role in big tests like the NAEP and PISA, making sure that the scoring is fair. HRM also provides valuable information on how well raters do their jobs. This information helps create effective training programs for raters, offering constructive feedback so they can get better at their tasks (Eckes, 2019). It's also crucial for international research, as it allows the study of cultural differences in how evaluators perceive things. This helps ensure that scores are comparable across different groups (Van de Vijver & Leung, 1997). Lastly, HRM is used to analyze how evaluators score essays and evaluate teachers.

A significant domain where HRMs exhibit substantial potential is in the improvement of formative and summative evaluations. Traditional methods often fail to effectively address rater effects, leading to biased assessments of student performance (Myford & Wolfe, 2003). HRMs can quantify and correct for these impacts, therefore yielding more precise estimations of student ability. Future implementations may incorporate HRMs into automated grading systems, facilitating real-time adjustment of rater biases and assuring equity in large-scale evaluations.

Patz et al. (2002) established the hierarchical rater model (HRM) to analyze polytomously scored item response data, tackling concerns of rater variability and consensus. This model has proved essential in extensive educational evaluations, providing a systematic method for analyzing item and rater effects (Patz et al., 2002). Subsequently, Wang and his collaguages (2021) introduced a new variation of HRM that specifically included rater centrality, thus broadening the established aspects of severity and consistency. The simulations conducted by them demonstrate the need of integrating rater centrality into the model to improve its fit.

Choi (2013) developed a model that brings together generalizability theory and item response theory to reduce bias from people who rate assessments. This model helps make assessments more accurate. Research, including works by Zupanc & Štrumbelj (2018) and Choi (2013), shows that hierarchical rater models (HRMs) are effective in reducing biases by raters and making evaluations more reliable across different situations. HRMs have improved to consider data collected over longer periods, which helps manage biases from raters as time passes. Casabianca et al. (2017) made significant progress with the development of a longitudinal hierarchical rater model (L-HRM). This model is used to estimate hidden traits from psychological tests that are rated by people over various intervals. It helps to reduce biases and inconsistencies among raters, providing more accurate assessments of the traits being measured.

There are two significant fields of study that have contributed to the development of HRMs. The first is multilevel modeling. This approach handles data that is layered, like when several raters evaluate students. Important works by Goldstein (1987) and Raudenbush and Bryk (2002) have established the basis for hierarchical linear models (HLMs), which are foundational for HRMs. These models break down variations into parts related to individuals, raters, and tasks. The second field is Item Response Theory (IRT), which assesses the probability of a correct response based on question characteristics and the latent traits of the person answering (Lord & Novick, 1968). Recent developments, such as the Many-Facet Rasch Model (Eckes, 2023), add more complexity by considering additional aspects of raters and tasks, thus aiding in the advancement of HRMs. Additionally, new advancements in Bayesian statistics have further improved the design of HRMs. Bayesian methods are adaptable in managing complex data structures and using prior information (Gelman et al., 2013), making them particularly useful when dealing with small sample sizes.

Nonetheless, it is only beneficial within the realm of educational study. The HRM technique has had three major phases of development over time. The adaptation of HRM to incorporate polytomous responses and multidimensional constructs (Patz et al, 2002) made it a feasible choice for diverse environments and various jobs. And brings flexibility to it. Moreover, the integration of HRMs with machine learning techniques enables the use of HRM in automated systems that can provide rater feedback to participants (Yang et al., 2020). This development presents the opportunity for HRM to be utilized in the automated evaluation of essay questions within an educational framework.

In their paper, Zupanc and Štrumbelj (2018) introduced a Bayesian hierarchical latent trait model to estimate rater bias and reliability in diverse performance assessment conditions. The examination of essay rating data collected ove five years indicated that the model accurately identified rater effects, showing that rater unreliability exerted a greater influence on final grades than rater bias.

Another article which introduce Hierarchical Rater Thresholds Model (HRTM), distinctly differentiates rater effects from item effects on the threshold parameters of categorical observable variables. This model is better from the prior HRMs and facilitates parameter estimation by Weighted Least Squares, thus improving computational efficiency and compatibility with conventional latent variable modeling software (De Gruyter, 2020).

Incorporating Human Rater Models (HRMs) into test creation becomes more effective with the use of Signal Detection Theory (SDT). Originating from psychophysics and decision theory, SDT aids understanding decisions in uncertain conditions. In rater evaluations, SDT views each decision as a balance between a "signal" (such as essay quality) and "noise" (like differences among raters and environmental factors). The key components of SDT are sensitivity, which measures how well changes in essay quality are detected, and bias, which reflects whether raters tend to give higher or lower scores. These elements are applied to understand rater performance better (Hautus et al., 2021). Bringing SDT and HRMs together capitalizes on

each method's strengths. HRMs are adept at illustrating individual rater effects, while SDT offers detailed insights into decision-making. This combination addresses many issues found in traditional Automated Essay Scoring (AES) models. SDT provides a thorough understanding of how raters judge essays, enhancing HRMs' ability to model rater effects (Hautus et al., 2021). The integrated approach pays particular attention to rater sensitivity and bias, helping to reduce unfairness and encouraging fairer scoring (Patz et al., 2002). Because SDT features like sensitivity and bias are straightforward to interpret, those involved in scoring can make better, more informed decisions (Gelman et al., 2013). This makes the scoring process clearer and promotes fairness, leading to improved outcomes.

A recently published study proposed a strategy for aligning (HRM with the automated essay scoring AES. Fink and his colleagues (2024) have proposed a new hierarchical rater model-based method to integrate predictions coming from various AES models, considering their differing scoring patterns. The aim of this strategy is to accumulate the strengths of specific models to improve the overall accuracy of scoring. The suggested approach was assessed utilizing data from a university essay-writing exam. The results showed that the integrated model had accuracy akin to the most efficient individual AES model. Additionally, the strategy lower the degree of differential item functioning (DIF) between human scoring and automated scoring. In this way It enhance measurement invariance. The authors stated that the successful accumulation of various AES models using HRM approach improves the reliability and validity of AES. This advancement presents the opportunity for enhanced efficiency and precision in AES processes in educational settings. This research significantly contribute to the domain of educational measurement via the a fore mentioned facts.

DISCUSSION AND CONCLUSION

Adding HRMs into the AES process marks a significant advancement in educational evaluation. HRMs address long-standing problems in essay grading including rater bias, inconsistency, and the subjectivity inherent in human judgment. HRMs provide a methodical framework that enhances the dependability and equity of assessments by clearly simulating evaluator behaviors and task problems (Engelhard, 2013; Myford & Wolfe, 2003).

HRMs are really skilled at noticing how people rate things, whether they are too strict, too lenient, or fair (Engelhard, 2013). These different rating styles can often create regular issues in Automated Essay Scoring (AES) systems, which in turn makes assessments less reliable and fair. While AES systems can process a lot of data, they aren't as capable as HRMs in spotting these biased rating patterns (Patz et al., 2002). Using HRMs alongside AES is very important, especially during significant decisions about students. HRMs also play a key role in showing how evaluators behave, ensuring they adhere to the established grading rules. This helps both automated systems and human evaluators remain consistent (Eckes, 2019).

Recent improvements in the HRM have made it more useful and efficient. By incorporating Signal Detection Theory (SDT) into HRM, we gain a better understanding of decision-making processes. SDT uses sensitivity and bias measurements to evaluate how both humans and automated systems assess essay quality. These measurements are crucial for ensuring fairness, resolving issues with automated Essay Scoring (AES) systems, and enhancing the accuracy of scoring models. Additionally, SDT increases transparency in HRM, making results easier to comprehend. This transparency is vital as more people depend on deep learning algorithms, which are often criticized for their "black box" nature, where the processes are not visible.

HRMs also excel when working with complex data. Recent advancements, such as integrating different response types and multiple factors into HRMs, have increased their usefulness in educational settings (Patz

et al., 2002). The use of Bayesian estimation techniques makes HRMs more scalable and adaptable, allowing them to work even when there is limited data or when traditional models can't be used due to assumptions (Gelman et al., 2013). Combining HRMs with machine learning makes it possible to update scoring algorithms and correct biases in real time, representing a significant development in AES. Research by Yang et al. (2023) shows that incorporating HRMs into AES systems significantly improves scoring accuracy.

One major benefit HRMs provide in the field is their great increase in capacity to integrate the results obtained from numerous AES types. Fink et al. (2024) show how HRMs can combine the special benefits of several AES systems by using forecasts from several systems, hence enhancing general score accuracy. This approach reduces discrepancies between human and computerized assessments and increases the accuracy of automated scoring. A key component in preserving fairness among diverse populations who was scored with different modes, the integration of HRMs improve measurement invariance across human scoring and AES. databases in line with this.

Despite progress, major challenges remain in using HRM-based AES systems on a large scale. This is particularly true for big tests that require real-time processing, as HRM systems are complex and hard to manage (Patz et al., 2002). They also need big, high-quality datasets to work well, and if we don't have these, using them effectively becomes more difficult. While HRMs can improve fairness, they can still show bias, though less than older AES models. Binns et al. (2018) observed that HRMs and AES models might be influenced by local language habits, cultural differences, and educational backgrounds. This means these systems might not perform equally well for everyone. To tackle these issues, more research is needed on making machine learning fair and developing better training datasets. Additionally, the complexity of HRMs makes them hard to understand. Although they offer useful insights into rating and decision-making, their complexity and lack of user-friendly software make them hard for non-experts to use, especially in smaller settings without technical help. For HRMs to gain greater acceptance in educational institutions, they must become simpler and more accessible.

HRMs have still another potential use: their application in formative assessments. Formative assessments have traditionally depended on human evaluators to score and provide comments that advances the student learning (Eckes, 2019). By concentrating on particular areas where students could need development, the inclusion of HRMs into AES systems could help to increase the diagnosis efficacy of these tests. HRMs can give careful assessments of essay elements including grammar, coherence, and argument. This helps to enable more precise and concentrated judgments and interventions in education. Furthermore, the capacity of HRMs to evaluate rater impacts guarantees the accuracy and equity of the feedbacks given by the raters, so promoting a more equitable environment for learning.

Improvements in natural language processing and machine learning will directly affect HRMs' future in automated essay assessment. Using NLP techniques—including semantic analysis and feature extraction—shows a considerable advancement in AES. With HRMs, these techniques could show notable increases in score accuracy and precision (Landauer et al., 1998). Deep learning models combined with HRMs show even more promise for spotting the intricate and complex elements of literature, including tone and audience awareness, that modern techniques still cannot assess with desired quality (Kenton & Toutanova, 2019).

Including HRMs into automated essay scoring systems marks the most recent advancement in the procedure of AES scoring and helps to construct dependable, objective, and adaptable essay scoring systems. HRMs can raise educational assessments by reducing rater biases, increasing scoring accuracy, and helping to create interpretable artificial intelligence models. Realizing this promise would, however, need addressing issues with computational complexity, dataset availability, and algorithmic bias. Future studies should

concentrate on improving HRM strategies by means of improved algorithms, analysis of their application in many educational environments, and use of natural language processing and machine learning to thus increase their capabilities. By use of modern AES approaches, the integration of HRMs helps teachers to produce improved learning results while preserving fair and consistent assessments.

Suggestions

- As known, in Turkey, automated scoring has not been used other than research purposes. It could be suggestted that, the government and private initiatives invest to the technological infrastsucture for automated scoring applications in schools. In addition, should be generalized to be used diverse testing conditions.

REFERENCES

Amabile, T. M. (1996). *Creativity in context*. Westview Press.

Anderson, L.W., Krathwohl, D.R., Airasian, P.W., Cruickshank, K.A., Mayer, R.E., Pintrich, P.R., … Wittrock, M.C. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives*. (Complete edition). Longman.

Andrade, H., & Du, Y. (2007). Student responses to criteria-referenced self-assessment. *Assessment and Evaluation in Higher Education, 32*(2), 159–181.

Attali, Y. (2015). Reliability-based feature weighting for automated essay scoring. *Applied Psychological Measurement*, *39*(4), 303-313.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3), 3–30.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54-74.

Bennett, R. E. (2018). Educational assessment: Tests and measurements in the age of accountability. *Educational Measurement: Issues and Practice, 37*(2), 21–28.

Biggs, J., & Tang, C. (2011). *Teaching for quality learning at university* (4th ed.). Open University Press.

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2018). *'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions.* Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 377.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans, Green.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1999). Assessing writing skill. *Review of Educational Research, 69*(2), 125–130.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Brown, G., & Knight, P. (1994). *Assessing learners in higher education*. Kogan Page.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine, 25*(3), 27–36.

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2017). Hierarchical rater models for longitudinal data. *Journal of Educational and Behavioral Statistics, 42*(3), 230–250.

Choi, J. (2013). *Advances in combining generalizability theory and item response theory* [Doctoral dissertation]. UC Berkeley.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated language assessment.

In *Quantitative data analysis for language assessment volume I* (pp. 153-175). Routledge.

Eckes, T. (2023). *Introduction to many-facet Rasch measurement*. Peter Lang.

Elbow, P. (1998). *Writing with power: Techniques for mastering the writing process*. Oxford University Press.

Engelhard, G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. Routledge.

Entwistle, N., & Ramsden, P. (1983). *Understanding student learning*. Routledge.

Fink, M., Ziegler, F., & Lee, H. (2024). Integrating automated scoring models into hierarchical frameworks. *Journal of Educational Measurement, 61*(2), 101–115.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). CRC Press.

Goldstein, H. (1987). *Multilevel models in educational and social research*. Oxford University Press.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hautus, M. J., Macmillan, N. A., & Creelman, C. D. (2021). *Detection theory: A user's guide*. Routledge.

Hyland, K. (2003). *Second language writing*. Cambridge University Press.

Kenton, J. D. M. W. C., & Toutanova, L. K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT* (Vol. 1, p. 2).

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259–284.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang.

Marton, F., & Säljö, R. (1976). On qualitative differences in learning: Outcome and process. *British Journal of Educational Psychology, 46*(1), 4–11.

Meadows, M. and Billington, L. (2005). *A review of the literature on marking reliability*. Report for the National Assessment Agency by AQA Centre for Education Research and Policy.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48,* 238–243.

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341-384.

Poe, M., & Elliot, N. (2019). Evidence of fairness: Twenty-five years of research in assessing writing. *Assessing Writing*, *42*, 100418.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping E-Rater: Challenging the validity of automated essay scoring. *ETS Research Report Series*, *2001*(1), i-44.

Ramsden, P., & Moses, I. (1992). Associations between research and teaching in Australian higher education. *Higher Education*, *23*(3), 273-295.

Raudenbush, S. W. (2002). Hierarchical linear models: Applications and data analysis methods. *Advanced Quantitative Techniques in the Social Sciences Series/SAGE*.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, *1*(2).

Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*(7), 807-826.

Şata, M., & Karakaya, I. (2022). Investigating the impact of rater training on rater errors in the process of assessing writing skill. *International Journal of Assessment Tools in Education*, *9*(2), 492-514.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. *Language Testing*, *22*(1), 1-30.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(7), 4-14.

Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation*. Routledge.

Shermis, M. D., & Hamner, B. (2013). Contrasting state-of-the-art automated scoring of essays. In *Handbook of automated essay evaluation* (pp. 313-346). Routledge.

Simkin, M. G., & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection?. *Decision Sciences Journal of Innovative Education*, *3*(1), 73-98.

Taghipour, K., & Ng, H. T. (2016, November). *A neural approach to automated essay scoring*. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1882-1891).

Van de Vijver, F., & Leung, K. (1997). Methods and data analysis of comparative research. *Handbook of Cross-Cultural Psychology*, *1*, 257-300.

Wang, S., Beheshti, A., Wang, Y., Lu, J., Sheng, Q. Z., Elbourn, S., ... & Galanis, E. (2021, June). *Assessment2Vec: learning distributed representations of assessments to reduce marking workload.* In International Conference on Artificial Intelligence in Education (pp. 384-389). Springer.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*, 2–13.

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English Language Arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, *100*, 94-109.

Wind, S. A., Stager, C., & Patil, Y. J. (2017). Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments. *Assessing Writing*, *34*, 1-15.

Yang, H., He, Y., Bu, X., Xu, H., & Guo, W. (2023). Automatic essay evaluation technologies in Chinese writing—A systematic literature review. *Applied Sciences*, *13*(19), 10737.

Zupanc, K., & Štrumbelj, E. (2018). A Bayesian hierarchical latent trait model for estimating rater bias and reliability in large-scale performance assessment. *Plos one*, *13*(4), e0195297.