Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309-6575

2024

Cilt: 15 - Sayı: Özel Sayı
Volume: 15 - Issue: Special Issue

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi

Journal of Measurement and Evaluation in Education and Psychology

ISSN: 1309 – 6575

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (EPOD) yılda dört kez yayımlanan hakemli uluslararası bir dergidir. Yayımlanan yazıların tüm sorumluğu ilgili yazarlara aittir.

Journal of Measurement and Evaluation in Education and Psychology (JMEEP) is an international refereed journal published four times a year. Authors are fully responsible for the content of their published articles.

Dizinleme / Abstracting & Indexing

Emerging Sources Citation Index (ESCI), DOAJ (Directory of Open Access Journals), SCOPUS, TÜBİTAK TR DIZIN Sosyal ve Beşeri Bilimler Veri Tabanı (ULAKBİM), Tei (Türk Eğitim İndeksi), EBSCO

Gülşen TAŞDELEN TEKER (Hacettepe Üni.)
Hakan KOĞAR (Akdeniz Üni.)
Hakan SARIÇAM (Dumlupınar Üni.)
Hakan Yavuz ATAR (Gazi Üni.)
Halil İbrahim SARI (Kilis Üni.)
Halil YURDUGÜL (Hacettepe Üni.)
Hatice Çiğdem BULUT (Northern Alberta IT)
Hatice KUMANDAŞ (Artvin Çoruh Üni.)
Hikmet ŞEVGİN (Van Yüzüncü Yıl Üni.)
Hülya KELECİOĞLU (Hacettepe Üni.)
Hülya YÜREKLI (Yıldız Teknik Üni.)
İbrahim Alper KÖSE (Bolu Abant İzzet Baysal Üni.)
İbrahim YILDIRIM (Gaziantep Üni.)
İbrahim UYSAL (Bolu Abant İzzet Baysal Üni.)
İlhan KOYUNCU (Adıyaman Üni.)
İlkay AŞKIN TEKKOL (Kastamonu Üni.)
İlker KALENDER (Bilkent Üni.)
İsmail KARAKAYA (Gazi Üni.)
İzettin Aydoğan (Van Yüzüncü Yıl Üni.)
Kadriye Belgin DEMİRUS (Başkent Üni.)
Kimeshia MYLES (US Air Force Command)
Kübra ATALAY KABASAKAL (Hacettepe Üni.)
Levent ERTUNA (Sakarya Üni.)
Levent YAKAR (Kahramanmaraş Sütçü İmam Üni.)
Mahmut Sami KOYUNCU (Afyon Üni.)
Mahmut Sami YİĞİTER (Ankara Sosyal B. Üniv.)
Mehmet KAPLAN (MEB)
Mehmet ŞATA (Ağrı İbrahim Çeçen Üni.)
Melek Gülşah ŞAHİN (Gazi Üni.)
Meltem ACAR GÜVENDİR (Trakya Üni.)
Meltem YURTÇU (İnönü Üni.)
Merve ŞAHİN KÜRŞAD (TED Üni.)
Metin BULUŞ (Adıyaman Üni.)
Mitchell CLARKE (WestEd)
Murat Doğan ŞAHİN (Anadolu Üni.)
Mustafa ASİL (University of Otago)
Mustafa İLHAN (Dicle Üni.)
Nagihan BOZTUNÇ ÖZTÜRK (Hacettepe Üni.)
Nail YILDIRIM (Kahramanmaraş Sütçü İmam Üni.)
Neşe GÜLER (İzmir Demokrasi Üni.)
Neşe ÖZTÜRK GÜBEŞ (Mehmet Akif Ersoy Üni.)
Nuri DOĞAN (Hacettepe Üni.)
Nükhet DEMİRTAŞLI (Emekli Öğretim Üyesi)
Okan BULUT (University of Alberta)
Onur ÖZMEN (TED Üniversitesi)
Ömer KUTLU (Ankara Üni.)
Ömür Kaya KALKAN (Pamukkale Üni.)
Osman TAT (Van Yüzüncü Yıl Üni.)
Önder SÜNBÜL (Mersin Üni.)
Özen YILDIRIM (Pamukkale Üni.)
Özge ALTINTAS (Ankara Üni.)
Özge BIKMAZ BİLGEN (Adnan Menderes Üni.)
Özlem ULAŞ (Giresun Üni.)
Paul EDELBLUT (Vantage Labs)
Recep GÜR (Erzincan Üni.)
Ragıp TERZİ (Harran Üni.)
Richard AMOAKO (Uni. of Tennessee Knoxville)
Sedat ŞEN (Harran Üni.)
Randy E. BENNETT (Educational Testing Service)
Recep Serkan ARIK (Dumlupınar Üni.)
Safiye BİLİCAN DEMİR (Kocaeli Üni.)
Selahattin GELBAL (Hacettepe Üni.)
Seher YALÇIN (Ankara Üni.)
Selen DEMİRTAŞ ZORBAZ (Ordu Üni.)
Selma ŞENEL (Balıkesir Üni.)
Seçil ÖMÜR SÜNBÜL (Mersin Üni.)
Sait ÇÜM (MEB)
Sakine GÖÇER ŞAHİN (University of Wisc. Mad.)
Sebahat GÖREN (Hacettepe Üni.)
Sedat ŞEN (Harran Üni.)
Sema SULAK (Bartın Üni.)
Semirhan GÖKÇE (Niğde Ömer Halisdemir Üni.)
Serap BÜYÜKKIDIK (Sinop Üni.)
Serkan ARIKAN (Boğaziçi Üni.)
Seval KIZILDAĞ ŞAHİN (Adıyaman Üni.)
Sevda ÇETİN (Hacettepe Üni.)
Sevilay KİLMEN (Abant İzzet Baysal Üni.)
Sinem DEMİRKOL (Ordu Üni.)
Sinem Evin AKBAY (Mersin Üni.)
Sungur GÜREL (Siirt Üni.)
Süleyman DEMİR (Sakarya Üni.)
Sümeyra SOYSAL (Necmettin Erbakan Üni.)
Şeref TAN (Gazi Üni.)
Şeyma UYAR (Mehmet Akif Ersoy Üni.)
Tahsin Oğuz BAŞOKÇU (Ege Üni.)
Terry A. ACKERMAN (University of Iowa)
Tuğba KARADAVUT (İzmir Demokrasi Üni.)
Tuncay ÖĞRETMEN (Ege Üni.)
Tülin ACAR (Parantez Eğitim)
Türkan DOĞAN (Hacettepe Üni.)
Ufuk AKBAŞ (Hasan Kalyoncu Üni.)
Ümmügül BEZİRHAN (Boston College)
Wenchao MA (University of Alabama)
Yavuz AKPINAR (Boğaziçi Üni.)
Yeşim ÖZER ÖZKAN (Gaziantep Üni.)
Yusuf KARA (Southern Methodist University)
Zekeriya NARTGÜN (Bolu Abant İzzet Baysal Üni.)
Zeynep ŞEN AKÇAY (Hacettepe Üni.)

*Ada göre alfabetik sıralanmıştır. / Names listed in alphabetical order.

Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi (Aralık 2024, Sayı: 15-Özel Sayı)

Journal of Measurement and Evaluation in Education and Psychology (December 2024, Issue: 15-Special Issue)

İÇİNDEKİLER / CONTENTS

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Editorial; 2024; 15(Special issue); 260-262

EDITORIAL

Opportunities and Challenges of AI in Educational Assessment

Alper SAHIN* **Nathan THOMPSON**** **Kadriye ERCIKAN*****

Abstract

In the past few years, as artificial intelligence (AI) and large language models (LLM) have rapidly entered our lives, we have witnessed groundbreaking innovations across numerous fields. The rapid pace of these changes has been met with excitement by some and apprehension by others. However, we all agree that they have made tremendous contributions so far and their contributions in the future will reshape our existence. The field of educational assessment is no exception. With this in mind, we issued a call for a special issue themed "Opportunities and Challenges of AI in Educational Assessment." which finally included seven distinguished articles on subthemes of fair and responsible use of AI in educational assessment, learning analytics, automated scoring, and real-life examples of AI and LLM.

Keywords: AI in Educational assessment, Large language models, fair and responsible use of AI and LLM, learning analytics, automated scoring, real-life examples of AI and LLM

For this special issue, we received 23 high-quality article proposals from seven different countries: Canada (1), Germany (1), Türkiye (1), South Africa (4), South Korea (1), the United Kingdom (1) and the United States (12). Of these proposals, we invited 12 authors to submit their articles for the special issue. Ultimately, we received 14 article submissions. After a rigorous blind review procedure and revisions, we are pleased to present seven meticulously selected articles in our special issue.

In line with our theme, our special issue begins with a systematic review, by Sato et al. (2024), of literature related to the fair and responsible use of artificial intelligence in educational assessment. In their article titled "Putting AI in Fair: A Framework for Equity in AI-driven Learner Models for Accessible and Inclusive Assessment", Sato et al. (2024) present us with an extensive literature review covering theoretical, empirical, ethical, and policy documents addressing the role of learner models in K-12 assessment. The authors sought answers to 5 important research questions regarding whether and how these models were used to promote accessibility and inclusivity, and they propose a framework that aspires to influence the equitable and valid of assessment of all students.

Following the opening article in the introduction section, the second section of our special issue includes two valuable articles by Guo et al. (2024) and Cavus & Kuzilek (2024) on *Learning Analytics*, which has the potential to directly impact the quality of education. In their very interesting and useful study titled "*Human-Centered AI for Discovering Student Engagement Profiles on Large-Scale Educational Assessments*", Guo et al. (2024) propose an artificial intelligence-supported model that combines response and process data to better reflect students' knowledge and test-taking processes using multi-source data to reveal their engagement profiles. We strongly recommend that you read this article where they described and tested this model, which is a first in the field that will allow educators to access deeper and more useful information about their students' test performances and knowledge levels. In their article titled "*An Effect Analysis of the Balancing Techniques on the Counterfactual Explanations*

* Assist. Prof. Dr., Department of Basic English, Atilim University, Ankara, Türkiye, alpersahin2@yahoo.com, ORCID ID: 0000-0001-7750-4408

** CEO, Assessment Systems Corporation, Minneapolis, MN, US, nthompson@assess.com, ORCID ID: 0000-0002-3981-7881

*** SVP of Global Research, ETS, Princeton, NJ, US, kercikan@ets.org, ORCID ID: 0000-0001-8056-9165

To cite this article:
Sahin, A., Thompson, N., & Ercikan, K. (2024). Opportunities and challenges of AI in educational assessment. *Journal of Measurement and Evaluation in Education and Psychology, 15,*(Special issue), 260-262. https://doi.org/10.21031/epod.1607441

of Student Success Prediction Models", Cavus & Kuzilek (2024) investigated the effectiveness and feasibility of using various counterfactual explanations to predict students' success to be better understood by students and parents and to increase their trust to these predictions. We believe you will enjoy this article.

The third section of our special issue includes studies by Chan et al. (2024) and Mo Zhang et al. (2024), who have undertaken two important studies on *Automated Scoring*, where AI and LLM have been widely used for a long time. Chan et al. (2024), in their study titled "*Integrating Metadiscourse Analysis with Transformer-Based Models for Enhancing Construct Representation and Discourse Competence Assessment in L2 Writing: A Systemic Multidisciplinary Approach*", address the important but somewhat neglected topic of discourse competence in Automated Essay Scoring. While doing this, the authors, who use Metadiscourse markers (MDM), test what can be done to expand the ability of automated scoring models to identify and classify MDM with 2000 texts at different CEFR levels and share their findings with us, providing a foundation for future research to expand the construct of L2 automated scoring models. In their article titled "*Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness*", Mo Zhang et al. (2024) investigated the effects of different sampling methods on the ability of AI to predict the scores given by human raters, and of the stratified sampling method on the fairness of AI's prediction ability, together with other methods. We strongly recommend that you read this article, which is one of the novel studies conducted with newly developed language models and yielded interesting findings.

Finally, in the fourth and final section of our special issue, there are two valuable studies by Bolender et al. (2024) and Ting Zhang et al. (2024), which evaluated the performance of *real-life examples of AI and LLM* that reached the end user. In their article titled "*Generative AI in K12: Analytics From Early Adoption*", Bolender et al. conducted three comprehensive case studies that included real-life use of Finetune's Finetune Generate, developed for item development using natural language models, and Finetune Catalog, developed to tag and align educational content to various standards and frameworks. Ting Zhang et al. (2024) contributed to our special issue with a multi-disciplinary study titled "*Ask NAEP: A Generative AI Assistant for Querying Assessment Information*", which, as the name suggests, includes an evaluation of the performance of the Ask NAEP chatbot, which was developed to provide accurate and comprehensive answers using publicly available National Assessment of Educational Progress (NAEP) data. You will not regret reading it.

All in all, we hope that you will read the articles in this special issue with pleasure and that this special issue will contribute significantly to the field of education assessment. On behalf of all researchers who conduct studies on educational assessment, we would like to thank the authors who have shown interest in our special issue, supported us with their article proposals and articles, and contributed to the publication of this magnificent special issue.

We would also like to express our gratitude to the journal editors and to our expert reviewers who have supported us so that the blind review process of this special issue ran smoothly and without any problems, to the layout editors, and to the journal's editorial team who have worked hard to prepare the articles in the special issue for publication.

References

Bolender, B., Vispoel, S., Converse, G., Koprowicz, N., et al. (2024). Generative AI in K12: Analytics From Early Adoption. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 361-377. https://doi.org/10.21031/epod.1539710

Cavus, M., & Kuzilek, J. (2024). An Effect Analysis of the Balancing Techniques on the Counterfactual Explanations of Student Success Prediction Models. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 302-317. https://doi.org/10.21031/epod.1526704

Chan, S., Sathyamurthy, M., Inoue, C., Bax M., et al. (2024). Integrating Metadiscourse Analysis with Transformer-Based Models for Enhancing Construct Representation and Discourse Competence Assessment in L2 Writing: A Systemic Multidisciplinary Approach. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 318-347. https://doi.org/10.21031/epod.1531269

Guo, H., Johnson, M., Saldivia, L., Worthington, M., et al. (2024). Human-Centered AI for Discovering Student Engagement Profiles on Large-Scale Educational Assessments. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 282-301. https://doi.org/10.21031/epod.1532846

Sato, E., Shyyan, V., Chauhan, S., Christensen, L. (2024). Putting AI in Fair: A Framework for Equity in AI-driven Learner Models and Inclusive Assessments. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 263-381. https://doi.org/10.21031/epod.1526527

Zhang, M., Johnson, M., & Ruan, C. (2024). Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 348-360. https://doi.org/10.21031/epod.1561580

Zhang, T., Patterson, L., Beiting-Parrish, M., Webb, B., et al. (2024). Ask NAEP: A Generative AI Assistant for Querying Assessment Information. *Journal of Measurement and Evaluation in Education and Psychology*, *15*(Special issue), 378-394. https://doi.org/10.21031/epod.1548128

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(Special issue); 263-281

Putting AI in Fair: A Framework for Equity in AI-driven Learner Models and Inclusive Assessments

Edynn SATO* Vitaliy SHYYAN** Swati CHAUHAN*** Laurene CHRISTENSEN****

Abstract

This paper delves into the critical role of learner models in educational assessment and includes a systematic review of recent literature on AI and K-12 education. This review brings to light gaps and opportunities in current practices and serves as a foundation for the Fair AI Framework, which centers on fairness and transformative justice, and aspires to influence AI applications to ensure they are inclusive of diverse learners. This paper concludes with a recommended path forward that underscores the critical importance of learner models in accessible, inclusive, equitable, and valid assessment for all learners.

Keywords: *artificial intelligence, K-12 education, assessment, validity, framework, equity, social justice, accessiblity, inclusion, students with disabilities, cultural diversity, linguistic diversity, English learners, policy, research, ethics*

Introduction

The field of educational measurement is experiencing significant advancements in methods and technologies, particularly through the integration of innovative tools that incorporate Artificial Intelligence (AI). These developments aim to create more efficient, personalized, and accurate evaluations of learning. This paper explores the implications of such advancements, focusing on AI-driven learner models and their potential to transform educational assessment practices within the U.S. Kindergarten through Grade 12 (K-12) assessment context. More specifically, this paper introduces a validity framework that centers fairness and transformative justice, addressing the critical need for equitable AI applications that are inclusive of students with disabilities, culturally and linguistically diverse students, and other currently and historically systemically marginalized and underserved student groups. The authors assert that learner models are fundamental to educational assessment and require meticulous consideration to ensure inclusivity and equity. Learner models reflect our understanding of learner characteristics in terms of how learners represent information and develop competence, and these models shape our definition of what is measured (constructs) as well as the criteria for evaluating demonstrations of knowledge, skills, and abilities (Mislevy, 2004; Pellegrino et al., 2001; Sato, 2024).

The first part of this paper delves into the critical role of learner models in educational assessment and includes a review of recent literature on AI and K-12 education. This review brings to light gaps and opportunities in current practices and serves as a foundation for the framework proposed in the second part of this paper, which centers on fairness and transformative justice. The framework aspires to influence AI applications to ensure they are inclusive of students with disabilities, culturally and linguistically diverse students, and other currently and historically systemically marginalized and underserved student groups. This paper concludes with a recommended path forward that underscores

* Dr., Sato Education Consulting LLC, San Francisco-US, e-mail: edynn@satoeducationconsulting.com, ORCID ID: 0000-0002-1706-6263
** Dr., WIDA at the University of Wisconsin, Madison-US, e-mail: vshyyan@wisc.edu, ORCID ID: 0009-0006-5262-3180
*** Ms., WIDA at the University of Wisconsin, Madison-US, e-mail: svchauhan@wisc.edu, ORCID ID: 0009-0001-1257-6328
**** Dr., WIDA at the University of Wisconsin, Madison-US, e-mail: llchristens2@wisc.edu, ORCID ID: 0000-0002-2765-1810

To cite this article:
Sato, E., Shyyan, V., Chauhan, S. & Christensen, L. (2024). Putting AI in fair: Toward a validity framework centering fairness and transformative justice in AI-driven learner models for accessible and inclusive assessment. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 263-281. https://doi.org/10.21031/epod.1526527

Received: 1.08.2024
Accepted: 23.11.2024

the critical importance of learner models in accessible, inclusive, equitable, and valid assessment for all learners.

The Essential Role of Learner Models in Valid Educational Assessment

An assessment cannot be designed and implemented and will not yield valid interpretations of student knowledge without appropriate and adequate consideration of a learner model reflective of a student's unique capabilities and needs (Marion & Pellegrino, 2006; Michel & Shyyan, 2024; Mislevy, 2004; Pellegrino, 2003; Pellegrino et al., 2001; Sato, 2024; Shyyan & Christensen, 2018). Without such a model, assessment results will not yield valid interpretations of what students know and can do.

The centrality of learner models for valid assessment is depicted in the assessment triangle (see Figure 1) which is a useful heuristic for examining the qualities and influence of learner models vis-a-vis assessment tasks and evaluative criteria. Learner models, assessment tasks, and evaluative criteria must be in congruence to yield a valid assessment (Marion & Pellegrino, 2006). The components of the assessment triangle heuristic are as follows (Mislevy, 2004; Pellegrino et al., 2001; Sato, 2024):

• Cognition: How information is represented, and competence is developed, including the learning theory and articulation of the knowledge being measured (learner model).

• Observation: How information is elicited, and the types of tasks that would best elicit demonstrations of understanding and knowledge (task model, assessment methods).

• Interpretation: How information is understood, including tools and methods for making sense of observed behaviors/responses (inferences, evaluative criteria).

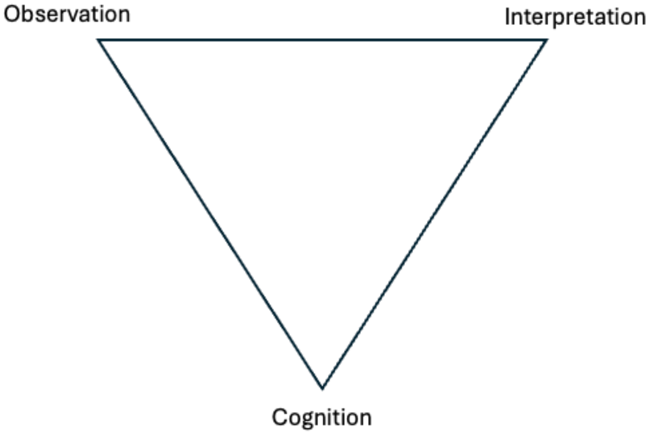


Figure 1. The Assessment Triangle (Pellegrino et al., 2001, p. 44)

With this heuristic in mind, the following discussion focuses on the Cognition vertex and learner models. Given the diverse population of learners in U.S. schools, understanding and implementing effective learner models are imperative for ensuring that all students are assessed accurately and fairly.

The development of learner models integrates theory and research from multiple disciplines including educational psychology and cognitive science, and pedagogy. This process involves extensive data collection and analysis, using both qualitative and quantitative methods, to identify learning patterns and individual variations (McDonald & West, 2021). Theoretical frameworks help interpret data, showing how students engage with content, process information, transfer knowledge, and represent understanding. Leveraging these models, educators can design responsive instruction that enhances access, engagement, understanding, and achievement (Darling-Hammond et al., 2019).

Similar to their role in instructional design, learner models can inform assessment design by providing a detailed map of expected learning progressions, and they can highlight critical considerations about the nature and conditions for students' learning and demonstrations of learning (Sato, 2024). By aligning assessment tasks (Observation vertex) with the learner model, assessment developers can ensure that tasks are appropriately challenging and supportive and accurately measure intended knowledge and skills in a manner appropriate for diverse students, without introducing bias or irrelevant difficulties. Learner models also support the interpretation of assessment outcomes (Interpretation vertex) by providing a framework for understanding student performance in terms of their cognitive and linguistic processes, learning experiences, and individual needs.

The Importance of Accounting for Diversity in Learner Models

Learner models are important because students learn and represent knowledge in various ways. There is a body of research showing that students' experiences and backgrounds affect their meaning making, learning, and representations of knowledge (e.g., de Klerk, 2008; Hall, 1983; Hofstede & Hofstede, 2005; Kulich, 2009; Levine, 1997; Lewis, 2006; Michel & Shyyan, 2024; Nisbett, 2003; Parrish & Linder-VanBerschot, 2010; Pearson & Garavaglia, 2003; Sato, 2017, 2024). Such research explains how students with different backgrounds, when presented with the same information, can have different interpretations of and responses to the information (Hammond, 2015; Ji et al., 2004; Masuda & Nisbett, 2001; Michel & Shyyan, 2024; Sato, 2024; Solano-Flores & Nelson-Barber, 2001; Wang & Leichtman, 2000). Evidence from such research suggests that there are background and experiential factors that are construct relevant and ought to be considered when designing and developing valid and fair assessments (Sato, 2017, 2024).

A mismatch between the expectations of an assessment (task design, administration conditions, evaluation criteria, and interpretations of performance outcomes) and the ways students learn (as shaped by their backgrounds and experiences) undermine assessment validity and can result in misrepresentations or underestimations of student knowledge (Montenegro & Jankowski, 2017). In the U.S. K-12 accountability context, assessments tend to privilege a Western orientation and values which generally reflect analytical and linear or sequential reasoning and typically place value on objectivity and individualism (Preston & Claypool, 2021). To the degree that subgroups of our diverse student population are either unfamiliar with the Western cultural orientation and values or have norms and values that differ, those students potentially may be unable to perform to the best of their abilities on the assessment (eCampusOntario, n.d.; Molle et al., 2015; Sato, 2024; Wexler, 2019, 2021). With more than 10 percent of students identified as English learners and roughly 15 percent of public school students receiving special education or related services under the Individuals with Disabilities Act (IDEA), commonly used assessments in our U.S. K-12 schools may not be accessible to the full range of these more than 12 million students (NCES, 2020, 2023, 2024; Montenegro & Jankowski, 2017). Learners from marginalized backgrounds or with diverse learning styles may be disproportionately affected when assessments do not align with students' ways of learning and understanding, perpetuating inequalities in educational outcomes and opportunities. Moreover, with such lack of alignment, students may feel disengaged, and their motivation and efficacy may be negatively impacted (Ryan & Weinstein, 2009; Usher, 2012). This can have long-term consequences for students' academic trajectories and overall well-being. It is, therefore, critical to develop learner models that reflect the diversity of our K-12 student population -- meticulous consideration of the range of ways students learn and demonstrate their learning is needed to develop sufficiently robust learner models that can support the design, development, and implementation of inclusive, equitable, and valid assessments.

The Promise of AI-Driven Learner Models in Assessment

While effective accessibility and inclusion solutions continue to emerge to support the learning and achievement of K-12 students (Cawthon & Shyyan, 2022; Michel & Shyyan, 2024), the integration of AI technology in education has the potential to significantly advance and transform how we understand and assess learner capabilities. Especially for students who are currently and have historically been systemically marginalized and underserved, AI-driven learner models offer the promise of more personalized, equitable, and inclusive educational experiences.

AI-driven learner models can help to address challenges faced by current K-12 assessments (Holmes et al., 2019; USED, 2023). For example, AI-driven learner models have the potential to support more student-centered assessment for diverse test takers through the analysis of student data, identification of learning patterns, and the leveraging of algorithms to adapt assessment content and format and match them to the capabilities and needs of individual test takers (Li et al., 2018). Accessibility can be enhanced by matching assessment content and formats (e.g., audio or tactile versions, translations, language adaptations to the student grade and age levels) to test taker needs so that each test taker is provided optimal conditions to demonstrate what they know and can do (Holmes et al., 2019; Li et al., 2022). Improving accessibility affects the accuracy of the measures of student knowledge and, subsequently, the validity of the interpretations of what students know and can do. Additionally, AI-driven learner models can help to ensure that assessments are as free from bias as possible and provide fair and equitable opportunities for all students (Grover, 2024). Bias can be mitigated through data analyses and the identification of patterns that indicate bias in assessment items and scoring algorithms, thereby supporting more inclusive and equitable assessment (Deshpande et al., 2023; Holmes et al., 2019).

Designing and implementing assessments at scale also poses a challenge in K-12 education, particularly given the number and diversity of students in our educational system (Holmes et al., 2019; Li et al., 2022). AI-driven learner models offer the potential for scalability by automating aspects of assessment design, development, and administration (Attali, 2018). Such models have the potential to generate more personalized assessment tasks, analyze large datasets efficiently, and provide timely feedback to students and educators, thereby streamlining the assessment process and reducing logistical and administrative burdens (Grover, 2024; Holmes et al., 2019).

By purposefully gathering information to understand the characteristics and preferences of learners (e.g., cultural backgrounds, language proficiency, learning styles, accessibility needs and preferences) and developing robust learner models that have the potential to be AI-driven and responsive to these characteristics and preferences, assessment designers can determine upfront the features necessary for accessible and engaging assessment tasks that place students in optimal conditions to demonstrate what they know and can do (Hansen & Mislevy, 2008; Mislevy, 2004; Sato, 2024). Developing such learner models, however, requires careful consideration of ethical, practical, and theoretical factors to ensure they meet the diverse needs of all students (Holstein et al., 2019; He & von Davier, 2016). The following section presents a review of literature with particular focus on the degree to which diverse learner characteristics currently are considered and incorporated into AI applications in U.S. K-12 education. More specifically the literature was evaluated with the intention of addressing the following questions:

Regarding the development of an AI-driven learner model:

1. In what ways can AI technology responsibly be leveraged to support a more robust understanding of K-12 learner capabilities and needs for assessment of students, especially those with disabilities, from culturally and linguistically diverse backgrounds, and who are currently and historically systemically marginalized and underserved?

2. What factors are needed to develop an AI-driven learner model that can accommodate a range of learning styles and minimize assessment bias to ensure inclusivity and equity?

Regarding the implementation of an AI-driven learner model:

3. How can AI-driven learner models be employed to improve decision-making processes in the areas of accessibility and inclusion in assessment (e.g., a priori matching of supports vis-a-vis student capabilities and needs)?

4. What are the potential successes and challenges of implementing AI-driven learner models in K-12 assessments? Given recent paradigm shifts in accessibility and inclusion, what intersectional opportunities with AI ought to be prioritized?

5. What are the ethical considerations associated with the use of AI in developing learner models for K-12 assessments, particularly with respect to fairness and validity?

Examining Learner Characteristics in AI Applications: A Review of Recent Literature on U.S. K-12 Education

This section describes a systematic literature review that examines how learner characteristics, particularly those relevant to students who are currently and have been historically systemically marginalized and underserved, are considered and incorporated into AI applications in U.S. K-12 education. Information from this review is used to address the questions listed above as well as informs the validity framework presented in a subsequent section of this paper.

Method

There were multiple steps involved in the systematic review of literature. First, a literature search of several electronic databases and online search engines, including ERIC, Google Scholar, Semantic Scholar, and PsychINFO was conducted. The list of search engines considered for this review is presented in Table 1. Keyword searches included but were not limited to terms such as “artificial intelligence,” “accessibility,” “equity,” and “inclusion.” Key topical areas such as empirical research, ethics, policy, and theory also were incorporated into the search. Table 1 provides the complete list of keywords used, both individually and in combination, for the literature review search. Inclusion and exclusion criteria were meticulously considered. Documents were required to be publicly available, published in English language journals or documents, and have publication dates ranging from 2014 to 2024. Documents needed to focus on one of the key topic areas—theoretical, empirical, policy, or ethical—and be framed within the context of the U.S. K-12 school setting. Additionally, journal articles had to be peer-reviewed. Any search findings that did not meet these criteria were excluded from the review. This literature search yielded an initial pool of documents that researchers considered for inclusion in their review of literature.

Second, each researcher selected one of the four topical areas of focus (i.e., theoretical, empirical, policy, ethical) and reviewed relevant documents in the initial pool. The researcher verified that a document met the inclusion criteria and should be included in the final analysis, and if it did, reviewed the document, extracting the following information:

- Theoretical documents: purpose; intended audience; underlying theory/theories; conceptual framework, models, and/or theory of action;
- Empirical documents: type of study; data source(s); subjects; n-size; research question(s)/purpose; factors/variables; analyses; key findings; key implications;
- Ethical documents: key considerations;
- Policy documents: by whom the policy was created; for whom the policy is intended; focus (e.g., principles, standards, guidelines); whether it is elective or required; and
- Additionally, for all documents, information related to fairness, equity, inclusion, and accessibility.

Each document was reviewed by a second researcher to verify inclusion in the final analysis as well as the information extracted from each document. If there was disagreement between the two reviews, a third researcher reviewed the document in question and made a consensus-based decision regarding the document's inclusion in the final analysis and the information extracted from the document.

Finally, data from each topic area were synthesized to surface and articulate general themes vis-a-vis fairness and accessibility in AI, as well as gaps and needs. Researchers conferred with each other throughout the process to ensure the accuracy and consistency of the interpretations. The syntheses for each topic area follow.

Table 1
Review of Literature: Summary of Sources, Key Words Searched, and Criteria

Sources	• ArXiv (post-print articles used only; confirmed with second academic search engine) • Elicit • ERIC • Google Scholar • Google search • Research Gate • Science Direct • Semantic Scholar • PubMed • PsychINFO • Digital Commons
Key words used (and combinations)	• Accessibility • Artificial Intelligence (AI) • Assessment • Equity • Empirical • Ethics • Fairness • Inclusion • K-12 education • Policy • Student learning • Theory • U.S. (schools, context) • Validity
Inclusion criteria	• Publicly available • English language journals/documents • Publication date range: 2014-2024, seminal work excepted • Theoretical, empirical, policy, ethical • K-12 • U.S. context • Peer reviewed (applies to articles/papers)
Exclusion criteria	• Not publicly available, fee/purchase required • Not an English publication • Publication date range before 2014 • Not peer reviewed (for articles/papers)

Findings

The initial search yielded 59 documents, all of which were recorded for tracking purposes. Of these, 23 documents met the criteria for inclusion in the final analysis (see Appendix A). In total, 5 empirical studies, 4 ethical texts, 10 policy-related documents, and 4 theoretical documents were analyzed for this literature review. Outcomes of the qualitative analysis of the documents and syntheses of information are summarized below.

Theoretical Documents

Four documents that address theoretical perspectives met the required inclusion criteria and were reviewed. Three of the documents address general K-12 educational contexts; one document focuses more specifically on language education. All four documents specifically address diverse learners, learning styles and preferences, and culturally and linguistically responsive approaches.

Song et al. (2024) present a framework based on Universal Design for Learning (UDL) to create inclusive AI education for K-12 students. This framework integrates AI learning design principles with UDL's multiple means of engagement, representation, and action and expression. It provides practical pedagogical examples and emphasizes project-based learning, collaborative learning, and interactive demonstrations. The framework aims to guide educators in designing AI curricula that cater to diverse learners' needs and promote fairness and accessibility.

Similarly, Mizumoto (2023) explores the integration of data-driven learning (DDL) and generative AI (GenAI), such as ChatGPT, in language learning. Mizumoto introduces the Metacognitive Resource Use (MRU) framework, which positions DDL within a broader ecosystem of language resources, including GenAI tools. The MRU framework emphasizes metacognitive knowledge and regulation, guiding learners to strategically use diverse language resources. The article suggests pedagogical strategies for enhancing learners' self-awareness and resource use and calls for future research to empirically assess the integrated DDL-GenAI approach and the MRU framework.

In considering how AI technologies can be utilized to enhance English language teaching for diverse learners, Anis (2023) outlines strategies for integrating AI tools such as language models and adaptive learning systems into educational practices. It emphasizes the potential of AI to address individual learning needs, offer personalized feedback, and support diverse learning styles. The article also discusses the implications of AI adoption in education, highlighting the importance of teacher training, ethical considerations, and the need for inclusive pedagogical frameworks to ensure equitable access to learning opportunities for all students.

Madaio et al. (2022) also critique the typical emphasis on performance gaps in AI fairness evaluations, pointing out that they overlook deeper systemic inequalities inherent in the development of the system itself. Drawing on critical theory and Black feminist scholarship, they show how educational AI technologies continue to reinforce historical injustices, even when the technologies seem to perform equally well. For example, the authors note that fairness approaches often focus on treating all groups the same, thereby reinforcing inequities because the algorithms fail to account for the societal complexities present within categories such as race and gender. The authors call for justice-oriented approaches and a complete redesign of educational AI to foster equity, stressing the importance of addressing and changing the structural inequalities that are built into these technologies. The authors argue that it is not enough to focus on identity and inclusion, but instead to address structural inequalities through participatory design.

All four documents emphasize the importance of creating inclusive and accessible learning environments using AI technologies. They highlight the potential of AI to provide personalized learning experiences, noting that AI can tailor educational content to meet the unique needs and preferences of individual students, enhancing engagement and learning outcomes. Each document introduces a framework or set of strategies for integrating AI in education. The documents address the need for responsible use of AI, ensuring data privacy, avoiding algorithmic bias, and promoting fairness and equity in educational practices. All documents call for ongoing research to evaluate the effectiveness of AI in education. They emphasize the importance of collaboration among educators, researchers, and policymakers to develop and implement effective AI-driven educational practices.

Policy-Related Documents

The search for policy and related documents yielded 10 documents that met the inclusion criteria. All of these documents included elective (rather than mandatory) guidance on AI considerations, with each framing these considerations as guidelines, and some also delineating principles (Burstein, 2023; TeachAI, 2023; UNESCO, 2021; UNICEF, 2020), standards (Burstein, 2023), and strategies (Roshanaei, 2023). The overarching intended audiences described in these publications included educators, policymakers, and researchers. Educational institutions were also specifically named as an intended audience in several documents (Burstein, 2023; TeachAI, 2023; UNESCO, 2021), while Cardona and Rodriguez (2024) defined their intended audience as developers of AI-enabled products and services in the educational sector, including product leads, innovators, designers, developers, customer-facing staff, and legal teams across research, nonprofit, and for-profit organizations.

Generally, these documents point out that the integration of AI in U.S. K-12 education presents both opportunities and challenges, particularly regarding the inclusion of diverse learners. Despite the potential benefits of personalized learning and enhanced assessment accuracy, current AI applications often lack comprehensive consideration of the diverse spectrum of learners, and this oversight can inadvertently reinforce existing biases, disproportionately affecting currently and historically systemically marginalized and underserved student groups. Several policy documents emphasize the importance of considering diverse learner characteristics in AI applications (e.g., Burstein, 2024; Cardona & Rodriguez, 2024).

The literature reviewed also highlights the challenges and potential biases in AI applications. White et al. (2024) advocates for "the adoption of new K-12 educational policies to ensure equitable access to AI education” (p. 1). Marino et al. (2023) note that while AI has the potential to revolutionize how students with disabilities learn, it also risks perpetuating existing biases if not carefully implemented. UNESCO (2021) and UNICEF (2020) underscore that it is essential that AI’s ethical deployment in education includes transparent and bias-minimizing practices to avoid exacerbating inequalities.

As AI tools continue to influence educational landscapes, their integration requires careful ethical and educational policy frameworks. Research suggests that machine learning may offer a more transparent alternative to certain AI applications, especially when considering the algorithmic oversight needed to maintain fairness and minimize bias (TeachAI, 2023). Roshanaei et al. (2023) note that AI has improved accessibility for students with disabilities by providing assistive technology solutions for them, such as screen readers and braille translators. They also state that “AI systems must be grounded in datasets reflecting diverse experiences and viewpoints to avoid biases and ensure fairness” (p. 138). Salas-Pilko et al. (2024) point out that AI technologies can enhance accessibility in education by providing personalized learning experiences that cater to individual student needs, including those with disabilities and multilingual learners. This personalized approach can help bridge the gap in educational outcomes for currently and historically systemically marginalized and underserved students.

The need for robust policy and ethical considerations is a recurring theme in the reviewed literature. UNESCO (2021) and UNICEF (2020) both emphasize the importance of developing policies that ensure the ethical use of AI in education, particularly regarding data privacy and the protection of children’s rights. To align with a broader goal of transformative justice in educational AI applications, policies must be designed to safeguard against the misuse of AI and ensure that all students benefit equitably from technological advancements, and they must include transparency and accountability to ensure that AI-driven decisions are fair and just.

Empirical Studies

Five relevant empirical studies met the required inclusion criteria and were reviewed. Park et al. (2022) investigated a visual interface designed to teach AI planning concepts to upper elementary students (grades 3-5), finding that while the interface showed promise in making AI concepts accessible to students, it also revealed usability challenges, particularly for students using different input devices. This study underscores the importance of designing AI-enhanced educational tools with accessibility in mind (e.g., resizable text and customizable color schemes). Similarly, Ali et al. (2021) focused on educating middle school students about deepfakes and misinformation, emphasizing the critical need for developing AI literacy in young students to navigate an increasingly AI-influenced information landscape.

In the realm of assessment and feedback, Li et al. (2018) and Hastings et al. (2018) explored the use of machine learning models to evaluate students’ writing. They developed models that demonstrate the potential for AI to provide automated feedback on complex writing tasks. Li et al. (2018) suggest that automated assessments of students’ language use could inform the development of personalized scaffolding to support learners with varying levels of academic language proficiency. Hastings et al.

(2018) investigated techniques to reduce the amount of human-annotated training data needed for such models, suggesting that AI could make sophisticated writing assessment and feedback more feasible across diverse educational contexts. In complement to these studies, Attali (2018) examined the large-scale deployment of automatic item generation for math assessment, finding that automatically generated items performed similarly to manually created ones. This approach has significant potential for providing more adaptive and personalized math assessments for learners with diverse abilities and backgrounds, which can be expanded to other content areas.

Ethical Texts

The search for articles with a focus on ethics yielded four relevant texts. Adams et al. (2023) identified several core ethical principles adapted for K-12 education, including justice and fairness, beneficence, and freedom and autonomy. They also uncovered principles unique to this context (e.g., pedagogical appropriateness and children's rights) that underscore the need for AI systems in education to be designed with the specific needs and rights of all students in mind. Bulathwela et al. (2024) further emphasize this point, arguing that while AI in education (AIEd) shows promise for personalized learning and improved access, it risks exacerbating existing inequalities if not implemented thoughtfully. They caution against "techno-solutionism" and stress the importance of addressing underlying political and social issues while developing AIEd solutions.

Dieterle et al. (2022) provide a framework for understanding these challenges by identifying five interrelated divides in AI education: access, representation, algorithms, interpretations, and citizenship. These divides can create either virtuous or vicious cycles in educational outcomes. Dieterle et al. (2022) propose strategies such as empowering diverse interest holder communities, infusing evidence-based decision making with cultural responsiveness, and building human capacity through professional development. These approaches align with Porayska-Pomsta and Holmes's (2023) emphasis on transparency, explicability, and human autonomy in AI educational systems. They argue that AIEd must critically examine its assumptions, involve diverse interest holders, and consider its broader societal impact to ensure ethical implementation.

Discussion

AI-driven learner models in U.S. K-12 education show promise for personalized learning and improved outcomes, particularly when integrated with UDL, offering real-time adjustments and individualized feedback (Mizumoto, 2024; Song et. al, 2023). When AI addresses student diversity, it supports inclusivity and equity, helping all students succeed.

Limitations

Despite a systematic and rigorous approach, this literature review has several limitations. Relying on English-language documents excludes insights from non-English publications, potentially limiting comprehensiveness. Focusing on peer-reviewed documents from 2014 to 2024 may exclude critical works outside this timeframe. The specific focus on U.S. K-12 education, while relevant, may exclude important international and post-secondary information. Additionally, excluding non-peer-reviewed documents, aimed at ensuring methodological rigor, might overlook innovative or emerging work. These limitations highlight the need for ongoing examination to understand AI applications in U.S. K-12 education promoting accessibility, inclusion, equity, and validity. Nonetheless, the reviewed literature underscores critical implications and gaps, particularly for currently and historically systemically marginalized and underserved groups.

Implications

AI has the potential to tailor educational experiences to individual students' needs, preferences, and learning styles. Frameworks such as UDL and MRU emphasize creating inclusive and accessible learning environments. These frameworks can guide educators in designing AI curricula that engage diverse learners through multiple means of engagement, representation, and action and expression, promoting fairness and accessibility (Mizumoto, 2023; Song et al., 2024).

The ethical integration of AI in education is paramount. Ensuring data privacy, avoiding algorithmic bias, and promoting fairness and equity in AI-driven educational practices are crucial. Theoretical perspectives advocate for justice-oriented approaches and the need to confront systemic inequities embedded in educational technologies (Madaio et al., 2022). Furthermore, policy documents emphasize the necessity of robust ethical guidelines and multi-interest-holder collaboration to ensure AI applications do not exacerbate existing biases and inequities (Burstein, 2023; UNESCO, 2021).

Effective implementation of AI in education requires significant investment in teacher training. Educators must be equipped with the knowledge and skills to leverage AI tools effectively while understanding their ethical implications and potential biases (e.g., Anis, 2023). The reviewed policy documents provide guidelines and principles for integrating AI in education, focusing on accessibility and inclusion of diverse learners. These documents underscore the importance of developing policies that ensure the ethical use of AI, safeguard data privacy, and protect children's rights. They advocate for AI systems that are tested and validated with diverse populations to ensure broad applicability and fairness (Cardona & Rodriguez, 2024; Roshanaei, 2023).

Gaps and Challenges

Despite the potential benefits of AI, there is a risk of perpetuating existing biases if AI systems are not carefully designed and implemented. Studies highlight the disproportionate impact of AI biases on marginalized communities and the exclusion of these groups from AI development processes (Marino et al., 2023; White et al., 2024). Ensuring that AI systems are developed using diverse datasets and are inclusive of all student groups is crucial to mitigating these risks. While theoretical and policy documents provide valuable guidelines, empirical studies are necessary to validate these approaches and understand their impact on diverse learners.

Addressing the structural inequalities that AI technologies may perpetuate is a significant challenge. Research by Madaio et al. (2022) call for a fundamental redesign of educational AI systems to promote equity and justice, emphasizing the need to confront and transform the structural inequalities embedded in these technologies. This requires a comprehensive approach that involves diverse interest holders in the design and implementation of AI-driven educational tools.

AI-driven learner models can enhance personalized learning and promote educational equity, but significant challenges remain. Addressing these requires ethical guidelines, empirical validation, and a commitment to inclusivity. Collaboration among educators, researchers, policymakers, and communities is crucial to harness AI's potential in education equitably. The following section presents a fairness- and transformative justice-based validity framework to ensure AI applications in K-12 assessments are inclusive of students with disabilities, culturally and linguistically diverse students, and other marginalized groups.

Validity Framework for Equitable AI Applications in K-12 Educational Assessment

To ensure the equitable application of AI in K-12 educational assessment, the authors propose the following validity framework, the Fair AI Framework. Centered on fairness and transformative justice, this framework is intended to prioritize equitable access to AI tools and ensure these tools do not perpetuate existing biases or inequalities. Generally, fairness refers to designing AI systems that treat all students justly, providing equal opportunities for success. Transformative justice goes further by actively aiming to address and dismantle systemic barriers and inequities within educational environments. This approach aims to prevent harm and create positive, inclusive changes that benefit currently and historically systemically marginalized and underserved student groups so that they can thrive. The framework includes five key components: Accessible and Inclusive Design, Ethical Implementation, Continuous Monitoring, Evaluation, and Improvement, Interest Holder Engagement, and Policy and Advocacy. Each component is grounded in theory and research and linked through a coherent theory of action.

Framework Components

Accessible and Inclusive Design: Involves designing AI tools that are responsive to the diverse visual, auditory, cognitive, and physical accessibility needs and preferences of students, as well as sensitive to their cultural and linguistic backgrounds. Creates an AI-driven accessible and inclusive learning environment that moves away from a deficit-based model that focuses on what students may be "missing" to an asset-based model leveraging the richness of student diversity and allows for diverse frames of reference, ways of knowing, and means of communication. Additionally, integrates assistive technologies to support students, including features like screen readers, voice recognition, and customizable interfaces. Relevant resources include: UDL principles to ensure AI tools provide multiple means of engagement, representation, and action and expression (CAST, 2018; Christensen et al., 2014; Christensen et al, 2023; Sato, 2023); the Sociocultural Dimensions Matrix (Sato, 2023, 2024) to systematically consider sociocultural factors that affect learners' understanding of information and their demonstration of knowledge; the Leading for Equity Framework (National Equity Project, 2024) that emphasizes inclusive design that considers equity, complexity, and user-centered approaches to address systemic oppression; and guidelines for reviewing demographic data for use in measuring "fairness and bias" in AI systems (Bogen, 2024).

Ethical Implementation: Involves ensuring AI algorithms are trained on diverse datasets and regularly audited for biases to maintain algorithmic fairness. Uses fairness-aware algorithms that minimize disparate impacts on different student groups (e.g., Ferrara et al., 2023). Establishes robust data governance policies to protect student data privacy and ensure that data collection, storage, and usage comply with ethical standards and legal regulations. Promotes transparency in AI decision-making processes by providing clear explanations of how AI tools make decisions and establishing accountability mechanisms for addressing any adverse impacts. Relevant resources include: guidance emphasizing fairness-aware AI algorithms, data governance policies protecting student privacy, regular auditing for biases, transparency in AI decision-making processes, and engagement with diverse interest holders to ensure ethical and equitable use of AI in educational settings (Council of the Great City Schools & Consortium for School Networking, 2023; Miao & Holmes, 2021; National Institute of Standards and Technology, 2023).

Continuous Monitoring, Evaluation, and Improvement: Involves conducting regular impact assessments to evaluate the effectiveness and fairness of AI applications, using both quantitative and qualitative data to measure educational outcomes and identify disparities. Establishes mechanisms for improvement that include (1) continuous feedback from students, educators, and other interest holders and (2) consideration of the emerging body of knowledge on diversity and innovations to iteratively improve AI tools and ensure they meet the evolving needs of diverse learners. Implements longitudinal studies to understand the effects of AI applications on student learning and equity, tracking educational outcomes over time to identify trends and areas for improvement. Relevant resources include: guidance and frameworks that focus on continuous monitoring and evaluation of AI applications in education, recommend regular impact assessments, mechanisms for interest holder feedback, longitudinal studies to understand long-term effects on student learning and equity, and engagement with diverse perspectives (Council of the Great City Schools & Consortium for School Networking, 2023; Miao & Holmes, 2021; National Institute of Standards and Technology, 2023).

Interest Holder Engagement: Includes involving a diverse group of interest holders in the design and implementation of AI tools, including educators, students, parents, community members, and experts. Ensures that the voices of currently and historically systemically marginalized and underserved groups are heard and valued. Provides ongoing professional development for educators on the ethical use of AI in assessments. Fosters partnerships with community organizations, advocacy groups, and local institutions, as appropriate, to support the inclusive implementation of AI, engaging these partners in co-creating and disseminating AI-driven educational assessment practices. Relevant resources include: The Emerging Technology Adoption Framework which provides a structured approach for engaging diverse interest holders, including educators, students, and families, throughout the process of evaluating, adopting, and implementing AI and emerging technologies in PK-12 education (Ruiz et al., 2022).

Policy and Advocacy: Includes advocating for policies that promote equity in AI applications in educational assessment, including funding for research on equitable AI, support for inclusive design practices, and regulations to prevent discriminatory practices. Develops and disseminates ethical guidelines for AI in educational assessment, informed by principles of fairness, justice, and inclusivity, to be adopted by educational institutions and technology developers. Raises awareness about the importance of ethical AI in educational assessment across interest groups and advocate for responsible and equitable AI adoption. Relevant resources include: The Education Technology Industry's Principles for the Future of AI in Education framework which advocates for implementing AI in education with purpose, transparency, and equity (Software & Information Industry Association, 2023).

The proposed validity framework operates within a theory of action that integrates its components to achieve equitable AI applications in K-12 educational assessment (see Figure 2). The starting point is the accessible and inclusive design of AI tools to meet the diverse needs of all students. Ethical implementation ensures that AI applications are fair, transparent, and secure, with algorithms regularly audited for biases and data privacy rigorously protected. Continuous monitoring, evaluation, and improvement provide critical insights into the impact of AI on student learning and equity, with feedback loops and longitudinal studies informing iterative improvements to AI tools. Active engagement of diverse interest holders ensures that AI tools are relevant and effective, supported by professional development and community partnerships that promote ethical AI use. Finally, equity-focused policies and ethical guidelines create a supportive environment for the fair and inclusive implementation of AI, with public awareness campaigns advocating for responsible AI adoption.

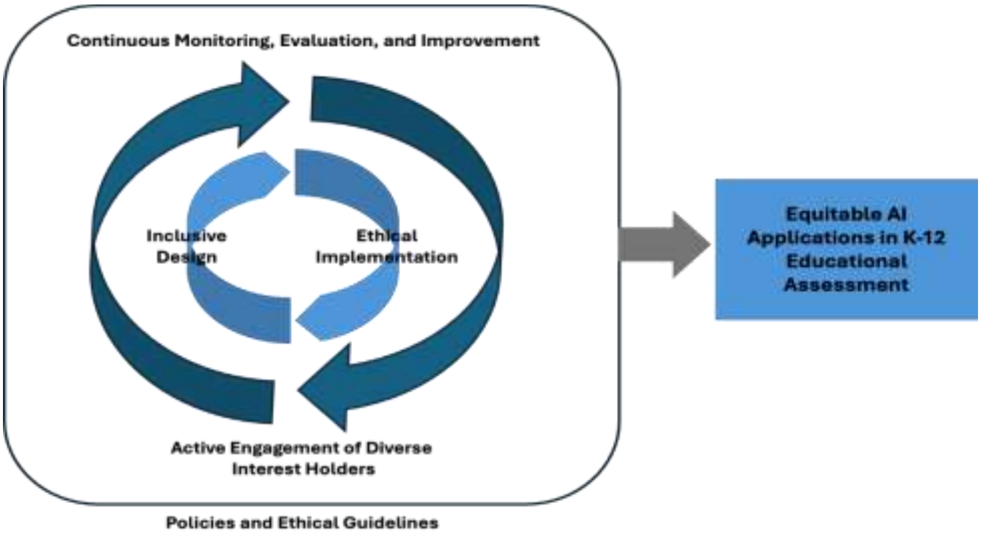

Continuous Monitoring, Evaluation, and Improvement
Inclusive Design
Ethical Implementation
Active Engagement of Diverse Interest Holders
Policies and Ethical Guidelines
Equitable AI Applications in K-12 Educational Assessment

Figure 2. Theory of Action: Fair AI Framework

By integrating these components, the framework aims to create a system where AI-driven tools are used ethically and inclusively, enhancing learning outcomes for all students. This approach aims to promote AI applications in educational assessment that contribute to transformative justice, promoting equity and fairness for diverse learners.

Recommendations for Next Steps

The importance of AI-driven learner models in promoting accessible, inclusive, equitable, and valid assessments for all learners necessitates a strategic and multifaceted approach. The authors recommend a path forward that includes specific considerations across research, policy, practice, and collaboration. The recommended next steps are designed to advance the development and implementation of AI technologies that address the diverse needs of students, particularly those from currently and historically systemically marginalized and underserved groups.

Research should involve a multidisciplinary (e.g., education, computer science, ethics) and holistic approach to consider the effects of socio-economic, cultural, and linguistic factors in educational assessment. It also should include input from various interest holders to ensure AI validity. Empirical studies must evaluate AI's effectiveness and fairness across varied contexts. Regular bias audits are crucial, and methodologies should be developed to detect and mitigate biases. Longitudinal studies are necessary to track the effects of AI-driven assessments on educational outcomes and equity. Scalable AI solutions adaptable to different contexts and accessible to schools with varying resources are essential.

Equity-focused policies at the federal, state/territory, and local levels should require rigorous testing for fairness and inclusivity of AI tools. Establishing and promoting ethical frameworks based on principles of fairness, transparency, accountability, and respect for student privacy and autonomy is essential. Securing funding for the research and development of equitable AI technologies and providing resources for schools and educators to implement and sustain inclusive and fair AI-driven learner models is vital.

Investment in professional development for educators should cover inclusive design principles, ethical considerations, and practical AI applications in the classroom, particularly vis-a-vis assessment. Promoting the adoption of inclusive design practices in developing AI tools is essential, ensuring these applications are co-designed with input from diverse interest holders. Employing AI-based language translation and adaptation applications is essential for supporting culturally and linguistically diverse students. Integrating assistive technologies into AI-driven assessments to support students with disabilities ensures these technologies are adaptable to various needs and are user-friendly.

Interest holder collaboration should focus on co-creating AI tools responsive to diverse learners' needs. Engaging communities, especially those currently and historically systemically marginalized and underserved, in developing and implementing AI-driven learner models ensures their voices are heard and their needs are addressed in design and implementation. Maintaining transparency in developing and using AI in education by clearly communicating the purposes, benefits, and risks of AI tools to all interest holders is essential.

Conclusion

This paper examined advancements in methods and technologies, particularly through the integration of innovative tools that incorporate AI, and the implications of such advancements, focusing on learner models that are AI-driven, and their potential to transform educational assessment practices within the U.S. K-12 assessment context. As a result of the literature review and development of the Fair AI Framework, responses to the five questions articulated at the beginning of this paper are as follows:

First Question: The literature underscores that AI technology can be responsibly used to enhance understanding of diverse learner capabilities by incorporating principles and practices related to UDL and socioculturally responsive pedagogy, for example. By leveraging AI to tailor assessments and support mechanisms based on individual needs, AI tools can provide more nuanced and effective educational support. The proposed validity framework further emphasizes integrating assistive technologies and socioculturally responsive design to ensure AI applications meet the diverse needs of all students.

Second Question: Developing an AI-driven learner model includes: the application of inclusive design principles, which support diverse learning styles and needs; ensuring algorithmic fairness and conducting bias audits to minimize assessment bias; and integrating feedback mechanisms and continuous evaluation processes to refine AI tools to promote inclusivity and equity, as well as address both the potential and limitations of AI technologies.

Third Question: AI-driven learner models can significantly enhance decision making in accessibility and inclusion by using data-driven insights to match educational supports with student needs proactively. The literature suggests that AI tools can help ensure that students receive appropriate accommodations based on their unique capabilities and needs, providing a more responsive and equitable assessment experience for students.

Fourth Question: The implementation of AI-driven learner models can enhance personalization and support for diverse learners; however, challenges include bias and equitable access. Recent paradigm shifts highlight the need for intersectional approaches that consider socio-economic, cultural, and linguistic diversity.

Fifth Question: The literature and framework highlight the value of fairness-aware algorithms, protecting data privacy, and maintaining transparency in AI decision-making processes. Ensuring that AI systems are regularly audited for biases and that ethical guidelines are followed is essential, aligning with the broader goals of transformative justice and equity.

Integrating AI-driven learner models in K-12 education can transform equity but requires addressing ethics, inclusivity, and fairness. The Fair AI Framework offers a comprehensive, research-informed approach, recommending interdisciplinary research, policy advocacy, collaboration, and evaluation for continuous improvement to ensure accessible, inclusive, and equitable educational assessments for all learners.

Declarations

Gen-AI use: The authors of this article declare (Declaration Form #: 2711240909) that ChatGPT, developed by OpenAI, and Claude, developed by Antropic, were used for Data Analysis and interpretation as well as for checking grammar, style, and coherence in up to 10% of this article. They further affirm that all content generated by GenAI has been carefully reviewed, and they assume full responsibility for its inclusion.

Author Contribution: Material preparation, writing, and review were supervised by Dr. Edynn Sato. All authors contributed to the manuscript's conceptualization, data curation and analysis, and writing.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Ethical Approval: We declare that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as this study uses data shared with the public.

Consent to Participate: All authors have given their consent to participate in submitting this manuscript to this journal.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). Ethical principles for artificial intelligence in K-12 education. *Computers and Education. Artificial Intelligence*, *4*, 100131-. https://doi.org/10.1016/j.caeai.2023.100131

Ali, S., DiPaola, D., Lee, I., Sindato, V., Kim, G., Blumofe, R., & Breazeal, C. (2021). Children as creators, thinkers and citizens in an AI-driven future. *Computers and Education. Artificial Intelligence*, *2*, 100040-. https://doi.org/10.1016/j.caeai.2021.100040

Anis, L. (2023). Leveraging artificial intelligence for inclusive English language teaching: Strategies and implications for learner diversity. *Journal of Educational Technology*, *45*(3), 234-250. https://doi.org/10.1234/jet.2023.00456

Attali, Y. (2018). Automatic item generation unleashed: an evaluation of a large-scale deployment of item models. *Artificial Intelligence in Education*, 17–29. https://doi.org/10.1007/978-3-319-93843-1_2

Bogen, M. (2024). Navigating demographic measurement for fairness and equity: AI governance in practice guide. Center for Democracy & Technology. https://cdt.org/insights/navigating-demographic-measurement-for-fairness-and-equity/

Bulathwela, S., Pérez-Ortiz, M., Holloway, C., Cukurova, M., & Shawe-Taylor, J. (2024). Artificial intelligence alone will Not democratise education: on educational inequality, techno-solutionism and inclusive tools. *Sustainability*, *16*(2), 781-. https://doi.org/10.3390/su16020781

Burstein, J. (2023). The Duolingo English Test Responsible AI Standards. [Updated March 29, 2024]. https://duolingo-papers.s3.amazonaws.com/other/DET%2BResponsible%2BAI%2BStandards%2B-%2B040824.pdf

CAST (2018). *Universal Design for Learning Guidelines version 2.2.* http://udlguidelines.cast.org.

Cawthon, S., & Shyyan, V. V. (2022). Routledge Encyclopedia of Education: Accessibility and Accommodations on Large Scale Assessments. https://doi.org/10.4324/9781138609877-REE52-1

Cardona, M. A., & Rodriguez, R. J. (2024). Designing for Education with Artificial Intelligence: An Essential Guide for Developers. U.S. Department of Education. https://tech.ed.gov/files/2024/07/Designing-for-Education-with-Artificial-Intelligence-An-Essential-Guide-for-Developers.pdf

Christensen, L., Shyyan, V., & Johnstone, C. (2014). Universal design considerations for technology-based, large-scale, next-generation assessments. *Perspectives on Language and Literacy*, *40*(1).

Christensen, L., Shyyan, V., & MacMillan, F. (2023). Developing an Accessibility Review Process for English Language Proficiency Tests. *Language Testing*, 02655322231168386.

Council of the Great City Schools, & Consortium for School Networking. (2023, October 11). *K-12 generative AI readiness checklist questionnaire (Version 1.1)*. https://www.cgcs.org/genaichecklist

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2019). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2). https://www.tandfonline.com/doi/full/10.1080/10888691.2018.1537791

de Klerk, G. (2008). Cross-cultural testing. In M. Born, C.D. Foxcroft & R. Butter (Eds.), *Online readings in testing and assessment, International Test Commission*. http://www.intestcom.org/Publications/ORTA.php

Deshpande, D.S., Shanmugapriya, I., Choudhary, R.K., Patil, S.S., & Sing, A. (2023). An empirical study on the impact of artificial intelligence in education with reference to teaching and learning. *Asian and Pacific Economic Review*, *16*(1), 1350-1355. https://doi.org/10.5281/zenodo.1234567

Dieterle, E., Dede, C., & Walker, M. (2024). The cyclical ethical effects of using artificial intelligence in education. *AI & Society*, *39*(2), 633–643. https://doi.org/10.1007/s00146-022-01497-w

eCampusOntario (n.d.). Designing and developing high quality student-centered online/hybrid learning experiences. https://opentextbooks.uregina.ca/qualitycourses/

Ferrara, C., Sellitto, G., Ferrucci, F., Palomba, F., & De Lucia, A. (2023). Fairness-aware machine learning engineering: How far are we? *Empirical Software Engineering*, *29*(9). https://link.springer.com/article/10.1007/s10664-023-10402-y

GAO (2022, June). K-12 education: Student population has significantly diversified, but many schools remain divided along racial, ethnic, and economic lines. https://www.gao.gov/assets/gao-22-104737.pdf.

Grover, S. (2024). Teaching AI to K-12 learners: Lessons, issues, and guidance. Proceedings of the 55th ACM Technical Symposium on Computer Science Education, March 20-23, Portland, OR. https://doi.org/10.1145/3626252.3630937

Hall, E.T. (1983). *The dance of life*. New York, NY: Doubleday.

Hammond, Z. L. (2015). *Culturally responsive teaching and the brain*. New York: Corwin Press.

Hansen, E.G., & Mislevy, R.J. (2008). Design patterns for improving accessibility for test takers with disabilities. https://files.eric.ed.gov/fulltext/EJ1111295.pdf

Hastings, P., Hughes, S., & Britt, M. A. (2018). Active learning for improving machine learning of student explanatory essays. *Artificial Intelligence in Education*, 140–153. https://doi.org/10.1007/978-3-319-93843-1_11

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment CBA PIAAC NLP LSA. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.). *Handbook of Research on Technology Tools for Real-World Skill Development, Volume II*. Hersey, PA: Information Science Reference, pp. 749-776.

Hofstede, G., & Hofstede, G.J. (2005). *Cultures and organizations: Software of the mind* (2nd ed.). New York: McGraw-Hill.

Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial intelligence in education: Promise and implications for teaching and learning*. Boston, MA: Center for Curriculum Redesign.

Holstein, K., McLaren, B.M., & Aleven, V. (2019). Co-designing a real-time classroom orchestration tool to support teacher-AI complementarity. *Journal of Learning Analytics*, *6*(2). 27-52.

Ji, L-J., Zhang, Z., & Nisbett., R.E. (2004). Is it culture or is it language? Examination of language effects in cross-cultural research on categorization. *Journal of Personality and Social Psychology, 87* (1), 57-65.

Kulich, S.J. (2009). Values theory: Sociocultural dimensions and frameworks. In S.W. Littlejohn & K.A. Foss (Eds.), *Encyclopedia of communication theory*. Thousand Oaks, CA: SAGE Publications, Inc.

Levine, R. (1997). *A geography of time*. New York: Basic Books.

Lewis, R.D. (2006). *When cultures collide: Leading across cultures* (3rd ed.). Boston: Nicholas Brealey International.

Li, L. (2022). A literature review of AI education for K-12. *Canadian Journal for New Scholars in Education, 13*(3). https://doi.org/10.5206/cjnse.v13i3.12345

Li, H., Gobert, J., Dickler, R., & Morad, N. (2018). Students' academic language use when constructing scientific explanations in an intelligent tutoring system. *Artificial Intelligence in Education*, 267–281. https://doi.org/10.1007/978-3-319-93843-1_20

Marino, M. T., Vasquez, E., Dieker, L., Basham, J., & Blackorby, J. (2023). The future of artificial intelligence in special education technology. *Journal of Special Education Technology, 38*(3), 404-416.

Marion, S.F., & Pellegrino, J.W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25* (4), 47-57.

Masuda, T., & Nisbett, R.E. (2001). Attending holistically vs. analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, 81, 922–934.

McDonald, J., & West, R.E. (2021). *Design for learning: Principles, processes and praxis*. EdTech Books, Provo, UT. https://edtechbooks.org/id

Michel, R., & Shyyan, V. (2024). Accessibility as a Core Value for Locally Responsive Assessments. [Manuscript in preparation] In Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy. R.E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.). Routledge.

Miao, F., & Holmes, W. (2021). Artificial intelligence and education: Guidance for policy-makers. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000376709

Madaio, M., Blodgett, S. L., Mayfield, E., & Dixon-Román, E. (2024). *Beyond "fairness:" Structural (in)justice lenses on AI for education.* Microsoft Research. https://www.microsoft.com/research/publication/beyond-fairness-structural-injustice-lenses-on-ai-for-education

Mislevy, R. J. (2004). A Brief Introduction to Evidence-Centered Design (Technical). Los Angeles: National Center for Research on Evaluation, Standards, and Student Testing (CRESST). https://files.eric.ed.gov/fulltext/ED483399.pdf

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics, 3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

Molle, D., Sato, E., Boals, T., & Hedgspeth, C.A. (Eds.) (2015). *Multilingual learners and academic literacies: Sociocultural contexts of literacy development in adolescents*. New York: Routledge.

Montenegro, E., & Jankowski, N.A. (2017). Equity and assessment: Moving towards culturally responsive assessment. https://files.eric.ed.gov/fulltext/ED574461.pdf

National Equity Project. (2024). Leading for equity framework. https://www.nationalequityproject.org/framework

NCES (2020). Projections of education statistics to 2028. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2020024.

NCES (2023). Condition of education. https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2023144REV

NCES (2024). Condition of education 2024. https://nces.ed.gov/programs/coe/pdf/2024/CGG_508c.pdf

Nisbett, R. E. (2003) *The geography of thought: How Asians and Westerners think differently, and why*. New York, NY: Free Press.

National Institute of Standards and Technology. (2023). AI risk management framework: AI RMF (1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Park, K., Mott, B., Lee, S., Gupta, A., Jantaraweragul, K., Glazewski, K., Scribner, J. A., Ottenbreit-Leftwich, A., Hmelo-Silver, C. E., & Lester, J. (2022). Investigating a visual interface for elementary students to formulate AI planning tasks. *Journal of Computer Languages (Online)*, *73*, 101157-. https://doi.org/10.1016/j.cola.2022.101157

Parrish, P., & Linder-VanBerschot, J.A. (2010). Cultural dimensions of learning: Addressing the challenges of multicultural instruction. *International Review of Research in Open and Distance Learning*, 11(2), 1-19.

Pearson, P., & Garavaglia,D. (2003). *Improving the information value of performance items in large scale assessments: NAEP validity studies*. Working Paper Series. ED Pubs.

Pellegrino, J.W. (2003, Winter). "Knowing What Students Know." *Issues in Science and Technology 19*(2).

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (2001). Knowing what students know: The science and design of educational assessment., Washington, DC: National Academies Press.

Porayska-Pomsta, K., & Holmes, W. (2023). Conclusions: Toward ethical AIED. In *The Ethics of Artificial Intelligence in Education* (1st ed., pp. 271–281). Routledge. https://doi.org/10.4324/9780429329067-14

Preston, J. P., & Claypool, T. (2021). Analyzing assessment practices for Indigenous students. *Frontiers in Education 6*, 679972.

Roshanaei, M., Olivares, H., & Lopez, R. R. (2023). Harnessing AI to foster equity in education: Opportunities, challenges, and emerging strategies. *Journal of Intelligent Learning Systems and Applications, 15*(4), 123-143. https://doi.org/10.4236/jilsa.2023.154009

Ruiz, P., Richard, E., Chillmon, C., Shah, Z., Kurth, A., Fekete, A., Glazer, K., Pattenhouse, M., Fusco, J., Fennelly-Atkinson, R., Lin, L., Arriola, S., Lockett, D., Crawford-Meyer, V., Karim, S., Hampton, S., & Beckford, B. (2022). Emerging technology adoption framework: For PK-12 education. Digital Promise. https://doi.org/10.51388/20.500.12265/161

Ryan R. M., & Weinstein N. (2009). Undermining quality teaching and learning: A self-determination theory perspective on high-stakes testing. *Theory Res Educ*. 7:224–233

Sato, E. (2017). Culture in fair assessment practices. In H. Jiao & R.W. Lissitz (Eds.) *Test fairness in the new generation of large-scale assessments.* Maryland Assessment Research Center Conference. College Park, MD.

Sato, E. (2023, April). Equity-minded Assessment: A Framework for Born Socio-culturally Responsive Assessment. Paper presented at the National Council on Measurement in Education. Chicago, IL.

Sato, E. (2024). Born socioculturally responsive assessment: An approach to design and development. [Manuscript in preparation] In Socioculturally Responsive Assessment: Implications for Theory, Measurement, and Systems-Level Policy. R.E. Bennett, L. Darling-Hammond, & A. Badrinarayan (Eds.). Routledge.

Shyyan, V. V., & Christensen, L. L. (2018, September). A framework for understanding English learners with disabilities: Triple the work (ALTELLA Brief No. 5). Madison, WI: University of Wisconsin–Madison, Alternate English Language Learning Assessment (ALTELLA). Retrieved from University of Wisconsin–Madison, Wisconsin Center for Education Research: http://altella.wceruw.org/resources.html

Software & Information Industry Association. (2023). *Education Technology Industry's Principles for the Future of AI in Education*. https://edtechprinciples.com/principles-for-ai-in-education/. https://edtechprinciples.com/

Solano-Flores, G., & Nelson-Barber, S. (2001). On the cultural validity of science assessments. *Journal of Research in Science Teaching*, 38(5), 553-573.

Song, Y., Weisberg, L. R., Zhang, S., Tian, X., Boyer, K. E., & Israel, M. (2024). A framework for inclusive AI learning design for diverse learners. *Computers & Education*.

TeachAI. (2023). AI guidance for schools toolkit: Principles. TeachAI. https://www.teachai.org/toolkit-principles

UNESCO. (2021). Ethics of artificial intelligence: Towards a global framework. United Nations Educational, Scientific and Cultural Organization. https://unesdoc.unesco.org/ark:/48223/pf0000381137

UNICEF. (2020). Policy guidance on AI for children: Draft 1.0. United Nations Children's Fund.

U.S. Department of Education, Office of Educational Technology. (2023). Artificial intelligence and future of teaching and learning: Insights and recommendations. Washington, DC. https://www2.ed.gov/documents/ai-report/ai-report.pdf

Usher, A. (2012). Student motivation: An overlooked piece of school reform. Center on Education Policy. https://files.eric.ed.gov/fulltext/ED532666.pdf

Usher, A. (2012). Student motivation: An overlooked piece of school reform. Center on Education Policy. https://files.eric.ed.gov/fulltext/ED532666.pdf

Wang, Q., & Leichtman, M.D. (2000). Same beginnings, different stories: A comparison of American and Chinese children's narratives. *Child Development, 71* (5), 1329-1346.

Wexler, N. (2019). Why we're teaching reading comprehension in a way that doesn't work. https://www.forbes.com/sites/nataliewexler/2019/01/23/why-were-teaching-reading-comprehension-in-a-way-that-doesnt-work/?sh=59996f0c37e0.

Wexler, N. (2021). New data shows building knowledge can boost reading comprehension. https://nataliewexler.substack.com/p/new-data-shows-building-knowledge?s=w

White, S. V., Childs, J., Koshy, S., & Scott, A. (2024). Policy Implementation in the Era of Responsible Artificial Intelligence (AI) use in K-12 Education. Proceedings of the 2024 on RESPECT Annual Conference, 81–85. https://doi.org/10.1145/3653666.3656097

Woodruff, K., Hutson, J., & Arnone, K. (2023). Perceptions and barriers to adopting artificial intelligence in K-12 education: A survey of educators in fifty states.

Appendix A

Documents That Met the Full Inclusion Criteria and Were Reviewed

Document name	Theoretical	Empirical	Policy	Ethical
Adams et al., 2023				x
Ali et al., 2021		x		
Anis, 2023	x			
Attali, 2018		x		
Bulathwela et al., 2024				x
Burstein, 2023			x	
Cardona & Rodriguez, 2024			x	
Dieterle et al., 2024				x
Hastings et al., 2018		x		
Li et al., 2018		x		
Madaio, 2024	x			
Marino et al., 2023			x	
Mizumoto, 2023	x			
Park et al., 2022		x		
Porayska-Pomsta & Holmes, 2023				x
Roshanaei et al., 2023			x	
Salas-Pilco et al., 2022			x	
Song et al., 2024	x			
TeachAI, 2023			x	
UNESCO, 2021			x	
UNICEF, 2020			x	
White et al., 2024			x	
Woodruff et al., 2023			x	

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(Special issue); 282-301

Human-Centered AI for Discovering Student Engagement Profiles on Large-Scale Educational Assessments

Hongwen GUO* Matthew JOHNSON** Luis SALDIVIA***
Michelle WORTHINGTON**** Kadriye ERCIKAN*****

Abstract

Large-scale assessments play a key role in education: they provide insights for educators and stakeholders about what students know and are able to do, which can inform educational policies and interventions. Besides overall performance scores and subscores, educators need to know how and why students performed at certain proficiency levels to improve learning. Process/log data contain nuanced information about how students engaged with and acted on tasks in an assessment, which hold promise of contextualizing a performance score. However, one isolated action event observed in process data may be open to multiple interpretations. To address this challenge, in the current study, use of multi-source data (performance and process) was proposed to integrate sequential process data with response data to create engagement profiles to better reflect students' test-taking processes and knowledge states. Most importantly, AI algorithms were used to assist and amplify human expertise in the creation of students' engagement profiles, so that the information extraction from the multi-source data can be scaled up to enhance the value of large-scale assessments in teaching and learning. Various machine learning techniques were leveraged to develop the general framework of the human-centered AI (HAI) approach to help human experts efficiently and effectively make sense of the multi-source data. Using a mathematics item block from the National Assessment of Educational Progress (NAEP) for illustrations, data from over 14,000 students resulted in ten preliminary profiles, more than half of which were associated with low performing students. Such HAI approaches and data insights are expected to generate rich and meaningful feedback for educators and stakeholders.

Keywords: Multi-source data, machine learning, human-in-the-loop, visualization, Large-scale assessment

Introduction

Large-scale assessments (LSAs) play a crucial role in assessing and improving the quality of education at state-, national-, and international-levels. These measures inform educators and stakeholders on what students know and can do, so that they can prepare for education policies and interventions in teaching and learning (Gordon, 2020; Pellegrino, 2020). These assessments may also help guide resource allocation in education (NAGB, 2024b). However, for educators to use these large-scale assessment results in a classroom, a performance score may not be enough, particularly for low performing students. Educators need to know how and why these students got low scores, so that they can prepare targeted and effective interventions. In the rapidly evolving landscape of educational technologies, many LSAs are administered on digital platforms, where students' digital footprints (i.e., process/log data) are collected (Ercikan & Pellegrino, 2017; Ercikan et al., 2023). These process data contain nuanced information about how students solved the tasks and how they navigated through the assessment, which may reflect students' cognitive thinking processes, affective states, and test-taking strategies, holding promise of providing contextual information beyond a performance score.

* Principal Research Scientist, ETS, Princeton-New Jersey, US, hguo@ets.org, ORCID ID: 0000-0002-1751-0918
**Principal Research Director, ETS, Princeton-New Jersey, US, msjohnson@ets.org, ORCID ID: 0000-0003-3157-4165
*** Research Strategic Advisor, ETS, Princeton-New Jersey, US, lsaldivia@ets.org, ORCID ID: 0009-0007-3482-7654
**** Assessment Development Manager, ETS, Princeton-New Jersey, US, mworthington@ets.org, ORCID ID: 0009-0006-0480-3769
***** SVP of Global Research, ETS, Princeton-New Jersey, US, kercikan@ets.org, ORCID ID: 0000-0001-8056-9165

To cite this article:
Guo, H., Johnson, M., Saldivia, L., Worthington, M. & Ercikan, K. (2024). Human-Centered AI for discovering student engagement profiles on large-scale educational assessments. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 282-301. https://doi.org/10.21031/epod.1532846

Received: 14.05.2024
Accepted: 24.09.2024

As described in prior literature (Ercikan et al., 2023), the key uses of process data in assessments include score validation (Wise, 2021), assessment design improvement (Pellegrino, 2020), evidence for the targeted construct (Johnson & Liu 2022; Levy, 2020; Pohl et al., 2021), group comparison (Ercikan et al., 2020; Guo & Ercikan, 2021a, 2021b; Rios & Guo, 2020), and feedback enrichment (Guo et al., 2024; Zoanetti & Griffin, 2017). Many features have been extracted from process data, among which time on task is one of the most-commonly used features. Item response times have been shown to be significantly associated with performance on LSAs (Ercikan et al., 2020; Guo & Ercikan, 2021a, 2021b; Rios et al., 2017; Wise, 2017, 2021). To solve an item, an appropriate amount of time needs to be spent in understanding the question and working towards its solution. A hard item usually takes a longer time to solve, and an easy item a shorter time. On LSAs, certain rapid responding behaviors associated with guessing are often observed, which may compromise score validity. However, such behavior can be associated with varied factors, such as low-test motivation, specific test-taking strategies (e.g., skipping hard items), and speededness because of time pressure. A rapid response may also be observed from a student simply because of their high proficiency and efficiency. Without context, it is challenging to explain why students rapidly respond to an item on a test. Other process data features face similar challenges in interpretation as well, since one isolated behavior observed during the test-taking process may be open to multiple interpretations (Ercikan et al., 2023; Greiff et al., 2016; Guo et al., 2024).

To address these challenges, the current study proposes to integrate sequential process data with response data (also called multi-source data in the current study) to create engagement profiles to better reflect students' test-taking processes for rich insights beyond their knowledge and skills in a knowledge domain. More specifically, in the current study, the multi-source data for each student (i.e., the item navigation sequence, the response time sequence associated with each item navigation, and the response score sequence corresponding to the items) are used to create preliminary profiles. These profiles would inform educators and stakeholders not only what a performance level of a student or a student group reached, but also how they worked through the assessment, which in turn would help to shed light on why they performed at this level. Such information and data evidence are particularly useful for helping low performing students.

Most importantly, given the large sizes of data students produced on LSAs, we propose to use AI/machine learning algorithms to assist and amplify human expertise in the creation of engagement profiles, so that the information extraction from these multi-source data can be scaled up to enhance the impact of large-scale assessments in teaching and learning. Therefore, one goal of the current study is to propose a general framework that uses AI to augment human experts in uncovering data insights and expediting the development of student profiles on a large scale. The engagement profiles created in the study may reflect students' navigation processes, affective states, and test-taking strategies, among others. Note that the engagement profiles use students' sequential information in the response and process (i.e., timing and navigation) data when they interacted with the assessment platform, which provide richer context than the commonly used engagement indices (such as the response time effort proposed by Wise and colleagues, 2005, 2017, 2021), but requires more intensive computational demands.

To create engagement profiles, our research questions are:

- RQ-1. What are the considerations in data preprocessing? This includes the creation of meaningful and explainable process data features and data visualization to assist human experts.
- RQ-2. How to start from scratch for human experts to discover engagement profiles? Since the proposed engagement profiles are novel, it is necessary to discover them from data. Given the expected large sizes of students on LSAs and the volume of process data students produced, we need an efficient approach to select a manageable and informative sample of students' data for human experts to examine and discover the initial engagement profiles.

- RQ-3. How to scale up the engagement profile creation? That is, how to combine the unique strengths of AI algorithms and human knowledge, thereby improving overall performance and productivity in the profile creation for all students.

The paper is structured as follows. In the Method section, we introduce the large-scale assessment, response data, and process data used in the study. Then we present the proposed Human-centered AI (HAI) framework for data analysis and profile creation. Three major steps in the HAI architecture are described in detail to show how human knowledge plays a crucial role in the profile creation, how to leverage AI algorithms (such as machine learning, deep learning, and active learning methods) to enhance data analysis and pattern identification, and how to combine AI power and human expertise to create the profiles. In the Results section, we present results obtained from each of the three major steps in the HAI architecture. In the last section, we discuss the potential uses of the engagement profiles, the implications and significance of the HAI general framework, and limitations of our current work.

Methods

Research Design

HAI approaches have been strongly recommended in education to make decisions based on established, modern learning principles, wisdom of educational practitioners, and human knowledge in the educational assessment community (Baker, 2021; Guo et al., 2024; Miao et al., 2021). In this study, the application of HAI is intended to assist and amplify (rather than displace) human expertise in understanding students' knowledge, skills, and abilities (KSAs) "beyond a sole focus on students' core academic performance measured by large-scale assessments, to support students and teachers with actionable feedback that nurtures the broader skills students need to succeed and thrive" (Office of Educational Technology, 2023).

In this study, data from the National Assessment of Educational Progress (NAEP) Grade 8 Mathematics assessment were used for illustration. NAEP provides important information about student academic achievement and learning experiences in various subjects and has provided meaningful results to improve education policy and practice in the US.

The NAEP mathematics assessment was first administered digitally in 2017. This digital administration allowed for the collection of process data, including information on how long students spent on the assessment questions (commonly referred to as timing data), how they navigated through items, and how students used onscreen assistive digital tools to develop their responses. NAEP also releases samples of process data. Interested researchers can consult their website for more information (NAGB, 2020).

Five broad content areas in the NAEP mathematics assessment are number properties and operations; measurement; geometry; data analysis, statistics and probability; and algebra, which are measured using a variety of item types including selected responses (e.g., single-and multiple-selection multiple choice, and matching), and short or extended constructed response (CR). Items are also classified on three levels of cognitive complexity (Low, moderate, and high), based on the items' demands on students' thinking process (NAGB, 2024a).

Response Data

For this study, we selected one item block in the 2022 NAEP 8th Grade Math assessment, because it contained many publicly released items. Detailed information on content of the released items and scoring rules can be found on the National Center of Educational Statistics (NCES) website (NCES, 2022). This item block has 14 items, and students can have 30 minutes to work on it (refer to Table 1). In the "Item" column of Table 1, items with * are released items. In the "Skill" column, Data stands for Data Analysis, Statistics, and Probability; Number stands for Number Properties and Operations. In the "Item Type" column, SR stands for selected response, and CR stands for short or extended constructed response.

Table 1.

Item information

Item	Skill	Item Type	Max Score	Item Difficulty
1	E) Algebra	SR	1	Very Easy
2*	D) Data	CR	2	Medium
3	E) Algebra	SR	1	Very Easy
4	B) Measurement	SR	1	Very Easy
5*	D) Data	SR	1	Easy
6*	E) Algebra	CR	1	Easy
7	C) Geometry	SR	2	Medium
8	A) Number	SR	1	Easy
9*	E) Algebra	SR	1	Hard
10*	C) Geometry	SR	2	Easy
11	A) Number	MS	2	Medium
12*	E) Algebra	SR	1	Easy
13*	C) Geometry	CR	4	Hard
14*	B) Measurement	SR	1	Medium

The maximum item score varies from 1 point to 4 points, as shown in the “Max score” column in Table 1. The total maximum raw score on the block is 21 points. For example, Item 13 is a CR item, with a maximum score of 4. A student can get a score of 0 for completely incorrect responses, a credit of 1, 2, or 3 for partial correct responses, or a score of 4 for full credit.

Process Data

NAEP digitally based assessments offer a testing environment that makes it possible to record students' interaction with the digital platform when students solve the tasks. Figure 1 shows a screenshot of the testing environment of one released item (NAGB,2020).

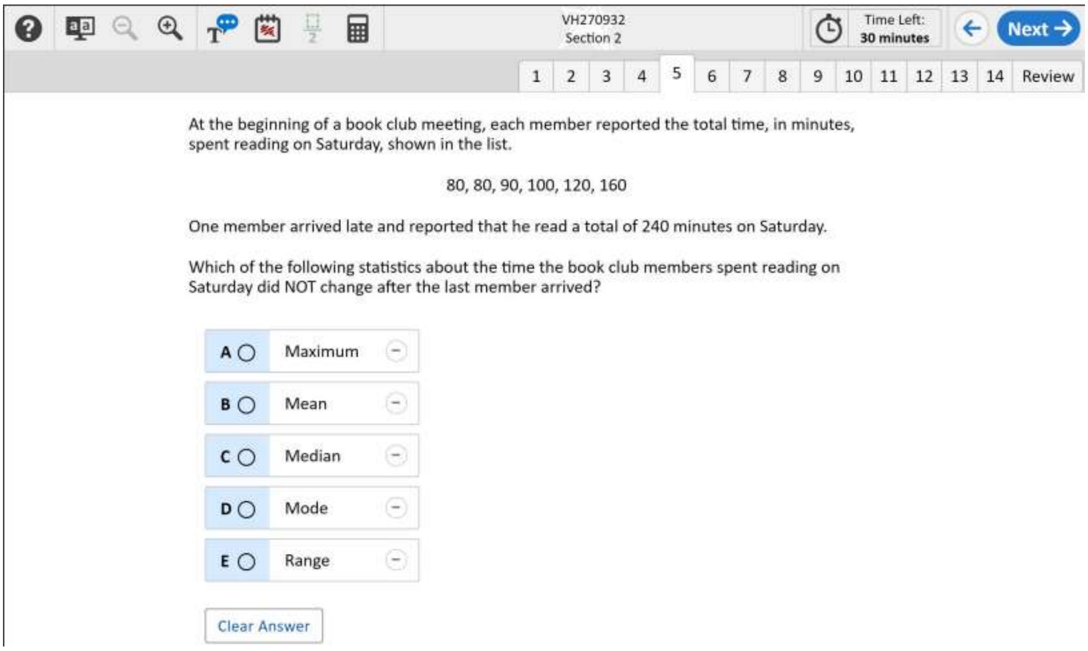
Starting from the upper left corner of the screen in Figure 1, the digital tools include help (a question mark), color contrast and theme change, zoom-in/out, text-to-speech, scratch work, equation editor, calculator (note that the studied item block allows the use of a calculator). On the upper left corner of the screen, students can monitor their session time (a clock icon), check their progress on the items, and move forward or backward of the pages/items by clicking on the item tags or using the ‘Next’ button. The ‘Review’ button allows students to get an overview of which items they had responded to and which they had not. Students' interactions with the testing environment were logged and collected to produce process data.

The process data contain logs of response processes collected from each student, such as item response time, use of the digital tools, and which items students were working on and for how long. Because of space limitations, please refer to NAGB (2020) for detailed information on the process data variables.

After removing students with irregular response time and abnormal completion on the selected item block, the sample size in the study is N=14,008.

Figure 1.

The NAEP testing environment in the 2022 Math Assessment (using one released as an illustration).



At the beginning of a book club meeting, each member reported the total time, in minutes, spent reading on Saturday, shown in the list.

80, 80, 90, 100, 120, 160

One member arrived late and reported that he read a total of 240 minutes on Saturday.

Which of the following statistics about the time the book club members spent reading on Saturday did NOT change after the last member arrived?

A ○ Maximum

B ○ Mean

C ○ Median

D ○ Mode

E ○ Range

Clear Answer

Data Analysis and Procedures

The proposed general framework of the human-centered AI (HAI) architecture (refer to Figure 2) attempts to amplify human knowledge, maximize AI power, and minimize redundancy of human labor, so that the data can be effectively and efficiently annotated to address the big data challenges.

There are three major steps in the architecture in Figure 2, each of which relies on human knowledge and decisions.

Step-1 (data preprocessing & feature engineering) includes data cleaning and feature engineering. This process is informed by insights gleaned from prior research, literature, and experiences on process data and test-taking behaviors on educational assessments.

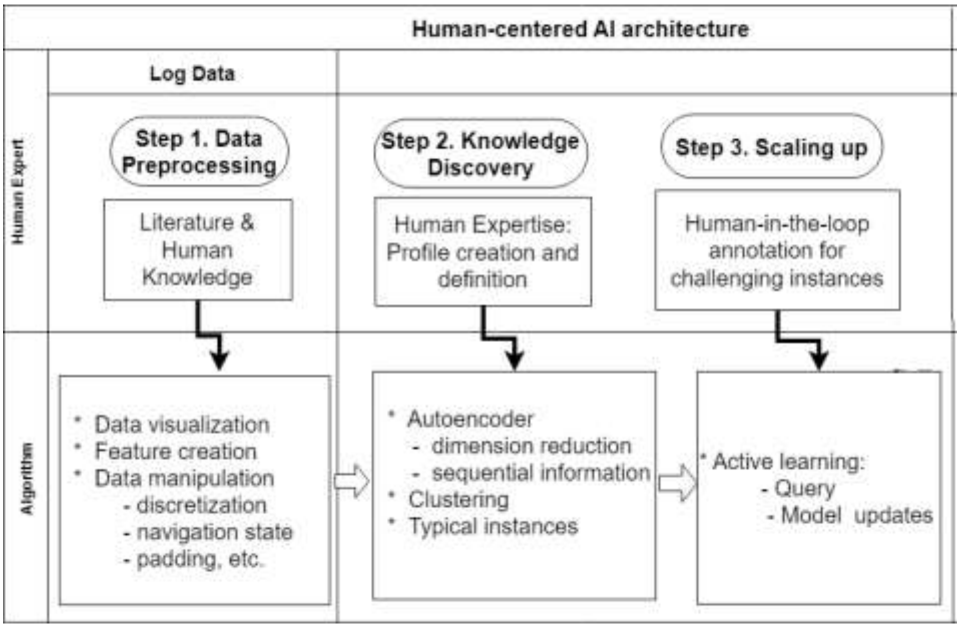
Step-2 (Knowledge Discovery) contains two parts: Part 1 uses an autoencoder (a self-supervised deep learning model) to compress the input sequential data (item responses, response times, and item navigation states) into a low-dimensional space (also called the latent space or the code space). Part 2 uses a clustering method to select typical data patterns for human experts to discover engagement profiles.

Step-3 (Scaling up) uses the active learning method to amply human experts' knowledge to unlabeled data.

A similar HAI architecture was applied in a recent study that investigated digital assistive tool uses, response times, and performance on the assessment platform (Guo et al., 2024). In the following paragraphs, we provide more details for each step.

Figure 2.

The general framework of the proposed HAI architecture



Step-1. Data Preprocessing and Feature Engineering

One of the prominent features extracted from process data is item response time. As discussed in the introduction, a rapid or prolonged response time may imply unexpected behaviors on the assessment. A rapid response is likely to reflect random guessing, which adds noise to response data and does not reflect students' true knowledge and skills (Guo & Ercikan, 2021a, 2021b; Guo et al., 2017; Rios & Guo, 2020; Wise, 2021). Therefore, we created six-time categories (refer to Table 2 below) to help to interpret the meaning of the item response time, in terms of whether a student spent reasonable time on an item (Guo & Ercikan, 2021a).

Table 2.

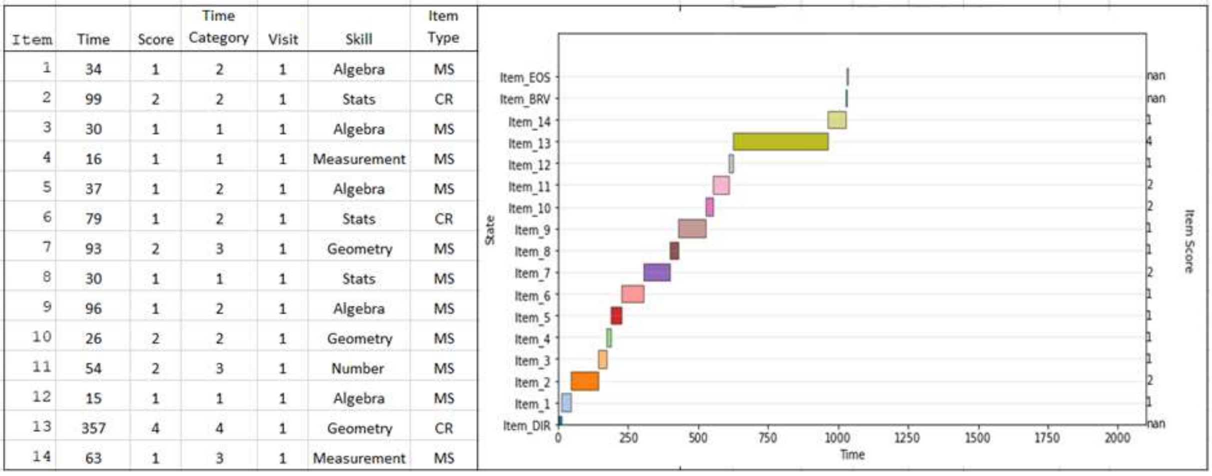
Definition of Time Categories and Their Possible Implications

Time Category	Definition	Implication
0	$T = 0$	Student did not work on the studied item.
1	$0 < T \leq$ Threshold*	Rapid responding (likely associated with random guessing)
2	Threshold* $\leq T < Q_1$	Student may spend sufficient time (but still low).
3	$Q_1 < T \leq Q_3$	Student spent sufficient time (in the middle quartiles).
4	$Q_3 < T \leq 95^{th}$ percentile	Student spent sufficient time (in the upper quartiles).
5	$T > 95^{th}$ percentile	Student spent prolonged time (likely facing challenges)

Notes. *In Column 2, T stands for item response time of the studied student on the studied item. The threshold* of response time for each item is derived using the hybrid method in Guo & Ercikan (2021a) to flag response times that indicate rapid responding behaviors. The quartiles Q_1, Q_2, Q_3 and 95th percentile are determined by the item response time distribution of the studied item for the sample (N = 14,008).*

Figure 3.

Data visualization for one instance



Item	Time	Score	Time Category	Visit	Skill	Item Type
1	34	1	2	1	Algebra	MS
2	99	2	2	1	Stats	CR
3	30	1	1	1	Algebra	MS
4	16	1	1	1	Measurement	MS
5	37	1	2	1	Algebra	MS
6	79	1	2	1	Stats	CR
7	93	2	3	1	Geometry	MS
8	30	1	1	1	Stats	MS
9	96	1	2	1	Algebra	MS
10	26	2	2	1	Geometry	MS
11	54	2	3	1	Number	MS
12	15	1	1	1	Algebra	MS
13	357	4	4	1	Geometry	CR
14	63	1	3	1	Measurement	MS

Notes. *The student had a total score of 21 out of 21, a total time of 1029 out of 1800 seconds and a total number of visit states of 17. The item-level summary information (item score, item response time, response time category, and number of item visits) is presented for the student in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student visited items linearly one item at a time and was in the highest performing profile.*

To illustrate, the extracted data for each student are presented at three levels: block level, item level, and granular navigation level (refer to Figure 3 as an example for one student). In Figure 3, the item block level summary is provided in the caption. For this student, the total score received is 21 out of the maximum 21 points; the total time spent is 1029 out of the maximum 1800 seconds; and the total number of visit/navigation states, which is 17, including reading direction (Item_DIR) at the beginning of the session, reviewing the block (Item_BRV) at the end of the session, and ending the session (Item_EOS). The navigation plot on the right side of the figure is a visualization of three sequences (navigation state, time on the state, and score received). Each colored rectangle shows the time spent on an individual navigation state. In the plot, the *x*-axis stands for the testing time, the *y*-axis on the left stands for the item state (i.e., what item the student was working on) and other navigation states, and the *y*-axis on the right stands for the item score the student obtained.

Figure 3 shows that this student worked linearly through the items by the item presentation order from Item 1 to Item 14. The table on the left-hand side of the navigation plot provides the item level information, regarding time category (refer to Table 2 for definition) on an item, total time spent on the item (in seconds), item score received, number of visits on the item, skill measured, and item type. Please refer to Figures A2 to A6 in Appendix for more examples.

In the data pre-processing step, we emphasized preserving sequential information and integration of response data and process data, because one isolated event was often open to multiple interpretations as to what generated it. For example, for a low performing student, a rapid response observed at the beginning of the assessment (refer to Figure A1 in Appendix for one example) and one observed at the middle of the assessment (refer to Figure A3 in Appendix for another example) clearly contain different meanings: the earlier rapid responding behavior is likely to be an indicator of the low test-taking motivation of the student, and the latter one is likely to be an indicator of applying a test-taking strategy of skipping a question on which the student may lack knowledge. On the other hand, for a high performing student with a perfect score, a rapid response may indicate high efficiency in solving the problems (refer to Figure 3 above for one example).

Data gathered for each student, as presented in Figure 3, ensures that the features are meaningful and interpretable for human understanding and annotation, to addresses RQ-1.

Step-2. Knowledge Discovery

In order to help human experts to start the profiling process from scratch, we used an autoencoder to compress the navigation sequences into a low-dimensional space. Autoencoders possess the ability to acquire compact representations of input data, operating in a self-supervised manner wherein data labels are absent (Geron, 2017). An effective autoencoder demonstrates proficiency in reconstructing input data autonomously upon decoding the code space. Within the autoencoder architecture, we implemented the long short term memory (LSTM) layers to maintain sequential information to capture the sequence nature of students' navigation of the item block and enhance data interpretation, particularly addressing RQ-2. Step-2 also includes a key component of knowledge discovery, for which, a clustering method (an unsupervised machine learning method) was applied to identify typical and representative instances in the big data, so that human experts could work on these typical instances to make sense of students' data, discover patterns, and define profiles. In this step, a large number of clusters was chosen on purpose to help with knowledge discovery. More specifically, in the current study, the number of clusters was 30; in each cluster, three representative instances were selected, and the visualization of each instance as in Figure 2 was presented to human experts to review and create profiles. Note that additional extreme cases (such as those with the highest/lowest score, with the longest/shortest time, and with the largest/smallest number of visits) were also presented to human experts to assist in profile creation.

Step-3. Scaling up

Based on the human labeled data, in this step, we applied an active learning approach integrated with a semi-supervised learning (AL&SSL) to predict the profiles for the unlabeled students' data (Guo et al., 2024; Rizve et al., 2021; Xie et al., 2019; Zhu et al., 2003), which addresses RQ-3.

Figure 4.

The active learning framework (Image from Radwan, 2019)

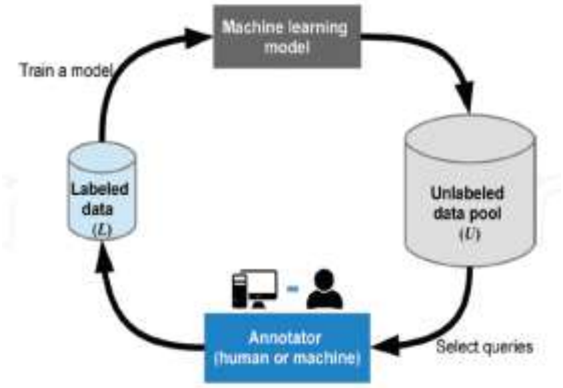
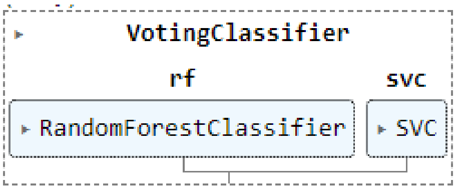


Figure 5.

The ensemble classifier



More specifically, there are two components in the AL&SSL framework (refer to Figure 4): a classifier (i.e., supervised machine learning model) and an oracle (i.e., human experts). There are four steps in one iteration in the AL&SSL framework as shown in Figure 4. We started from the "labeled data", which were obtained from Step-2 in our study. To build a "machine learning model", we used an ensemble voting classifier with a soft voting mechanism by combining a random forest classifier and a support vector machine (SVC) classifier and then trained and initialized the model with the labeled data (refer to Figure 5).

Using the trained model, we predicted pseudo labels/profiles for instances in the "unlabeled data pool". We then selected instances that were challenging to the model (i.e., instances with low confidences/probabilities for the pseudo label prediction) and asked human experts to annotate them (i.e., "human annotator" labeled data). At the same time, instances for which the model was accurate, and the prediction had high confidences, adopted the pseudo labels (i.e., "machine annotator" labeled data). The newly labelled data were then used to update both the training data and the model, and a new iteration started again. The iteration process in Figure 4 could end when all instances were labeled with satisfactory accuracy of the model and high prediction confidence of the pseudo labels.

Results

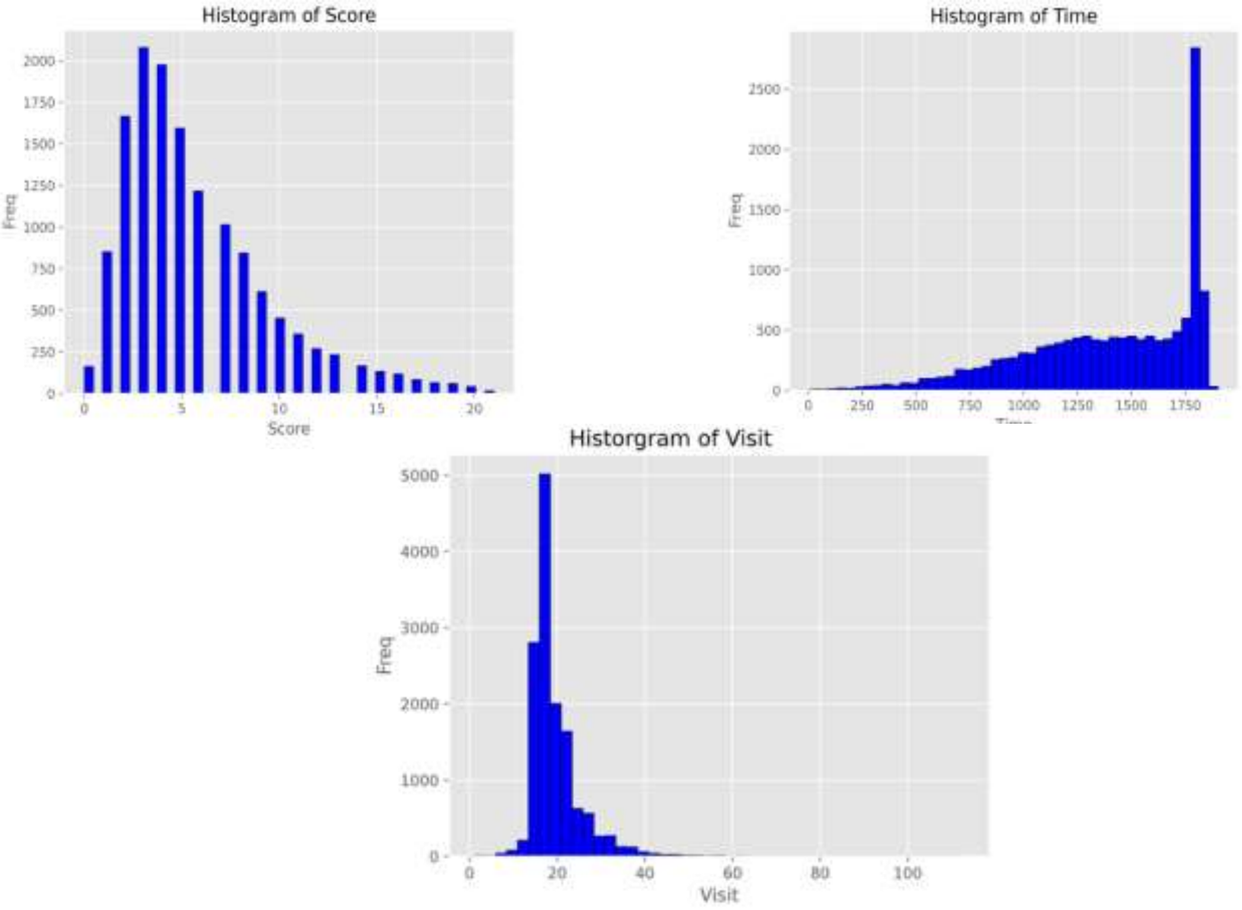
In this section, we first briefly show results from Step-1, and then we focus on the resulting engagement profiles from Step-2 and Step-3. The Python and TensorFlow libraries (Abadi et al., 2015) were used in producing the results.

Data Preprocessing Results – Step 1

As discussed in the methods section, in Step-1, we preprocessed the process data, extracted meaningful process features, and created visual presentations (as in Figure 3). In addition, Figure 6 below shows the histograms of the test-level variables (total score, total time, total number of visits) for the N=14008 students on the studied item block.

Figure 6.

The histograms of the test-level variables.



The histograms in Figure 6 show that all the test-level variables have skewed distributions: the total scores are concentrated on 3 and 4 points and have a long right tail; the total response times peak at the maximum allowed time (1800 seconds); and the total number of visits has a mode around 15 (note again that the number of items in the block is 14).

Table 3.

Item-level statistics

Item	Average Item Score	Item Difficulty	Median Time	95%tile Time	Rapid Responding Threshold	Max Score
1	0.69	0.69	43	126	13	1
2	0.99	0.50	121	310	17	2
3	0.65	0.65	65	206	38	1
4	0.50	0.50	73	233	42	1
5	0.16	0.16	69	207	42	1
6	0.25	0.25	128	336	32	1
7	0.58	0.29	90	238	25	2
8	0.17	0.17	52	144	31	1
9	0.48	0.48	148	351	45	1
10	0.27	0.14	68	207	7	2
11	0.23	0.12	40	134	20	2
12	0.41	0.41	32	94	16	1
13	0.33	0.08	195	456	24	4
14	0.10	0.10	60	202	5	1

Table 3 shows the item-level summary statistics, which shows that Item 13 (worth 4 points in total with an average item score of 0.33) is the most difficult item (difficulty is 0.08 = 0.33/4) and most time consuming (the median response time is 195 seconds); Item 1 is the easiest item (difficulty is 0.69) and second to the least time consuming (the median time is 43 seconds).

Item 12 is the least time-consuming item (the median time is 32 seconds). The 95%tile time (Time Category 5) shows that, again, Item 13 is the most time-consuming item (456 seconds), and Item 12 is the least (94 seconds). Also shown in Table 3, the thresholds for flagging rapid responses (Time Category 1) are the longest for Item 9 (45 seconds) and the shortest for Item 14 (5 seconds), respectively. For each student, the data were prepared and visualized as in Figure 3.

Knowledge Discovery Results – Step 2

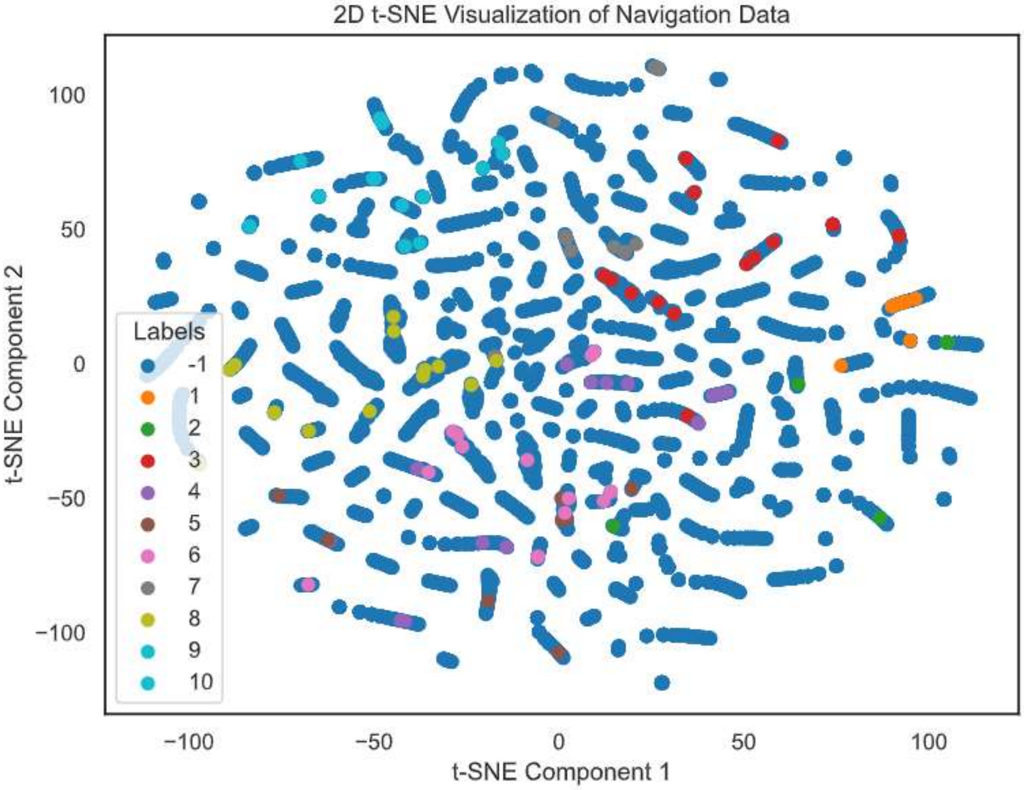
As noted, we had no labeled data on students' engagement with NAEP assessments, so it was necessary for human experts to discover such knowledge (i.e., engagement profiles). Exploration of autoencoder models with the long-short term memory layers (i.e., LSTM that preserve sequential information) led to the selection of a code space with eight dimensions. The code space, with summary statistics of total score, total response time, and the total number of item visits on the item block, a total of eleven variables, were used in clustering. Given the size of data (N=14008 by 11), we used the K-means method for easy processing. Note again that the purpose of clustering is to select representative instances for human experts' annotation and for discovery of possible engagement profiles. Other clustering methods are feasible as well.

To help human experts annotate the data, we purposely chose a number of clusters larger than necessity (in our case, the number of clusters selected was 30) to avoid missing potential engagement profiles. From each cluster, three representative instances closest to the centroid of a cluster were selected. Each representative instance is displayed as in Figure 3, as well as complimentary information about raw data sequences (such as item difficulty, item type and content), to produce a full picture of the student's engagement with the assessment for human annotation.

Human experts reviewed these representative instances, as well as extreme instances (such as highest/lowest scores, longest/shortest total times, and largest/least numbers of navigation states), aggregated and dissected the clusters and obtained ten differentiable profiles. Figure 7 is a 2-dimensional visualization of the profile distribution of about 150 initially labeled instances mapped into a 2-dimensional space, using the t-SNE techniques. Note that t-SNE is a dimensionality reduction technique commonly used for visualizing high-dimensional data in a lower-dimensional space (Van der Maaten & Hinton, 2008).

Figure 7.

Visualization of the ten profiles mapped into a 2-dimensional space



In Figure 7, the solid small dark blue circles (labeled as -1) are unlabeled instances; points with other colors are the initial labeled instances, which is about 1% of the studied sample. The order of labels (from 1 to 10) is roughly corresponding to the order of raw scores from low to high.

The preliminary ten profiles are described in Table 4. Again, refer to Figures A2 to A6 in Appendix for more examples with detailed discussion.

Table 4.

Description of the ten preliminary profiles created in the study

Label	Brief Descriptor	Profiles	Freq
1	Attempted little to no items	Unengaged group	1.77%
2	Very low score, low/regular time, and regular visit behavior	Low engagement with very low performance, navigated through most items with low time	11.02%
3	Low score, low/regular time, and regular visit behavior	Low engagement with low performance, navigated through most items with low time	14.00%
4	Low score, full/regular mixed time, and regular visit behavior	Engaged with low performance, navigated through most items, used mixed strategies	11.74%
5	Low or very low score, unregulated and/or speeded, with high visit behavior	Engaged with low performance, navigated through the items with high revisit rates, in some cases seemingly unpredictably, irregular navigation patterns with without speededness	14.16%
6	Low score, full/regular time with some prolonged item response times	Engaged with low performance, navigated through most items, spent a large amount of time on a small number of items, with or without speededness	7.67%
7	Medium score, regular time and visit behavior	Medium performing group in all dimensions	13.86%
8	Medium score, full/regular time with some prolonged item response times, and regular visit behavior	Medium performing, show strategic engagement behaviors (such as strategical response times)	18.50%
9	High score, regular time and visit behavior	High performing group, expected navigation patterns	5.43%
10	Very high score, regular time and visit behavior	Highest performing group, expected navigation patterns	1.87%

In Table 4, the very low, low, medium, high, and highest performing scores correspond roughly to the cutoffs of the lowest 10%, 1st quartile, between 2nd and 3rd quartile, 3rd quartile and the top 10% of the score range. For the last column in Table 4, please refer to the next section.

Overall, there were more profiles associated with low performing students than with high performing students. The first six profiles are low-performing ones, and they reflected different levels of engagement with the assessment from not engaged at all, low engagement, to engaged, which may reflect students' different levels of knowledge and skills, motivation in taking the assessment, affective states, time management, and/or test-taking strategies. On the other hand, the four profiles associated with medium and high scores show more engaged and expected test-taking behaviors.

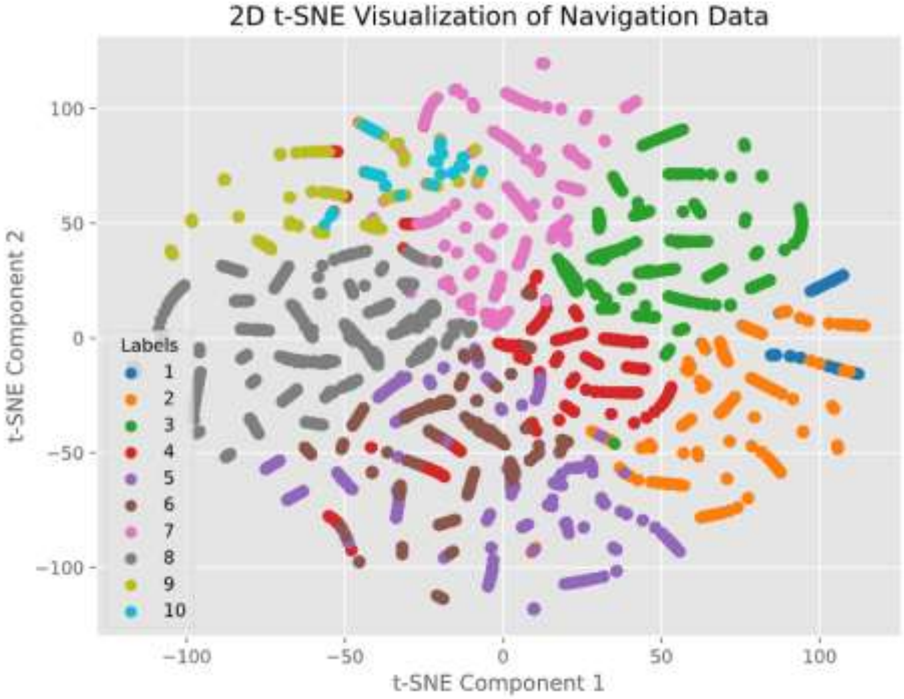
Scaling Up Results – Step 3

In Step-3, the ensemble model was applied to the initial labeled data in Step-2 to predict unlabeled data. Based on the model prediction, the least confident instances were selected for human manual labels, and then added to the training data. At the same time, based on the model accuracy and label confidence trade off, instances with the pseudo label confidence larger than 0.75 were added to training data as well (users could select other thresholds to experiment). The iteration process stopped when the model accuracy could not be improved. Figure 8 shows the fully labeled data in the 2-dimensional space using the t-SNE algorithm.

The proportions of students in each preliminary engagement profile in the fully labeled data are shown in the last column of Table 4. We observed that there were very small numbers of students (about 300 out of 14008) in either Profile 1 (the unengaged group) or Profile 10 (the highest performing group), and relatively large numbers of students in the middle profiles. Overall, about 60% of students in the studied sample were in the low- or very-low-score profiles, and 40% were in the medium- or higher-score profiles.

Figure 8.

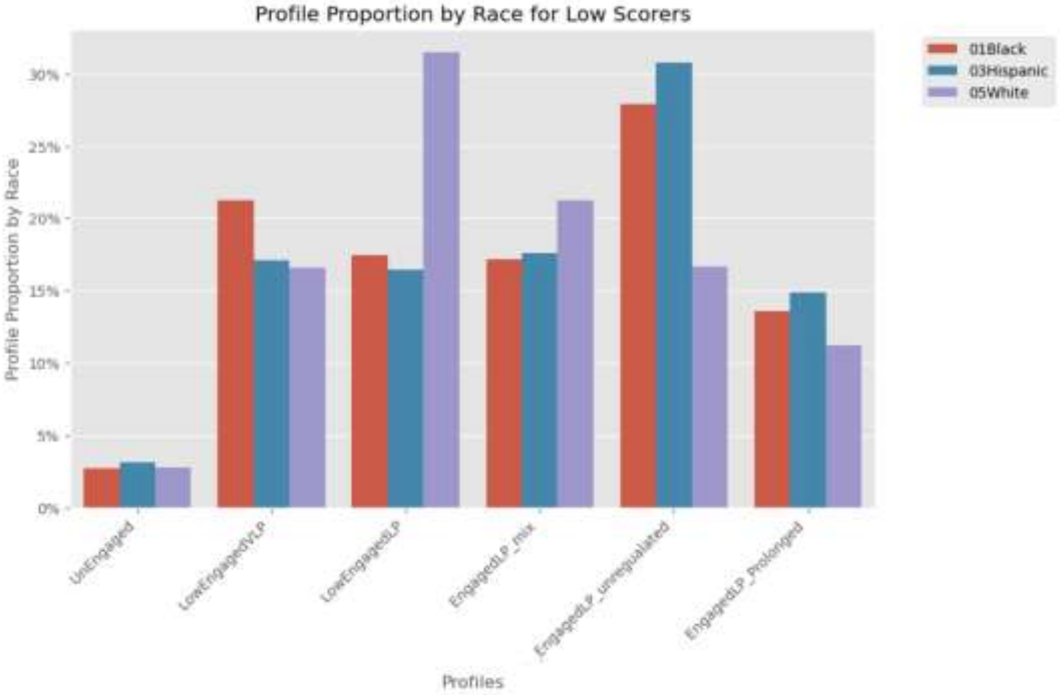
The 2-dimentional visualization of the fully labeled data, with different colors representing different profiles



These engagement profiles provide more contextualized information about test-taking processes and engagement status for individual students than the performance scores alone. Aggregation of the profiles can also shed light on student group differences. For example, among all the low performing students, Figure 9 shows that there are differences in profile proportions among different race groups (Black, n=1894; Hispanic, n=2206; and White, n=3515). A much higher proportion of white students is in Profile 3 (labeled as 'LowEngagementLP') than the black or Hispanic students, but much higher proportions of black and Hispanic students are in Profile 5 (labeled as 'EngagedLP_unregulated') than the white student group.

Figure 9.

The comparison of profile distributions by race for low performing students



Discussion

As evidenced in many recent studies (Ercikan & Pellegrino, 2017; Gordon, 2020; Guo et al., 2024; Pohl et al., 2021), digitally based assessments provide rich data about students' engagement with the assessments, which afford opportunities to investigate students' cognitive processes and problem-solving strategies, and to develop innovative assessments that better measure learning and support teaching.

In the current study, we used data from the NAEP math assessment to demonstrate how such large-scale data and AI can help generate students' engagement profiles beyond performance scores to support teaching and learning in the digital age (Office of Educational Technology, 2023). Preliminary results of the study show that there were more engagement profiles associated with low performing students, and these engagement profiles were differentiable from those with high performing students. The engagement profiles provide a holistic view of students' knowledge and skill, how they engaged with the assessment, their affective states, their test-taking strategies, and time management, etc. These profiles might suggest clues for understanding why students performed at certain levels, shed light on potential issues in their learning (such as lack of knowledge, low motivation, poor time management, difficulty with attention, focus, and organization, or other deficiency in learning and learning skills), particularly for the low-performing students (NASEM, 2018; NRC, 2000). This knowledge about students might help educators prepare differentiable intervention strategies for students in different profiles and help provide data evidence for making educational policy decision for improving teaching and learning.

Most importantly, given the large sizes of data collected from large-scale assessments, in this study, the general framework for the human-centered AI approach can support and amplify human ability in new knowledge discovery, so that useful information extraction from performance and process data can be scaled up to potentially enhance the impact of large-scale assessments. Our findings demonstrate the potential of advanced AI tools in facilitating a better understanding of students' test-taking processes and performance in context and minimizing potential false positive flags in detecting students' engagement in existing literature (Ercikan et al., 2023; Wise, 2017). The current approach allows for the exploration of innovations in assessments through harnessing AI power in analyzing extensive educational datasets to uncover patterns, trends, and insights that help inform instructional strategies and educational policies.

The significance of our innovation in analyzing large-scale assessment data is twofold. First, the proposed generic human-centered AI architecture is applicable for mining un-labeled and partially labeled complex and large-scaled educational data for insights. Human expertise and knowledge are used in every step of the work to ensure that the results are explainable and meaningful. This architecture can help to build and accelerate the creation of large and rich benchmark data sets in education for research and practice. Second, the work takes advantage of the rich process data from large-scale assessments to explore meaningful, and potentially actionable, data-based information that may complement and enhance the impact of large-scale assessments. Students' engagement profiles with the visualizations, combined with other complementary information about the students, for example, would help educators to prepare meaningful conversations with students who have different profiles for further interventions. Aggregation of engagement profiles for groups of students within a region, a school district, or a school, would also help stakeholders to make informed educational policy decision, when compared with student bodies of similar racial/ethnic composition (NAGB, 2024b).

The current exploration work has a few limitations. First, the preliminary profiles need more refinement and improvement by involving educators and stakeholders. The second limitation is that only one item block was used from the NAEP Grade 8 Math Assessment. Further work should explore the HAI framework that can create engagement profiles across multiple item blocks and overcome the challenge of feature differences in different item blocks. Further investigation also needs to explore alternative and explainable approaches (such as new process features and different machine learning algorithms) to better capture how human experts reason to create the engagement profiles.

Declarations

Gen-AI Use: The authors of this article declare (Declaration Form #: 2611241642) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Author Contribution: The first author led the study and contributed to conceptualization, methodology, data modeling, analysis, and visualization, interpretation, and writing. All the other authors played critical roles in shaping the study by contributing to concept, methodology, data annotation, interpretation, or revision.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Funding: This project has been funded at least in part with Federal funds from the National Center for Education Statistics in the U.S. Department of Education. The content of the publication does not necessarily reflect the views or policies of the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endowment of the U.S. Government.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., . . . Zheng, X. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Google. https://www.tensorflow.org/

Baker, R. (2021). *Artificial intelligence in education: Bringing it all together*. In S. Vincent Lancrin (Ed.), Pushing the frontiers with AI, blockchain, and robots (pp. 43–54). OECD.

Ercikan, K., Guo, H., & He, Q. (2020). Use of response process data to inform group comparisons and fairness research. *Educational assessment, 25(3),* 179–197. https://doi.org/10.1080/10627197.2020.1804353

Ercikan, K., Guo, H., & Por, H.-H. (2023). *Uses of process data in advancing the practice and science of technology-rich assessments.* Innovating Assessments to measure and support complex skills (N. Foster & M. Piacentini, Eds.). OECD Publishing. Retrieved from https://www.oecd-ilibrary.org/education/innovating-assessments-to-measure-and-support-complex-skills_7b3123f1-en

Ercikan, K., & Pellegrino, J. (2017). *Validation of score meaning in the next generation of assessments: The use of response processes*. Routledge.

Geron, A. (2017). *Hands-on machine learning with scikit-learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. Sebastopol, CA: O'Reilly Media.

Gordon, E. (2020). Toward assessment in the service of learning. *Educational Measurement: Issues and Practice, 39(3),* 72–78. Retrieved from https://doi.org/ 10.1111/emip.12370

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36-46.

Guo, H., & Ercikan, K. (2021a). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation, 26(5-6),* 302-327. https://doi.org/10.1080/13803611.2021.1963941

Guo, H., & Ercikan, K. (2021b). Using response-time data to compare the testing behaviors of English language learners (ells) to other test-takers (non-ells) on a mathematics assessment. *ETS Research Report, 2021(1),* 1-15. https://doi.org/10.1002/ets2.12340

Guo, H., Johnson, M., Ercikan, K., Saldivia, L. & Worthington, M. (2024). Large-scale assessments for learning: A huma-centered AI approach to contextualize test performance. *Journal of Learning Analytics*, *11(2), 229-245.* https://doi.org/10.18608/jla.2024.8007

Guo, H., Rios, J., Haberman, S., Liu, O., Wang, J. & Paek, I. (2017). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education. 29(3). 173 – 183. http://doi.org/10.1080/08957347.2016.1171766*

Johnson, M. S., & Liu, X. (2022). *Psychometric considerations for the joint modeling of response and process data [Paper presentation]*. International Meeting of Psychometric Society, Bologna, Italy.

Levy, R. (2020). Implications of considering response process data for greater and lesser psychometrics. *Educational Assessment, 25(3),* 218–235.

https://doi.org/10.1080/10627197.2020.1804352

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Miao, F., Holmes, W., Huang, R., & Zhang, H. (2021). *AI and education: Guidance for policymakers*. UNESCO.

National Assessment Governing Board. (NAGB, 2020). *Response process data from the 2017 NAEP grade 8 mathematics assessment*. https://www.nationsreportcard.gov/process_data/

National Assessment Governing Board (NAGB, 2024a). *Mathematics assessment framework for the 2022 and 2024 National Assessment of Educational progress.* Retrieved from https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/mathematics/2022-24-nagb-math-framework-508.pdf

National Assessment Governing Board (NAGB, 2024b). *How states use and value the Nation's Report Card*. Retrieved from https://www.nagb.gov/about-us/state-and-tuda-case-studies.html

National Center for Education Statistics. (NCES, 2022). *NAEP questions tool*. Retrieved from https://nces.ed.gov/NationsReportCard/nqt/

National Research Council. (NRC, 2000). *How people learn: Brain, mind, experience, and school: Expanded edition*. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/9853

National Academies of Sciences, Engineering, and Medicine (NASEM, 2018). *How people learn II: Learners, contexts, and Cultures*. Washington, DC: The National Academies Press.

Office of Educational Technology. (2023). *Artificial intelligence and the future of teaching and learning: Insights and recommendations (Report)*. Washington, DC, 2023: U.S. Department of Education.

Pellegrino, J. W. (2020). Important considerations for assessment to function in the service of education. *Educational Measurement: Issues and Practice, 39(3),* 81- 85. Retrieved from https://doi.org/10.1111/emip.12372

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. Science, 372(6540), 338-340. Retrieved from https://doi.org/10.1126/science.abd3300

Radwan, A. M. (2019). *Human active learning*. In S. M. Brito (Ed.), Active learning (chap. 2). Rijeka: IntechOpen. Retrieved from https://doi.org/10.5772/ intechopen.81371

Rios, J., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? an analysis of differential noneffortful responding on an international college-level assessment of critical thinking ISLA. *Applied Measurement in Education, 33(4),* 263–279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the impact of careless responding on aggregated-scores: To filter unmotivated examinees or not? *International Journal of Testing, 17(1)*, 74–104. https://doi.org/10.1080/08957347.2020.1789141

Rizve, M. N., Duarte, K., Rawat, Y. S., & Shah, M. (2021). *In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning*. In International conference on learning representations. Retrieved from https://iclr.cc/media/iclr-2021/Slides/3255.pdf

Wise, S. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. Educational Measurement: Issues and Practice, 36(4), 52–61. https://doi.org/10.1111/emip.12165

Wise, S. (2021). Six insights regarding test-taking disengagement. Educational Research and Evaluation, 26(5-6), 328–338. https://doi.org/10.1080/13803611.2021.1963942

Wise, S. & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18 (2),* 163 – 183. https://doi.org/10.1207/s15324818ame1802_2

Xie, Q., Dai, Z., Hovy, E. H., Luong, M., & Le, Q. V. (2019). Unsupervised data augmentation for consistency training. CoRR, abs/1904.12848. Retrieved from http://arxiv.org/abs/1904.12848

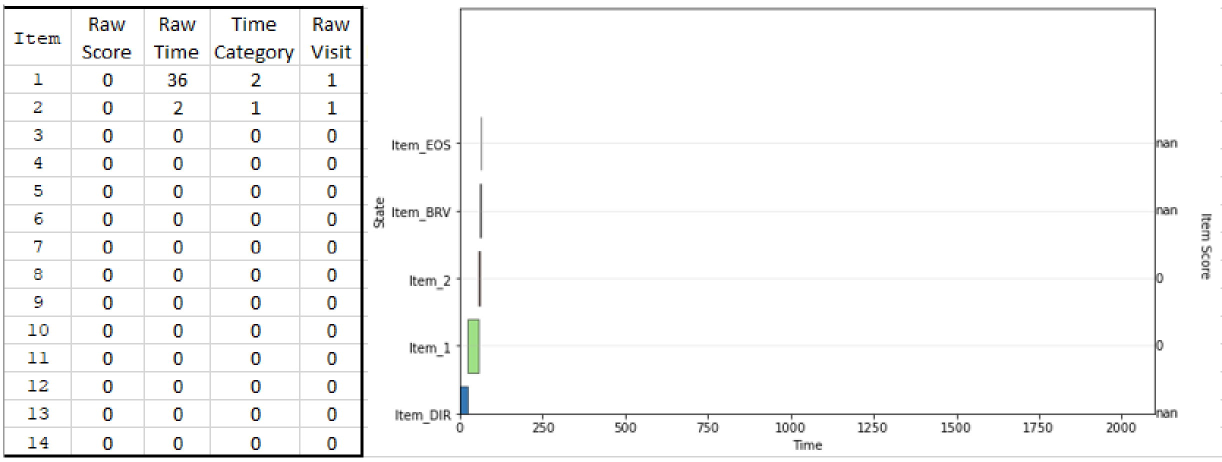
Zhu, X., Lafferty, J., & Ghahramani, Z. (2003). Combining active learning and semisupervised learning using gaussian fields and harmonic functions. In ICML 2003 workshop on the continuum from labeled to unlabeled data in machine learning and data mining (pp. 58–65).

Zoanetti, N., & Griffin, P. (2017). *Log-file data as indicators for problem-solving processes*. In B. Csapo & J. Funke (Eds.), The nature of problem solving: Using research to inspire 21st century learning (chap. 11). Paris: OECD Publishing. Retrieved from https://doi.org/10.1787/9789264273955-en

Appendix

Figure A1.

One instance in Profile 1 (unengaged)

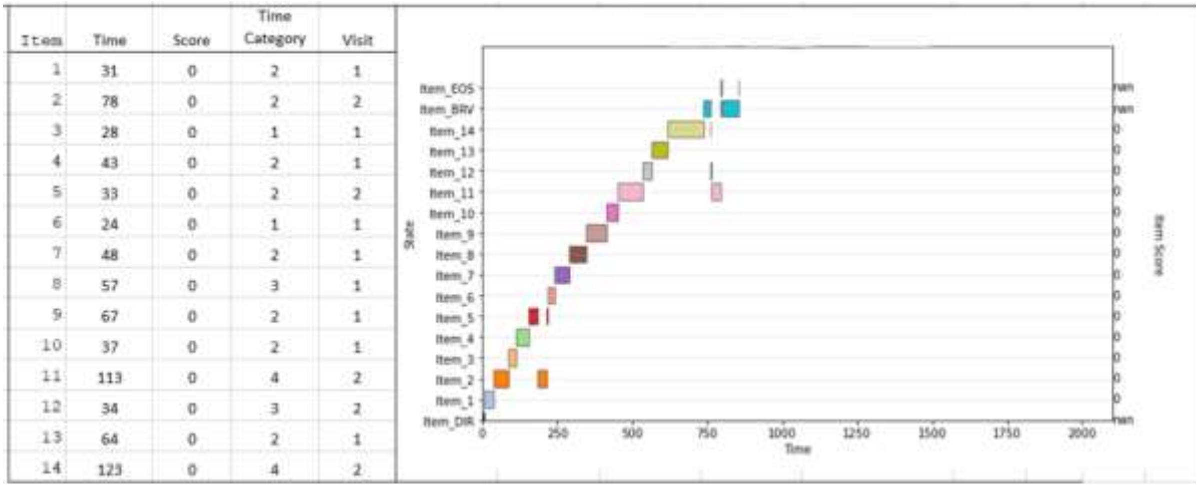


Item	Raw Score	Raw Time	Time Category	Raw Visit
1	0	36	2	1
2	0	2	1	1
3	0	0	0	0
4	0	0	0	0
5	0	0	0	0
6	0	0	0	0
7	0	0	0	0
8	0	0	0	0
9	0	0	0	0
10	0	0	0	0
11	0	0	0	0
12	0	0	0	0
13	0	0	0	0
14	0	0	0	0

Notes. *The student had a total score of 0, a total time of 38 seconds and a total number of visit states of 5. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student did not engage with the assessment.*

Figure A2.

One instance in Profile 2 (Low engagement with very low score)

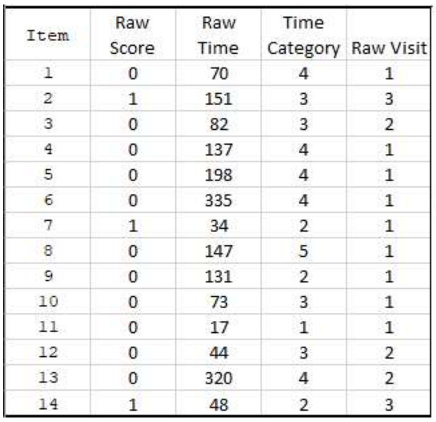


Item	Time	Score	Time Category	Visit
1	31	0	2	1
2	78	0	2	2
3	28	0	1	1
4	43	0	2	1
5	33	0	2	2
6	24	0	1	1
7	48	0	2	1
8	57	0	3	1
9	67	0	2	1
10	37	0	2	1
11	113	0	4	2
12	34	0	3	2
13	64	0	2	1
14	123	0	4	2

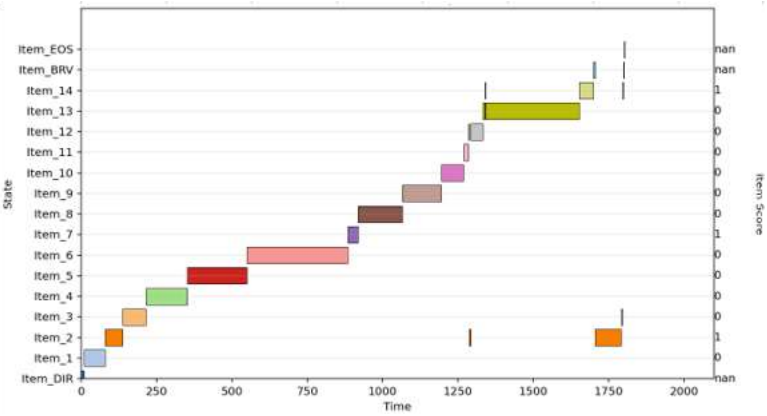
Notes. *The student had a total score of 0, a total time of 780 seconds and a total number of visit states of 25. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student did not engage with the assessment. The student worked through all the items but mostly without adequate effort.*

Figure A3.

One instance in Profile 4 (Low score, full/regular mixed time, and regular visit behavior)



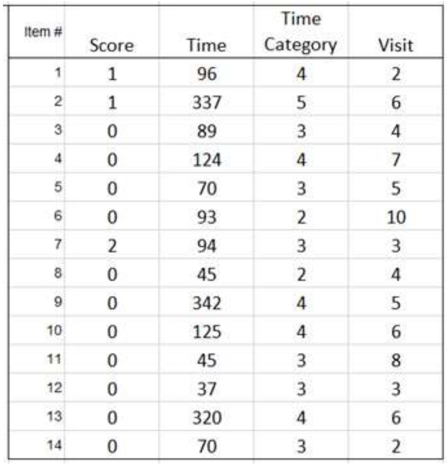
Item	Raw Score	Raw Time	Time Category	Raw Visit
1	0	70	4	1
2	1	151	3	3
3	0	82	3	2
4	0	137	4	1
5	0	198	4	1
6	0	335	4	1
7	1	34	2	1
8	0	147	5	1
9	0	131	2	1
10	0	73	3	1
11	0	17	1	1
12	0	44	3	2
13	0	320	4	2
14	1	48	2	3



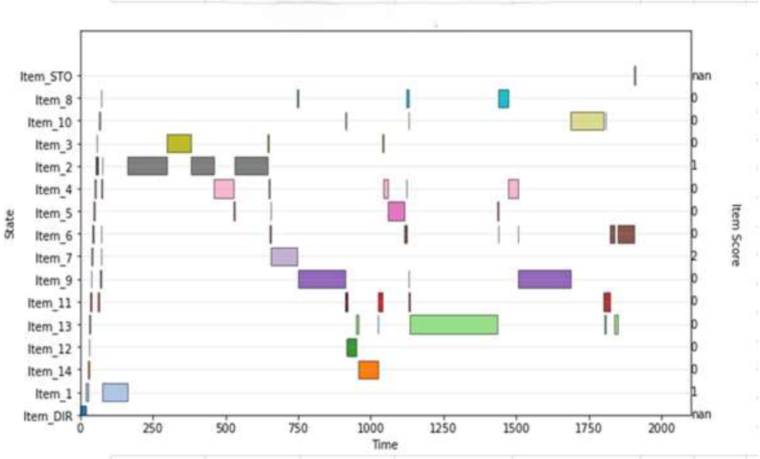
***Notes**. The student had a total score of 3 out of 21, a total time of 1790 seconds and a total number of visit states of 25. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student used nearly full time on the item block and adopted a mixed responding strategy (relatively prolonged time on Item 8 and relatively rapid responding on Item 11, for example).*

Figure A4.

One instance in Profile 5 (Low score, unregulated and/or speeded, with high visit behaviors)



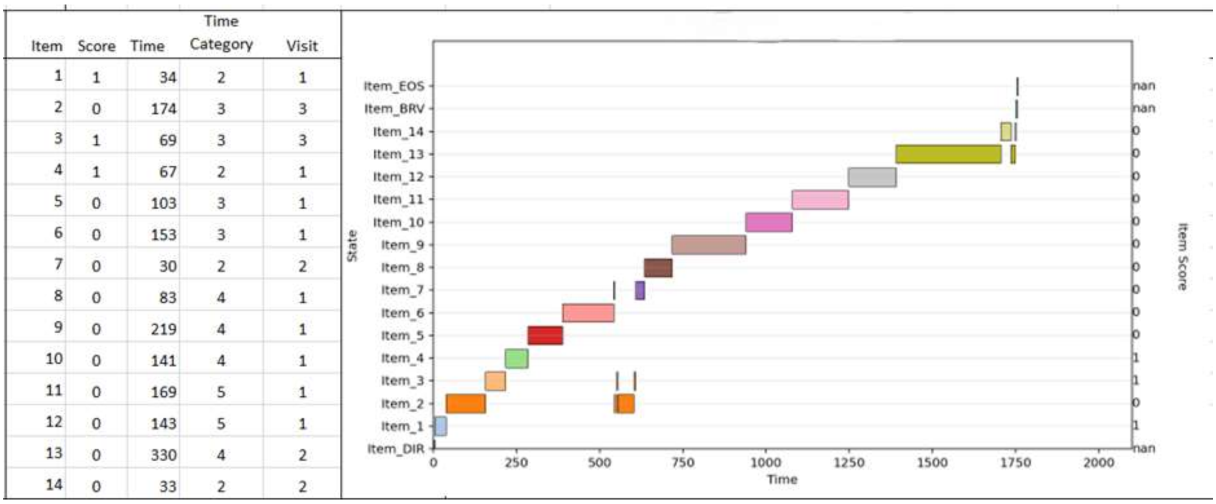
Item #	Score	Time	Time Category	Visit
1	1	96	4	2
2	1	337	5	6
3	0	89	3	4
4	0	124	4	7
5	0	70	3	5
6	0	93	2	10
7	2	94	3	3
8	0	45	2	4
9	0	342	4	5
10	0	125	4	6
11	0	45	3	8
12	0	37	3	3
13	0	320	4	6
14	0	70	3	2



***Notes**. The student had a total score of 4 out of 21, a total time of 1887 seconds and a total number of visit states of 72. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student visited items many times and irregularly, with a lot of item quick scanning behaviors and poor time management.*

Figure A5.

One instance in Profile 6 (Low score, full time with some prolonged item response times)



Item	Score	Time	Time Category	Visit
1	1	34	2	1
2	0	174	3	3
3	1	69	3	3
4	1	67	2	1
5	0	103	3	1
6	0	153	3	1
7	0	30	2	2
8	0	83	4	1
9	0	219	4	1
10	0	141	4	1
11	0	169	5	1
12	0	143	5	1
13	0	330	4	2
14	0	33	2	2

Notes. *The student had a total score of 3 out of 21, a total time of 1887 seconds and a total number of visit states of 72. The item summary information (item score, item response time, response time category, and number of item visits) is presented in the table on the left-hand side. The student's navigation pattern is presented in the plot on the right-hand side. This student visited items almost linearly with adequate or prolonged time effort on most items, but low performing.*

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(Special issue); 302-317

An Effect Analysis of the Balancing Techniques on the Counterfactual Explanations of Student Success Prediction Models

Mustafa CAVUS* Jakub KUZILEK**

Abstract

In the past decade, we have experienced a massive boom in the usage of digital solutions in higher education. Due to this boom, large amounts of data have enabled advanced data analysis methods to support learners and examine learning processes. One of the dominant research directions in learning analytics is predictive modeling of learners' success using various machine learning methods. To build learners' and teachers' trust in such methods and systems, exploring the methods and methodologies that enable relevant stakeholders to deeply understand the underlying machine-learning models is necessary. In this context, counterfactual explanations from explainable machine learning tools are promising. Several counterfactual generation methods hold much promise, but the features must be actionable and causal to be effective. Thus, obtaining which counterfactual generation method suits the student success prediction models in terms of desiderata, stability, and robustness is essential. Although a few studies have been published in recent years on the use of counterfactual explanations in educational sciences, they have yet to discuss which counterfactual generation method is more suitable for this problem. This paper analyzed the effectiveness of commonly used counterfactual generation methods, such as WhatIf Counterfactual Explanations, Multi-Objective Counterfactual Explanations, and Nearest Instance Counterfactual Explanations after balancing. This contribution presents a case study using the Open University Learning Analytics dataset to demonstrate the practical usefulness of counterfactual explanations. The results illustrate the method's effectiveness and describe concrete steps that could be taken to alter the model's prediction.

Keywords: explainable artificial intelligence, actionable explanations, imbalance learning, educational data mining, learning analytics

Introduction

For centuries universities have been collecting information about their students. With the rise of Information and Communication Technologies (Eurostat, 2023), the information collected and stored is transformed from paper-based collections to digital domains (Hilbert and López, 2011). The introduction of new digital education formats and the information collection shift resulted in storing vast amounts of student and study-related data including student demographics, assessment, learning design, and context. In combination with the advancement in Data Mining and Machine Learning (ML) research (LeCun et al., 2015; Vaswani, 2017), the collected data enabled new research exploring the educational domain. The most prominent research fields are Educational Data Mining (EDM) and Learning Analytics (LA), which explore the educational domain from two different perspectives (Siemens and Baker, 2012). More recently, the concerns about the use of Artificial Intelligence (AI) have become stronger uncovering the limitations and possible problems such as bias and explainability of models developed (Singer, N., 2014). As a consequence, new data and AI regulations such as the General Data

*Asst. Prof., Department of Statistics, Eskişehir Technical University, Eskişehir-Türkiye, mustafacavus@eskisehir.edu.tr, ORCID ID: 0000-0002-6172-5449
**Dr., Humboldt University of Berlin, Berlin-Germany, jakub.kuzilek@hu-berlin.de, ORCID ID: 0000-0002-8656-0599

To cite this article:
Çavuş, M. & Kuzilek, J. (2024). An Effect Analysis of the Balancing Techniques on the Counterfactual Explanations of Students Success Prediction Models. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 302-317. https://doi.org/10.21031/epod.1526704

Received: 1.08.2024
Accepted: 26.12.2024

Protection Regulation (GDPR[1]) and the Artificial Intelligence Act (AI Act[2]) in the EU have been established (Hoel et al., 2017). As a consequence trust in the analytical tools and AI methods in higher education has been reduced leading to the new approach in LA research called Trusted Learning Analytics (TLA) (Drachsler H., 2018). The TLA approach focuses on using intrinsically explainable `white box` AI models and systems. This significantly reduces the opportunity of using more "user-unfriendly" models such as Random Forests (RF) or Neural Networks. Luckily, the field of Explainable Artificial Intelligence (XAI) (Molnar, C., 2020) provides researchers with methods with the potential to unlock the `black box` models for use in TLA systems (Drachsler H., 2018). The trend of using XAI methods in the educational domain is highly resonating within the research community resulting in more research in the area in recent years (e.g., Human-Centric eXplainable AI in Education Workshop at 17th Educational Data Mining Conference[3]).

There are various tasks within the LA that focus on supporting learners and educators using various tools and methods. However, one of the most common objectives is the predictive modeling of learner success (with varying definitions of success), which focuses on the identification of the learners in need of help with their studies (Papamitsiou and Economides, 2014). Within the task of success prediction, the legacy learner and learning data are utilized for training the prediction model using the ML algorithm (Arnold and Pistilli, 2012; Waheed et al, 2020; Adnan et al., 2021). From the LA point of view, the prediction delivered by the ML model is used as a trigger for educational intervention. Thus the model itself is used as a tool by the lecturer, teaching assistant, or anyone responsible for supporting the students. Yet, there is a constant demand for providing not just the prediction itself, but also the "reasons behind the model decision" (Kuzilek et al., 2015). At this stage, again, the XAI comes into play and fosters the objectives of TLA (Drachsler H., 2018).

In the context of ML, predictive models pursue the highest predictive accuracy. The so-called `black-box` models frequently perform best, sacrificing the understanding of reasoning to deliver a concrete prediction. Thus, `black-box` models are preferred over the so-called `white-box` models, which, in addition to the prediction, provide intrinsically interpretable predictions. (Guidotti et al., 2018; Biecek et al., 2021; Holzinger et al., 2022). However, to enable the power of XAI for the `black-box` models the post-hoc methods can be used (Pinto and Paquette, 2024). The XAI methods are primarily categorized into global and local. At the global level, they reveal which variables are important in the model. In contrast, at the local level, they answer questions about the contributions of variables in generating individual predictions (Molnar et al., 2020; Cavus et al., 2023). However, commonly used global and local tools, while sufficient for understanding the prediction made for a particular observation, are not sufficient for generating a counterfactual understanding of an undesirable outcome. Commonly used XAI methods (both local and global) are adequate for understanding particular observation predictions and not for generating a counterfactual understanding of an undesirable outcome (e. g. negative class in a binary classification problem).

To improve understanding of the undesirable outcome the method of counterfactual explanations (CEs) has become popular. CEs are defined as the minimal change in the variable values to flip the model's prediction into the intended outcome (Artelt and Hammer, 2019). In the frame of learner success prediction, the models may indicate an unfavorable outcome, but they do not provide recommendations

[1] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons concerning the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A32016R0679

[2] Proposal for a Regulation of the European Parliament and of the Council laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts https://eur-lex.europa.eu/legalcontent/EN/TXT/?uri=CELEX%3A52021PC0206

[3] https://hexed-workshop.github.io/

to reverse the learner situation. Counterfactual explanations provide the extension of the baseline model and provide such a recommendation by highlighting necessary changes in the learner profile to reverse the negative outcome. Learners, teachers, and curriculum designers can be guided toward actions or measures to be taken through their generated explanations.

The use of counterfactual explanations in LA has been explored in several studies (Cavus and Kuzilek, 2024; Tsiakmaki et al., 2021; Zhang et al., 2023; Afrin et al., 2023). All of the research works focused on providing a frame for delivering actionable insights to relevant stakeholders using the CE. Facing numerous counterfactual explanations due to the nature of optimization problems requires selecting those explanations that fulfill specific criteria beneficial for the stakeholder. Each learner requires personalized counterfactuals because of their background, challenges, and differences in needs (Smith et al., 2022).

The research presented in this paper focuses on using CE measures for the evaluation of the effect of balancing techniques used on the raw imbalanced dataset. More specifically we focus on the following research questions:

RQ1: *What is the most appropriate method for generating the counterfactual explanations after balancing?*

RQ2: *How do balancing techniques affect the counterfactual explanations of student success prediction models?*

This study compares the qualities of different counterfactual generation methods for students whose success prediction model developed after balancing the training dataset anticipates failing. For the reproducibility of the developed approach, we used the Open University Learning Analytics Dataset (OULAD) (Kuzilek et al., 2017) as a raw data source. The study is essential in two ways: (1) because the missing evaluation of the counterfactual quality can lead to inefficient explanations, and this may compromise their trustworthiness (Artelt et al., 2021), (2) there is no uniformly better method for each domain (Dandl et al., 2023) and this is the first benchmark in the domain of LA, and (3) there are no many investigations on the effect of balancing methods on the counterfactual explanations (Gunonu et al, 2024).

The rest of the paper is organized using the following analysis approach. It examines the effect of balancing strategies on the quality of counterfactuals generated by the three most commonly used methods. Finally, the results are discussed.

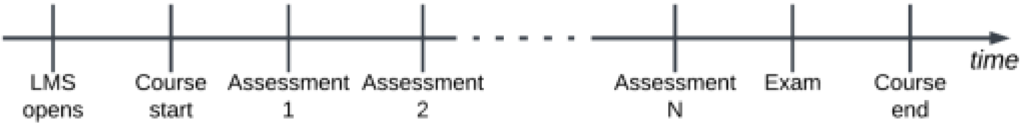
Method

This section contains details of the dataset, counterfactual explanations, resampling methods, and the experimental design.

Data

The OULAD dataset has been released by the Open University (OU). The OU is the largest distance-learning institution in the UK. It is utilized to analyze the impact of the balancing strategies on the counterfactual generation methods. The typical course duration at the Open University is nine months and includes multiple assignments and a final exam. The most important assignments are Tutor Marked Assignments (TMAs), which represent critical milestones throughout the course. Fig. 1 presents the timeline of the typical Open University course.

Figure 1.
The OU course timeline



The course registration opens several months before the course starts. The registration process involves several batch enrollment rounds, during which the students eligible to take the course are enrolled. In addition, students can register for the course by themselves. The interaction with the Moodle-like Learning Management System (LMS) starts up to four weeks before the official course starts. The students can test the course contents and decide if the course is worth attending. Since LMS opened student interactions in the form of daily aggregated click-stream logs were recorded. During the course, several assessments evaluate the gained knowledge. Before the official end of the course, the exam is scheduled. Students can deregister from the course at any time. The information about student interactions, demographics, assessment results, and course outcomes forms the core of the OULAD dataset.

For the analysis, the STEM course DDD and its 2013J and 2014J presentations studied by 3741 students have been selected. The course includes six TMAs. The final student result was used as the target variable for model training. Students with the result "Distinction" have been merged with students with the result "Pass". Reducing the prediction task to binary classification to classes: "Pass" and "Fail". We excluded the actively withdrawn students (n = 1328). The resulting dataset includes data from 2296 students.

The previous research with the OULAD and Open University data showed that the importance of the demographics is significantly reduced after the first LMS click-stream is recorded and included in the prediction modeling (Kuzilek et al., 2015). The first TMA has been identified as a strong predictor of student success in the course (Kuzilek et al., 2015). Thus, the importance of interaction data at the beginning of the course is even greater since they are strong predictors not just for the outcome prediction but also for the first TMA prediction (Kuzilek et al., 2015). In addition, the nature of the learning context of the Open University produces specific learning patterns within the student cohort, where most students prefer to study only in specific periods (Kuzilek et al., 2017). These periods tend to have a weekly repetition pattern. Thus, it makes sense to focus on weekly aggregated click-stream data.

The resulting dataset consists of 42 predictors, numerical variables containing the weekly summary of online interactions with the LMS, and the target variable representing the outcome for the student from the course. Table 1 provides descriptions of the selected variables.

Table 1.
The description of the variables used to train the student success prediction model

Variable	Description	Class	Value
final_result	student's final exam result	categorical	$\{Fail, Pass\}$
week_minus4	the number of clicks four weeks before the course starts	numeric	$[0, 493]$
week_minus3	the number of clicks three weeks before the course starts	numeric	$[0, 765]$
week_minus2	the number of clicks two weeks before the course starts	numeric	$[0, 745]$
week_minus1	the number of clicks one week before the course starts	numeric	$[0, 987]$
week_0	the number of clicks before the course starts	numeric	$[0, 1319]$
week_1	the number of clicks one week after the course starts	numeric	$[0, 525]$
…	…	…	…
week_37	the number of clicks thirty-seven weeks after the course starts	numeric	$[0, 50]$

Counterfactual Explanations

Counterfactual explanations (CE) illustrate "what-if" scenarios that emphasize the necessary alterations to the input data to change a model's output (Watcher et al., 2017). $X = [x_1, x_2, \ldots, x_p]$ represent a data matrix with n observations and p variables and y be the response vector. The objective is to identify a function $f: X \rightarrow y$ that minimizes the expected value of the loss function L in predictive modeling. A counterfactual $x' \epsilon R^p$ of observation $x \epsilon R^p$ is determined by solving the following optimization problem:

$$argmin_{x' \epsilon R^p} \, L[f(x'), y'] + d(x, x') \qquad (1)$$

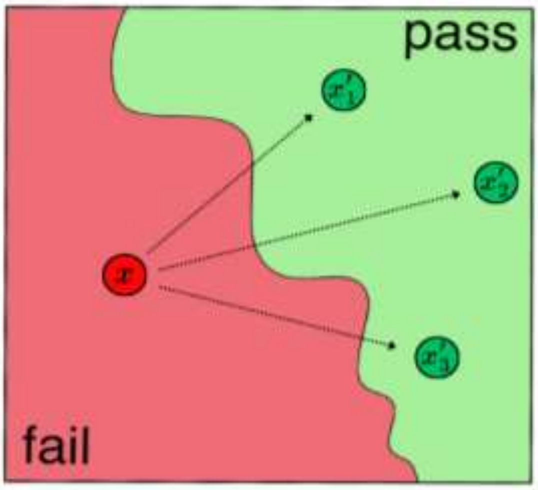
where R^p represents p-dimensional real space, L is a loss function that penalizes the difference between the prediction $f(x')$ and the desired outcome y', and d is a distance function between the observation x and x'. A CE specifies the necessary adjustments in one or more variables to change the model's prediction. The distance function d regulates the proximity between the original observation and the counterfactual.

Figure 1 visualizes an observation and its counterfactuals. Assume that f is a student success prediction model and x is a vector consists the variable values of a student. The prediction of the model f for the student x who has failed. The red zone shows the fail area, and the green one shows the pass area. They are divided by the decision boundary of the model. The CEs x'_1 ,x'_2 , x'_3 represent the ways how the student can pass.

Counterfactuals strive to minimize the distance between the target observation and the counterfactual; however, additional properties are essential for a counterfactual explanation (Wachter et al., 2017; Karimi et al., 2020). **Sparsity** suggests altering the minimal number of variables to keep the explanation straightforward. **Minimality** aims for the most minor feasible changes in the variable values. **Validity** is ensured by reducing the difference between the counterfactual instance x' and the original observation x while ensuring the model's output matches the desired label y'. **Proximity** emphasizes the necessity for a slight variation between the factual and counterfactual features. **Plausibility** requires that counterfactual explanations remain realistic and closely follow the underlying data distribution. Over 120 known counterfactual generation methods; see Warren et al. (2023) for further details. However, we focused on three widely used counterfactual methods *WhatIf Counterfactual Explanations*, *Multi-Objective Counterfactual Explanations*, and *Nearest Instance Counterfactual Explanations* to facilitate the comparison of counterfactual quality.

Figure 2.

The counterfactual explanations for an observation



What-if counterfactual explanations. The What-if method (WhatIf) finds the observations closest to the observation x from the other observations in terms of Gower distance, solving the following optimization problem (Wexler et al., 2019):

$$x' \in argmin_{x \in X}\, d(x, x') \qquad (2)$$

Multi-objective counterfactual explanations. The Multiobjective Counterfactual Explanations (MOC) method aims to generate counterfactual explanations by optimizing multiple objectives simultaneously (Dandl et al., 2020). These objectives often include validity, proximity, sparsity, and plausibility.

$$x' \in min_x(o_v(\hat{f}(x), y'), o_p(x, x'), o_s(x, x'), o_{pl}(x, X)) \qquad (3)$$

where o_v, o_p, o_s, o_{pl} are the objective functions for the desired properties *validity*, *proximity*, *sparsity*, and *plausibility*, respectively. Thus, it is expected that the counterfactuals generated by the MOC method are valid, proximity, sparse, and plausible.

Nearest instance counterfactual explanations. The Nearest Instance Counterfactual Explanations (NICE) method identifies observations that are most similar to a given observation using the heterogeneous Euclidean overlap method (Burghmans et al., 2023). This approach allows for two options in the objective function, depending on the properties of *proximity* and *sparsity*, offering flexibility in how it can be applied.

The WhatIf method produces counterfactuals that are valid, proximal, and plausible. It has been demonstrated that the MOC method generates a higher number of counterfactuals that are closer to the training data and require fewer feature changes compared to other counterfactual methods (Dandl et al., 2020). Additionally, NICE specifically generates proximity-focused counterfactuals. However, no single method consistently outperforms others across datasets from various domains (Dandl et al., 2023). Therefore, evaluating the quality of the generated counterfactuals is essential, and we will conduct experiments to evaluate this in the following section.

Balancing Techniques

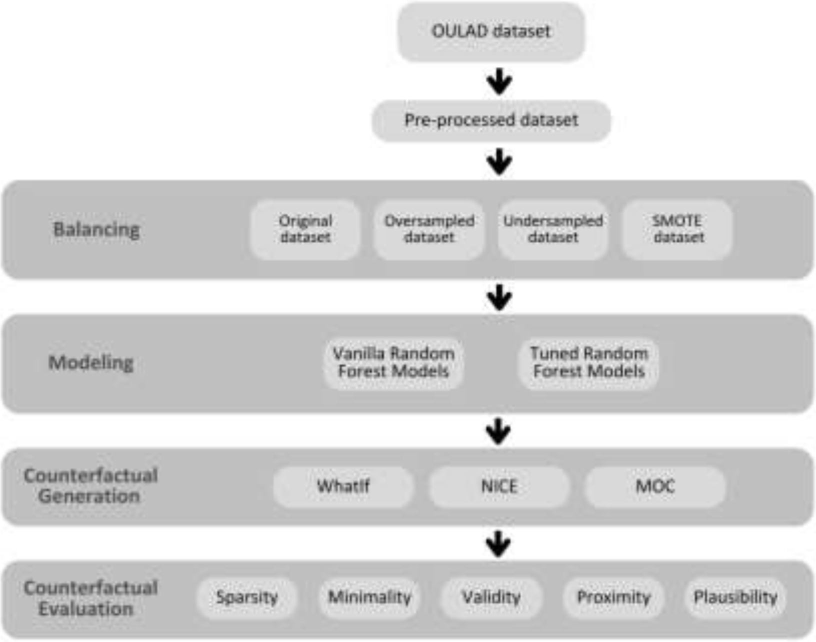
The most commonly encountered challenge in designing predictive models with high discriminatory performance is an imbalanced class distribution in the response variable. In the binary case, the imbalance problem occurs when one class is observed less frequently. Models with such response variables tend to be biased toward the majority class in their predictions. Consequently, when dealing with the imbalance problem, models often have a significantly lower performance in correctly predicting the minority class than the majority class. In real-world problems, the class of interest is generally the minority class. For example, in predicting student dropouts, students who drop out are observed less frequently than those who do not. In the classification problem of predicting whether a student will complete a specific educational material or content module, students who do not complete the material are observed less frequently than those who do. In learning analytics, when considering student success prediction models, students who fail are observed less frequently than those who succeed. In these examples, students who drop out, do not complete educational materials and fail constitute the minority class. Due to the nature of these problems, the focus is on identifying the minority class. The inaccurate models in correctly predicting the minority class is a problem that must be overcome in such scenarios.

Solutions to this problem are divided into three categories: data-based, model-based, and weighting-based methods. The most commonly used data-based methods involve balancing class distributions through random undersampling or oversampling and synthetic data generation techniques. In undersampling, a subset of the majority class is randomly selected to match the minority class, whereas, in oversampling, the number of observations in the minority class is increased through resampling to match the size of the majority class (Chawla, 2010). In synthetic data generation methods, new observations are artificially generated from the distribution of the minority class to balance the size with the majority class (Elyan et al., 2021; Liu, 2022). Model-based methods are specific models developed to address the imbalance problem (Yin et al., 2020; Gu et al., 2022). Weighting-based methods aim to achieve higher prediction performance by penalizing the model more for errors in predicting the minority class (Zong et al., 2013; Tao et al., 2019). Although there are many methods to solve the classification problem in unbalanced data, in recent years, it has been found that these methods generally need to be revised and have adverse effects on classification models (Junior and Pisani, 2022; Stando et al., 2024; Cavus and Biecek, 2024; Carriero et al., 2024). These criticisms, mainly focusing on oversampling, undersampling, and synthetic data generation methods, brought the cost-sensitive approach to the fore (Gunonu et al., 2024). This study used data-based and weighting-based methods due to the mentioned criticism, their practical applications, and their frequent usage in the literature.

Experimental Design

This paper focuses on which method provides the highest quality counterfactual explanations for the student success prediction model trained with and without hyperparameter tuning (i.e., vanilla model) regarding the imbalancedness problem using the OULAD dataset. Thus, the approach followed, which is given in Figure 2, is (1) balancing the dataset, (2) training the model with and without hyperparameter tuning, (3) generating the counterfactuals, and (3) evaluating the effect of the balancing techniques of the imbalancedness problem producing the evaluation criteria.

Figure 2.
The flow of the experiments



Balancing. Two balancing strategies are used. The dataset is balanced using several resampling methods such as undersampling, oversampling, and SMOTE, and the models are trained on the original dataset with the cost-sensitive approach.

Modeling. The random forest algorithm is used in modeling because tree-based models exhibit lower prediction performance than alternative complex models in classifying tabular datasets (Grinsztajn et al., 2022). It is trained with and without hyperparameter tuning to achieve higher prediction performance. The performance of the random forest models trained on imbalanced (i.e., Original), balanced datasets by the oversampling, undersampling, SMOTE, and also trained with the cost-sensitive approach are compared. The costs are chosen as 2.37931 for the minority class (i.e., failed students) and 1 for the majority class regarding the imbalance ratio. Moreover, to achieve better predictive performance the models are tuned in terms of hyperparameters mtry, splitrule, and min.node.size using the 10-fold repeated cross-validation in addition to the vanilla versions of the model which is trained with the default values of the hyperparameters.

Counterfactual generation. After the modeling phase, the counterfactuals are generated for failing students which are estimated by the models using MOC, sparsity-based NICE (NICE_sp), proximity-based NICE (NICE_pr), and What-If methods.

Results

In this section, the results are summarized. Firstly, the performance of the models is compared, and then the counterfactuals are evaluated to determine the best counterfactual generation method for the case considered in the paper.

Model performance. The performance of the random forest models trained on imbalanced and balanced datasets by the oversampling, undersampling, SMOTE, and cost-sensitive approach are given in Table 2. Accuracy, Area Under Curve (AUC), and F1 score are used to measure the model performance.

Accuracy represents the proportion of correct predictions made by the model out of all predictions. The AUC is a single number representing the area under the Receiver Operating Curve (ROC), ranging from 0 to 1. An AUC of 1 means the classifier perfectly distinguishes between positive and negative classes. The F1 score shows that the model correctly predicts all positive instances and doesn't produce false positives. The imbalance ratio of the test set is 2.41 (number of observations in the majority class/number of observations in the minority class), thus the performance evaluations should be using the F1 score as well as accuracy and AUC.

The vanilla Random Forest models generally outperform tuned models in terms of accuracy and F1 scores across most balancing strategies, particularly on original data and some resampling methods. Vanilla models demonstrate higher accuracy and more balanced F1 scores, especially under oversampling and SMOTE techniques. On the other hand, tuned models achieve slightly higher AUC values with cost-sensitive learning and SMOTE, indicating better classification discrimination. Sampling methods like oversampling and SMOTE improve performance for both vanilla and tuned models, while undersampling tends to decrease accuracy and F1 scores but maintains stable AUC values. Cost-sensitive learning offers balanced improvements, with both model types benefiting from enhanced AUC scores. Overall, while vanilla models excel in accuracy and F1 scores, tuned models show enhanced AUC values in specific conditions, highlighting the trade-offs between different performance metrics and modeling approaches. The tuned values of the hyperparameters for the models are given in Table A in the Appendix.

Table 2.
The performance of the random forest models on the test set over balancing strategies

	Vanilla Random Forests Models			Tuned Random Forests Models		
	Accuracy	AUC	F1	Accuracy	AUC	F1
Original	0.8196	0.8549	0.7040	0.8044	0.8480	0.6450
Oversampling	0.8402	0.8652	0.6840	0.8366	0.8658	0.6795
Undersampling	0.7741	0.8552	0.6560	0.7812	0.8558	0.6611
SMOTE	0.8286	0.8620	0.6900	0.8321	0.8621	0.6907
Cost-sensitive	0.8357	0.8643	0.6940	0.8339	0.8671	0.6910

Counterfactual evaluations. The counterfactual generation methods can generate more than one explanation for an observation, also each method may generate different explanations. The number of counterfactuals generated by the methods is given in Table 3. The MOC generates the highest number of counterfactuals independently from the balancing strategy and model while the NICE methods generate the lowest number of counterfactuals. The differences between the number of counterfactuals between the balancing strategies depend on the number of students that were predicted as failed by the models. The number of counterfactuals for the models is slightly different because of the difference between the models caused by the hyperparameter optimization.

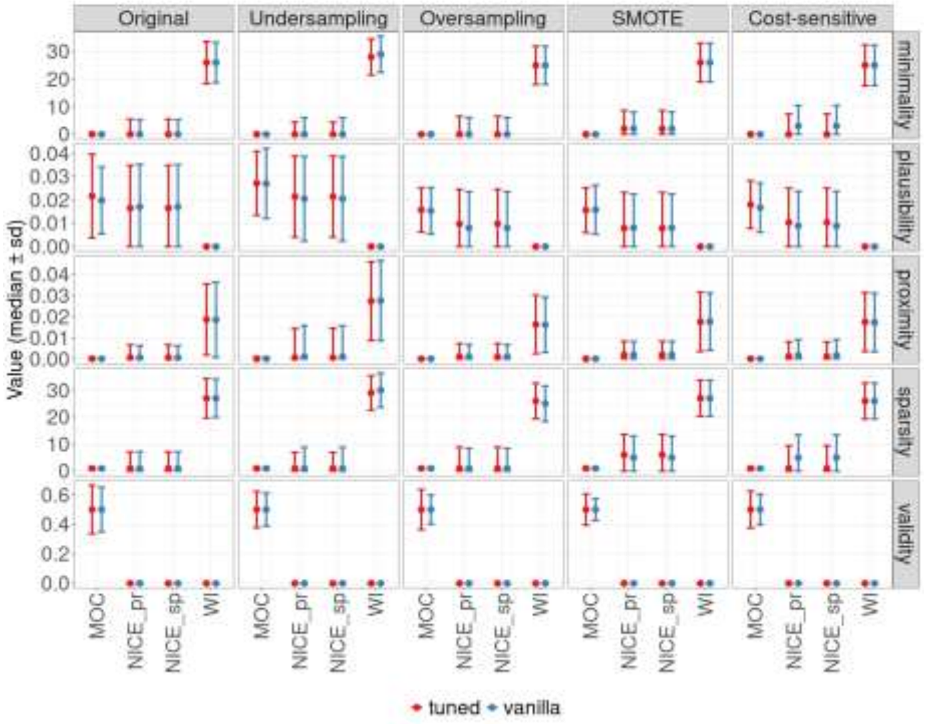
Table 3.
The number of counterfactuals generated by the methods across balancing strategies

	Model	Original	Undersampling	Oversampling	SMOTE	Cost-sensitive
WI	vanilla	2910	4050	2370	2890	2730
	tuned	2890	3950	2430	2800	2740
MOC	vanilla	23321	28287	15934	24100	19570
	tuned	24932	38262	15997	23687	19401
NICE_sp	vanilla	419	555	320	360	339
	tuned	390	530	327	336	530
NICE_pr	vanilla	419	555	320	360	339
	vanilla	2910	4050	2370	2890	2730

It is necessary to evaluate the quality or usefulness of the counterfactuals before deployment. Thus, we conduct a comparison study to analyze the effect of the conditions regarding the balancing and modeling strategies on the counterfactual quality. We aim to determine the best counterfactual generation method to find actionable insights from the student success prediction models trained on the OULAD dataset. The quality of counterfactuals is visualized using error bar plots as in Figure 3. An error bar plot shows the variability or uncertainty of data. It features error bars that extend above and below the median of the observations. Error bars can show measures of dispersion such as standard deviation, standard error, or confidence intervals, providing a visual indication of the reliability and precision of the data. Figure 3 demonstrates that each method exhibits varying performance regarding quality metrics across different balancing and modeling strategies. The error bars represent the median ± standard deviation, reflecting the variability in the performance of different counterfactual methods across various datasets and balancing techniques. The width of these error bars indicates how robust (or consistent) each method is in different scenarios.

Figure 3.

Evaluation of counterfactual generation methods across tuning and balancing strategies



NICE_sp and NICE_pr consistently demonstrate superior performance with the models trained on the original dataset. The minimality and plausibility values are particularly low, with medians near 0 and minimal variability, suggesting robust performance. On the other hand, MOC and WI show much higher values, especially in minimality where median values reach around 30, indicating suboptimal outcomes. Similarly, in metrics like proximity and sparsity, NICE_sp and NICE_pr maintain low values, whereas MOC and WI exhibit considerably higher values, suggesting that these methods struggle with the original data distribution.

When applying the Undersampling method, there is a general improvement in minimality values across all methods, though MOC and WI still trail behind NICE_sp and NICE_pr. While NICE_sp and NICE_pr continue to perform well with relatively low values across all metrics, the error bars suggest a slight increase in variability. MOC and WI, although showing some improvement, still exhibit higher plausibility and proximity values, indicating that undersampling does not fully mitigate their performance issues.

The Oversampling method highlights the strengths of NICE_sp and NICE_pr even further. These methods maintain low values across all metrics, particularly in minimality and plausibility, where their performance remains nearly flawless with median values close to 0. In contrast, MOC and WI continue to struggle, showing higher values across metrics such as proximity and sparsity, with only marginal improvements compared to the Original and Undersampling strategies. This suggests that while oversampling enhances performance for NICE_sp and NICE_pr, it does not sufficiently benefit MOC and WI.

Moving to SMOTE, NICE_sp, and NICE_pr once again emerge as the top performers, maintaining low values across all metrics. The proximity and sparsity values for these methods remain minimal, reflecting strong and consistent performance. MOC and WI, however, continue to display higher values in metrics like minimality and validity, suggesting that even with synthetic data generation, these methods are less effective. The error bars for MOC and WI also indicate greater variability, reinforcing the idea that SMOTE does not significantly improve their robustness.

Finally, the cost-sensitive approach shows that NICE_sp and NICE_pr maintain their strong performance, with median values remaining low across all metrics. Particularly in minimality and plausibility, these methods exhibit near-perfect performance, with minimal error bars indicating consistent results. MOC and WI show slight reductions in their median values for some metrics, but they still lag significantly, with higher values in proximity and sparsity indicating ongoing performance issues. The consistent superiority of NICE_sp and NICE_pr across different balancing strategies, including Cost-sensitive approaches, underscores their robustness and reliability.

Tuned models consistently show improved performance compared to their vanilla counterparts across various balancing strategies. Tuned models trained on the original dataset exhibit lower minimality and plausibility values, indicating enhanced performance. In the Undersampling strategy, the gap between tuned and vanilla models narrows slightly, but tuned models still outperform vanilla ones. With Oversampling and SMOTE, the advantage of tuning becomes more pronounced, as tuned models maintain lower values across all metrics, while vanilla models show higher variability. The cost-sensitive approach further highlights the superiority of tuned models, particularly in minimality and validity, where they consistently demonstrate lower values and greater consistency. Overall, tuning leads to better and more reliable performance across different data conditions and metrics.

When focusing on RQ1: "*What is the most appropriate method for generating counterfactual explanations after balancing?*" the analysis highlighted the consistent superiority of NICE_sp and NICE_pr across various balancing strategies and metrics, demonstrating their robustness and reliability. To answer RQ2: "*How do balancing techniques affect the counterfactual explanations of student success prediction models?*" we find out that the impact of different data balancing strategies, such as SMOTE and the cost-sensitive approaches, further underscores the adaptability of these methods compared to MOC and WI, which generally underperform. Additionally, tuned models outperform their vanilla counterparts across all conditions, emphasizing the importance of model optimization in achieving optimal performance across diverse balancing strategies.

Conclusion

This study explored the impact of various balancing techniques on the generation of counterfactual explanations within student success prediction models. Our analysis reveals that **NICE_sp** and **NICE_pr** consistently outperform other counterfactual explanation methods across various balancing strategies, including Original, Undersampling, Oversampling, SMOTE, and Cost-sensitive approaches. These methods demonstrate superior performance in terms of key metrics like minimality, plausibility, proximity, sparsity, and validity, showing lower variability (narrower error bars) and higher robustness across different datasets. This consistent superiority indicates that NICE_sp and NICE_pr are more reliable and effective in generating high-quality counterfactual explanations, regardless of the balancing strategy applied. The results indicate that the choice of balancing strategy significantly influences the quality and characteristics of the counterfactuals generated by different methods, such as Multi-Objective Counterfactual Explanations (MOC), Nearest Instance Counterfactual Explanations (NICE), and WhatIf.

Effectiveness of balancing techniques. The results suggest that certain balancing techniques improve the validity and plausibility of counterfactuals, aligning them more closely with realistic scenarios that educators and students can act upon. For example, balancing methods that mitigate class imbalances not only enhanced the performance of the predictive models but also resulted in more actionable and sparse counterfactual explanations. These findings are consistent with previous research, which emphasizes the importance of balancing in training robust models for educational predictions (Artelt et al., 2021).

Effect analysis of counterfactual generation methods. Among the methods tested, MOC consistently produced counterfactuals that were closer to the original data distribution, showing a higher degree of plausibility and sparsity. This is particularly valuable in educational settings where changes to multiple variables might not be feasible. In contrast, the NICE method, which focuses on proximity, often generated explanations that were more straightforward but potentially less realistic. This trade-off highlights the need to select counterfactual generation methods based on the specific requirements of the educational context.

Implications for educational interventions. The insights gained from this study have significant implications for how educational institutions might use counterfactual explanations to inform interventions. By understanding how different balancing techniques affect the characteristics of counterfactuals, educators can better choose models and explanations that are not only accurate but also actionable and interpretable for students and staff.

This study contributes to the growing field of explainable artificial intelligence in education by demonstrating the critical role of balancing techniques in generating effective counterfactual explanations. These findings pave the way for more refined and targeted educational interventions, ultimately contributing to more personalized and supportive learning environments.

Limitations and Future Work

While this study provides a comprehensive analysis, some limitations warrant further investigation. The focus on a single dataset and specific counterfactual methods may limit the generalizability of the results. Future research should explore these effects across different datasets containing educational data with similar and different contexts (López-Pernas, 2024); and additional machine learning such as neural networks or support vector machines (Murphy, K., 2022) and counterfactual methods (Guidotti, R., 2022). Moreover, the long-term impact of using such explanations on student outcomes should be evaluated to better understand their practical utility in educational settings. This involves conducting the research study with the lecturers and learners on the usability and acceptance of the method together with the evaluation of the learning gains and study outcomes similar to the studies conducted to evaluate the influence of the predictive modeling on student outcomes (e. g., Herodotou, 2019).

Declarations

Gen-AI use : The authors of this article declare (Declaration Form #: 2611241720) that Gen-AI tools have NOT been used in any capacity for content creation in this paper.

Acknowledgments: The work in this paper is supported by the German Federal Ministry of Education and Research (BMBF), grant no. 16DHBKI045.

Conflict of Interest: No potential conflict of interest was reported by the authors.

Ethical Approval: It is declared that all ethical guidelines for authors have been followed by all authors. Ethical approval is not required as this paper uses data shared with the public.

Supplemental Materials

The materials for reproducing the experiments performed and the dataset are accessible in the following anonymized repository: https://github.com/mcavs/JMEEP_paper.

References

Adnan, M., Habib, A., Ashraf, J., Mussadiq, S., Raza, A. A., Abid, M., ... & Khan, S. U. (2021). Predicting at-risk students at different percentages of course length for early intervention using machine learning models. IEEE Access, 9, 7519-7539. https://doi.org/10.1109/ACCESS.2021.3049446

Afrin, F., Hamilton, M., & Thevathyan, C. (2023, June). Exploring counterfactual explanations for predicting student success. In International Conference on Computational Science (pp. 413–420). Springer. https://doi.org/10.1007/978-3-031-36021-3_44

Arnold, K. E., & Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12) (pp. 267–270). ACM. https://doi.org/10.1145/2330601.2330666

Artelt, A., & Hammer, B. (2019). On the computation of counterfactual explanations: A survey. arXiv preprint arXiv:1911.07749. https://doi.org/10.48550/arXiv.1911.07749

Artelt, A., Vaquet, V., Velioglu, R., Hinder, F., Brinkrolf, J., Schilling, M., & Hammer, B. (2021). Evaluating the robustness of counterfactual explanations. In 2021 IEEE Symposium Series on Computational Intelligence (pp. 01–09). IEEE. https://doi.org/10.1109/SSCI50451.2021.9660058

Biecek, P., & Burzykowski, T. (2021). Explanatory model analysis: Explore, explain, and examine predictive models. Chapman and Hall/CRC. https://doi.org/10.1201/9780429027192

Brughmans, D., Leyman, P., & Martens, D. (2023). NICE: An algorithm for nearest instance counterfactual explanations. Data Mining and Knowledge Discovery, 1–39. https://doi.org/10.1007/s10618-023-00930-y

Carriero, A., Luijken, K., de Hond, A., Moons, K. G., van Calster, B., & van Smeden, M. (2024). The harms of class imbalance corrections for machine learning-based prediction models: A simulation study. arXiv preprint arXiv:2404.19494. https://doi.org/10.48550/arXiv.2404.19494

Cavus, M., & Biecek, P. (2024). An experimental study on the Rashomon effect of balancing methods in imbalanced classification. arXiv preprint arXiv:2405.01557. https://doi.org/10.48550/arXiv.2405.01557

Cavus, M., & Kuzilek, J. (2024). The actionable explanations for student success prediction models: A benchmark study on the quality of counterfactual methods. Joint Proceedings of the Human-Centric eXplainable AI in Education and the Leveraging Large Language Models for Next Generation Educational Technologies Workshops (HEXED-L3MNGET 2024), co-located with the 17th International Conference on Educational Data Mining (EDM 2024), 1-10. https://ceur-ws.org/Vol-3840/HEXED24_paper1.pdf

Cavus, M., Stando, A., & Biecek, P. (2023). Glocal explanations of expected goal models in soccer. arXiv preprint arXiv:2308.15559. https://doi.org/10.48550/arXiv.2308.15559

Chawla, N. V. (2010). Data mining for imbalanced datasets: An overview. In Data Mining and Knowledge Discovery Handbook (pp. 875–886). Springer. https://doi.org/10.1007/978-0-387-09823-4_45

Dandl, S., Hofheinz, A., Binder, M., Bischl, B., & Casalicchio, G. (2023). Counterfactuals: An R package for counterfactual explanation methods. arXiv preprint arXiv:2304.06569. https://doi.org/10.48550/arXiv.2304.06569

Dandl, S., Molnar, C., Binder, M., & Bischl, B. (2020). Multi-objective counterfactual explanations. In Proceedings of the International Conference on Parallel Problem Solving from Nature (pp. 448–469). Springer. https://doi.org/10.1007/978-3-030-58112-1_31

Drachsler, H. (2018). Trusted learning analytics. Synergie, 6, 40–43. https://doi.org/10.25657/02:19141

Elyan, E., Moreno-Garcia, C. F., & Jayne, C. (2021). CDSMOTE: Class decomposition and synthetic minority class oversampling technique for imbalanced-data classification. Neural Computing and Applications, 33(7), 2839–2851. https://doi.org/10.1007/s00521-020-05130-z

Eurostat. (2023). Glossary: Information and communication technology (ICT). Eurostat: Statistics Explained. https://tinyurl.com/eust-ict

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? Advances in Neural Information Processing Systems, 35, 507–520. https://doi.org/10.48550/arXiv.2207.08815

Guidotti, R. (2022). Counterfactual explanations and how to find them: Literature review and benchmarking. Data Mining and Knowledge Discovery, 1–55. https://doi.org/10.1007/s10618-022-00831-6

Gunonu, S., Altun, G., & Cavus, M. (2024). Explainable bank failure prediction models: Counterfactual explanations to reduce the failure risk. arXiv preprint arXiv:2407.11089. https://doi.org/10.48550/arXiv.2407.11089

Gu, Q., Tian, J., Li, X., & Jiang, S. (2022). A novel Random Forest integrated model for imbalanced data classification problem. Knowledge-Based Systems, 250, 109050. https://doi.org/10.1016/j.knosys.2022.109050

Herodotou, C., Rienties, B., Boroowa, A., et al. (2019). A large-scale implementation of predictive learning analytics in higher education: The teachers' role and perspective. Education Tech Research Dev, 67, 1273–1306. https://doi.org/10.1007/s11423-019-09685-0

Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. Science, 332(6025), 60–65. https://doi.org/10.1126/science.1200970

Hoel, T., Griffiths, D., & Chen, W. (2017). The influence of data protection and privacy frameworks on the design of learning analytics systems. In Proceedings of the Seventh International Learning Analytics & Knowledge Conference (pp. 243–252). ACM. https://doi.org/10.1145/3027385.3027414

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., & Samek, W. (2022). Explainable AI methods: A brief overview. In International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (pp. 13–38). Springer. https://doi.org/10.1007/978-3-031-04083-2_2

Junior, J. D. S. F., & Pisani, P. H. (2022). Performance and model complexity on imbalanced datasets using resampling and cost-sensitive algorithms. In Fourth International Workshop on Learning with Imbalanced Domains: Theory and Applications (pp. 83–97). PMLR.

Karimi, A. H., Barthe, G., Balle, B., & Valera, I. (2020). Model-agnostic counterfactual explanations for consequential decisions. In Proceedings of the International Conference on Artificial Intelligence and Statistics (pp. 895–905). PMLR. https://doi.org/10.48550/arXiv.1905.11190

Kuzilek, J., Hlosta, M., & Zdrahal, Z. (2017). Open university learning analytics dataset. Scientific Data, 4(1), 1–8. https://doi.org/10.1038/sdata.2017.171

Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z., Vaclavek, J., & Wolff, A. (2015). OU student data from a MOOC environment. Data in Brief, 5, 759–761. https://doi.org/10.1016/j.dib.2015.11.014

LeCun, Y., Bengio, Y. & Hinton, G. (2015).Deep learning. Nature 521, 436–444. https://doi.org/10.1038/nature14539

Liu, J. (2022). Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. Soft Computing, 26(3), 1141-1163. https://doi.org/10.1007/s00500-021-06532-4

López-Pernas, S., Saqr, M., Conde, J., Del-Río-Carazo, L. (2024). A Broad Collection of Datasets for Educational Research Training and Application. In: Saqr, M., López-Pernas, S. (eds) Learning Analytics Methods and Tutorials. Springer, Cham. https://doi.org/10.1007/978-3-031-54464-4_2

Molnar, C. (2020). Interpretable machine learning. Lulu.com.

Murphy, K. P. (2022). Probabilistic Machine Learning: An Introduction. MIT Press.

Papamitsiou, Z., & Economides, A. (2014). Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society, 17(4), 49–64. http://www.jstor.org/stable/jeductechsoci.17.4.49

Pinto, J. D., & Paquette, L. (2024). Towards a unified framework for evaluating explanations. arXiv preprint arXiv:2405.14016. https://doi.org/10.48550/arXiv.2405.14016

Siemens, G., & Baker, R. (2012). Learning analytics and educational data mining: Towards communication and collaboration. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK '12) (pp. 252–254). Association for Computing Machinery. https://doi.org/10.1145/2330601.2330661

Singer, N. (2014). InBloom student data repository to close. The New York Times, April 21, 2014. Available under: https://uhh.de/2rgnb [11.07.2018].

Smith, B. I., Chimedza, C., & Bührmann, J. H. (2022). Individualized help for at-risk students using model-agnostic and counterfactual explanations. Education and Information Technologies, 1–20. https://doi.org/10.1007/s10639-021-10661-6

Stando, A., Cavus, M., & Biecek, P. (2024, June). The effect of balancing methods on model behavior in imbalanced classification problems. In Fifth International Workshop on Learning with Imbalanced Domains: Theory and Applications (pp. 16–30). PMLR. https://doi.org/10.48550/arXiv.2307.00157

Tao, X., Li, Q., Guo, W., Ren, C., Li, C., Liu, R., & Zou, J. (2019). Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification. Information Sciences, 487, 31–56. https://doi.org/10.1016/j.ins.2019.02.062

Tsiakmaki, M., & Ragos, O. (2021). A case study of interpretable counterfactual explanations for the task of predicting student academic performance. In 2021 25th International Conference on Circuits, Systems, Communications and Computers (CSCC) (pp. 120–125). IEEE. https://doi.org/10.1109/CSCC53858.2021.00029

Vaswani, A. (2017). Attention is all you need. Advances in Neural Information Processing Systems. https://doi.org/10.48550/arXiv.1706.03762

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*, 841–872. https://doi.org/10.48550/arXiv.1711.00399

Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting the academic performance of students from VLE big data using deep learning models. Computers in Human Behavior, 104, 106189. https://doi.org/10.1016/j.chb.2019.106189

Warren, G., Keane, M. T., Gueret, C., & Delaney, E. (2023). Explaining groups of instances counterfactually for XAI: A use case, algorithm, and user study for group-counterfactuals. arXiv preprint arXiv:2303.09297. https://doi.org/10.48550/arXiv.2303.09297

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2019). The what-if tool: Interactive probing of machine learning models. IEEE Transactions on Visualization and Computer Graphics, 26(1), 56–65. https://doi.ieeecomputersociety.org/10.1109/TVCG.2019.2934619

Yin, J., Gan, C., Zhao, K., Lin, X., Quan, Z., & Wang, Z. J. (2020). A novel model for imbalanced data classification. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 6680–6687). https://doi.org/10.1609/aaai.v34i04.6145

Zhang, H., Dong, J., Lv, C., Lin, Y., & Bai, J. (2023). Visual analytics of potential dropout behavior patterns in online learning based on counterfactual explanation. Journal of Visualization, 26(3), 723–741. https://doi.org/10.1007/s12650-022-00899-8

Zong, W., Huang, G. B., & Chen, Y. (2013). Weighted extreme learning machine for imbalance learning. Neurocomputing, 101, 229–242. https://doi.org/10.1016/j.neucom.2012.08.010

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2024; 15(Special issue); 318-347

Integrating Metadiscourse Analysis with Transformer-Based Models for Enhancing Construct Representation and Discourse Competence Assessment in L2 Writing: A Systemic Multidisciplinary Approach

Sathena CHAN* Manoranjan SATHYAMURTHY** Chihiro INOUE***
Michael BAX**** Johnathan JONES***** John OYEKAN******

Abstract

In recent years, large-scale language test providers have developed or adapted automated essay scoring systems (AESS) to score L2 writing essays. While the benefits of using AESS are clear, they are not without limitations, such as over-reliance on frequency counts of vocabulary and grammar variables. Discourse competence is one important aspect of L2 writing yet to be fully explored in AEE application. Evidence of discourse competence can be seen in the use of Metadiscourse Markers (MDM) to produce reader-friendly texts. The article presents a multidisciplinary study to explore the feasibility of expanding the construct representation of automated scoring models to assess discourse competence in L2 writing. Combining machine learning, automated textual analysis and corpus-linguistic methods to examine 2000 scripts across two tasks and five proficiency levels, the study investigates (1) in addition to frequency and range, whether accuracy of MDM is worth pursuing as a predictive feature in L2 writing, and (2) how identification and classification of MDM use might be fed into developing an automated scoring model using machine learning techniques. The contributions of this study are three-fold. Firstly, it offers valuable insights within the context of Explainable AI. By integrating MDM usage and accuracy into the scoring framework, this research moves beyond frequency-based evaluation. This study also makes significant contributions to the current understanding of L2 writing development that even lower-proficiency learners exhibit evidence of discourse competence through their accurate use of MDMs as well as their choice of MDMs in response to genre. From the perspective of expanding the construct representation in automated scoring systems, this study provides a critical examination of the limitations of many AEE models, which have heavily relied on vocabulary and grammar features. By exploring the feasibility of incorporating MDMs as predictive features, this research demonstrates the potential for construct expansion of L2 AEE. The results would support test providers in developing competence tests in various contexts and domains including manufacturing, medicine and so on.

Keywords: L2 Writing, Metadiscourse Markers, Automated Essay Scoring, Large Language Models

* Assoc. Prof. Dr., University of Bedfordshire, CRELLA, UK, sathena.chan@beds.ac.uk, ORCID ID: 0000-0002-7852-6737
** Researcher, University of Oxford, UK, manosathya98@gmail.com, ORCID ID: 0000-0001-8928-2689
*** Assoc. Prof. Dr., University of Bedfordshire, CRELLA, UK, chihiro.inoue@beds.ac.uk, ORCID ID: 0000-0003-1927-6923
**** Researcher, Weblingua LTD, UK, michael@textinspector.com, ORCID ID: 0000-0002-2753-1990
***** Dr., University of Bedfordshire, CRELLA, UK, johnathan.jones@beds.ac.uk, ORCID ID: https://orcid.org/0000-0003-4158-7971
****** Senior Lecturer, University of York, UK, john.oyekan@york.ac.uk, ORCID ID: 0000-0001-6578-9928

To cite this article:
Chan, S., Sathyamurthy, M., Inoue, C., Bax, M., Jones, J., & Oyekan, J. (2024). Integrating Metadiscourse Analysis with Transformer-Based Models for Enhancing Construct Representation and Discourse Competence Assessment in L2 Writing: A Systemic Multidisciplinary Approach. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 318-347. https://doi.org/10.21031/epod.1531269

Received: 13.08.2024
Accepted: 16.12.2024

Introduction

In recent years, large-scale language test providers have developed or adapted automated essay scoring systems (AESS) to score second language (L2) writing essays. For example, Educational Testing Service uses Natural Language Processing based e-rater® Scoring Engine and Pearson uses Intelligent Essay Assessor™ through a combination of Latent Semantic Analysis (LSA) and other methods. While the benefits of incorporating AEE applications in the scoring systems are clear, they are not without limitations. Early systems were criticized for their over-reliance on frequency counts of vocabulary and grammar variables (Chapelle and Chung, 2010). Current state-of-art AESS have incorporated scoring features such as content and organization. However, discourse competence as one important aspect of L2 writing is yet to be fully explored in AESS. Discourse competence "concerns the ability to design texts, including generic aspects like thematic development and coherence and cohesion as well as … cooperative principles and turn-taking" (CoE, 2018, p.138). In writing, evidence of discourse competence can be seen in the use of metadiscourse markers (MDM) to produce reader-friendly texts. Such competence is typically expected from higher-proficiency L2 writers learners, especially at the CEFR B2 level or onwards (CoE, 2018), when they have mastered linguistic accuracy and basic writing skills Nevertheless, in the increasingly multicultural contexts we live in, discourse competence which underpins effective communication is relevant to L2 learners across the proficiency spectrum, arguably more so for lower-proficiency learners who need to build meaningful connections and achieve educational/professional goals. The article presents a multidisciplinary study to explore the feasibility of expanding the construct representation of AESS to assessing discourse competence in L2 writing. This would improve the way tests are developed and assessed across various contexts, domains and sectors including manufacturing, construction, medicine and so on thereby supporting low-skilled to highly skilled labor in these areas.

Use of MDMs in L2 writing tests

Metadiscourse markers (MDM) are defined in this study as "those aspects of the text which explicitly refer to the organization of the discourse or the writer's stance towards either its content or the reader" (Hyland, 2005, p. 109). The use of MDMs has two major functions. Firstly, skilled writers use MDMs to signal the organization of a text and provide cohesion between ideas in a text, e.g., to indicate conjunctive and/or additive, adversarial, causal and temporal relationships in the text (Schiffrin et al., 2001, p.55). Secondly, MDMs are used to state the attitude of the writer (Burneikaite, 2008). Skilled writers use MDMs to provide an explicit organizational structure within a text and to guide the reader to their attitude on the topic. Appropriate use of MDMs makes a text more reader-friendly, especially for L2 readers (Camiciottoli, 2003). Despite the importance of discourse competence in the development of L2 writing proficiency, especially when learners progress to CEFR B2 or upwards (CoE, 2001), evaluation of the use of MDMs in L2 writing is typically reduced to a holistic judgment of the number and/or range of cohesive devices used under the criterion of "cohesion and coherence" in human scoring schemes (for example see the Aptis Guide, 2019). This approach might be limited to reveal the nuanced developmental features of the use of MDMs by L2 writers.

We now review the previous studies on the use of MDMs in L2 writing. Most of these studies focused on the use of MDMs by upper-intermediate L2 writers, comparing their academic essays with those of L1 writers (e.g., Adel, 2006; Crompton, 2012; Hyland, 2005; Lee & Deakin, 2016). Their findings are clearly inconclusive and contradictory at times. Some studies found that higher-proficiency writers use more MDMs overall than lower-proficiency writers (Sanford, 2012). Others reported higher use of certain MDMs (such as endophoric markers and evaluative markers) among higher-proficiency writers (Burneikaitė, 2008). In contrast, some reported that higher-proficiency writers use fewer logical connectives than lower-proficiency writers but used a wider range of MDMs (Carlsen, 2010).

Only a handful of studies investigated the use of MDMs by L2 learners in standardized writing tests. In Knoch et al.'s (2014) study on TOEFL writing test, lower-proficiency writers used more MDMs overall than more proficient writers. Bax et al. (2019) conducted the first large-scale study to examine L2 test takers' use of MDMs. 900 writing scripts produced by L2 test takers at CEFR B2-C2 levels were examined. They found that higher-proficiency writers used fewer MDMs but a significantly wider range of MDMs than lower-proficiency writers. Barkaoui (2016) investigated only interactional MDMs among repeaters of IELTS and found no significant effects in test taker group on the overall use of interactional MDMs. However, test takers scoring IELTS 6.0 (indicating CEFR level B2 according to test providers' information) tended to use more hedges and boosters but fewer self-mentions than did test takers with lower initial writing scores. Owen et al. (2021) expanded on Bax et al.'s work to include test takers from CEFR A levels. The results showed that each of the 13 MDM categories used in their study discriminated across at least one CEFR boundary. The overall deployment of MDMs changed significantly in transitioning from A0 to A2 levels and from B1 to C levels. The range of MDMs also rose across CEFR thresholds, with significant differences obtained across A1-A2 and A2-B1 thresholds. As a result, they argued that the use of MDMs should be operationalized separately from vocabulary (grammatical competence) as part of discourse competence (Bachman and Palmer, 2010).

These studies clearly show differences in frequency and range of MDMs used by L2 writers, indicating that increasing (or sometimes decreasing) use of MDMs may signal test takers' ability to manage textual and interpersonal complexity in discourse. However, the findings are inconclusive in at least two aspects. First, the direction of the relationship between the use of MDMs (frequency and range) and L2 writing proficiency is inconclusive. Second, differences in frequency and range of MDMs seem observable between some levels but not others. As a result, simple frequency and range counts of MDMs might not be the most suitable way of distinguishing between L2 writing proficiency levels, especially for writing tests which target multiple proficiency levels.

Potentials and challenges of Automated Essay Scoring Systems

Automated essay scoring systems (AESS) have become increasingly prevalent in the assessment of L2 writing. A range of lexical and some syntactic measures have been shown to consistently discriminate across score boundaries in large-scale testing. Lexical complexity can be measured in terms of rarity, variability and disparity (Jarvis, 2013). For example, word frequency counts in relation to threshold levels of vocabulary use based on various wordlists, e.g. English Vocabulary Profile, Academic Word List and New General Service List (Brezina & Glabasova, 2013) are commonly used in L2 writing AESS. However, most AESS rely heavily on frequency and range of lexical use based on frequency wordlists.

The performance of pre-trained transformers on various NLP tasks is well documented, however this does not necessarily translate to good out-of-the-box performance on all downstream tasks presented to the model (Lin et al., 2022). Currently pre-trained transformers have been used to obtain word embeddings; after which a classifier has been trained to perform our binary classification task. We can build upon the knowledge that the pre-trained transformer has learnt by fine-tuning the model using our labelled dataset. In one such fine-tuning method, we can alter parameters in a given number of layers in the transformer architecture that we wish to fine-tune, leaving parameters in all other layers untouched (Lialin et al., 2023). This can, depending on the level of fine-tuning, potentially be a reasonably resource consuming task; it is, however, capable of boosting performance for particular tasks.

Taken together, research is needed to explore whether AESS can be extended to detect frequency and range of MDM as measures of discourse competence in L2 writing and whether MDM accuracy can serve as a predictive feature in L2 writing.

Methods

Through a multidisciplinary approach combining machine learning, automated text analysis and corpus-linguistic methods, we investigated whether MDM accuracy is a predictive feature in L2 writing and the feasibility of building an automated scoring model to identify use of MDMs and to distinguish between accurate and inaccurate use of MDMs. Three research questions guided this study:

RQ1: How do learners use MDMs across proficiency levels?

RQ2.1: To what extent can a transformer-based AI model classify correctly whether or not a word or a phrase is a MDM?

RQ2.2: To what extent can a transformer-based AI model classify correctly the accurate and inaccurate use of MDMs?

The Research Questions were addressed in two phases. Phase 1 involved human coding to identify and examine frequency and accuracy of use of MDMs by test takers taking a large-scale proficiency writing test, and to explore that use across a range of CEFR levels. Phase 2 involved use of machine learning of the human-coded data to investigate which machine learning algorithms could be used to develop an automated model to replicate expert judgement on detection and accuracy of MDMs.

Tasks and data set

The dataset for the study consists of 2,003 sample scripts from the corpus of Aptis candidates' writing. Aptis is a standardized multi-level English Proficiency Test. The Writing component of the Aptis test consists of four parts. Part 4 (Formal and Informal Writing) was used in the current study. Aptis writing is evaluated by trained and certified human examiners. Although Aptis employs a single-rating approach, different raters are assigned to each task, ensuring that multiple observations are made of a single candidate's response. The inter-rater reliability on benchmark Writing responses is at 0.97 (O'Sullivan, Dunlea, Spiby, Westbrook, and Dunn, 2020). Since Aptis is taken by candidates in different international contexts, candidates are allowed to use any standardized version of English (e.g., American, Australian, British, Singaporean) in the writing test as long as it is consistent, especially in the formal writing task.

The scripts used in this study were from the two email tasks in Part 4 of the Aptis writing test. The two tasks were thematically-linked. Task 1 required the candidates to write an email (40-50 words) to a classmate friend about a class cancellation in a cooking school as the teacher is going on a holiday. Task 2 required the candidates to write an email (120-150 words) to the manager of the cooking school to complain about the cancellation. They had 20 minutes to finish each task. Each script was operationally tagged with a CEFR level based on the candidate's test scores received on the task (as part of the standard test procedure in Aptis), and the breakdown of the numbers of scripts at the five CEFR proficiency levels is shown in Table 1.

Table 1

Numbers of scripts used for analysis in this study

	A1	A2	B1	B2	C	Total
Task 1	175	210	190	187	234	996
Task 2	173	206	197	193	238	1007

MDM Categories

We used the categories of MDM shown in Table 2 (Hyland, 2005, modified by Bax et al., 2019) (see Appendix 1 for the full list).

Table 2

Categories of MDM

	Category analyzed		Function	Examples
Textual metadiscourse	Logical connectives		Express semantic relation between main clauses	In addition / but / thus / and
	Frame markers:	Sequencing	Explicitly refer to discourse acts or text stages	Finally / to repeat / here, we try to
		Label Stages		
		Announce goals		
		Topic shift		
	Code glosses		Help readers grasp meanings of ideational material	Namely / such as / e.g. / i.e.
	Endophoric markers		Refer to information in other parts of the text	Noted above / see figure X
	Evidentials		Refer to source of information from other texts	According to X, … / 1990 / X argues that…
Interpersonal metadiscourse	Attitude markers		Expressing opinion of propositional content	I agree that… / X claims that…
	Hedges		Withhold writer's full commitment to statements	Might / perhaps / possible
	Relational markers		Explicitly refer to or build relationship with reader	Frankly / note that / as you can see…
	Person markers		Explicit reference to author	I / we / mine / our
	Emphatics		Emphasize force or certainty in message	Definitely / in fact / it is certain that…

Procedures for RQ1 (Use of MDM at different proficiency levels)

A total of 996 Tasks 1 and 1007 Tasks 2 scripts were manually coded using the procedures described below. The manual coding results were then used to build a labelled training dataset as the first step for developing a transformer-based AI model to identify and assess accuracy of MDM in RQ2.

Automated tagging of MDM and data cleaning

Text Inspector, a web-based tool allowing users to analyze features of texts, was used to provide an initial tagging of MDM using categories of Hyland's (2005) list. Adopted from the procedures used in Owen et al. (2021), we cleaned the tagged dataset as follows:

- Full stops and exclamation marks were removed, since the units of analysis were not sentences;
- Special symbols were removed or replaced with correct ones; and
- Spelling errors were corrected to improve the accuracy of automated classification.

During the initial coding of the dataset (i.e., 30% with over 100,000 words in total for each task), we found that more than half of the inaccurate uses of MDMs were spelling errors[1] (e.g., 102 out of 200 inaccurate MDMs in Task 1). A decision was made to correct them for two reasons. First, the focus of the study is about the frequency, range and accuracy of MDM use by L2 writers. As argued previously, this is related to their discourse competence (Bachman and Palmer, 2010) to signal the organization and/or the author's stance in a text for its reader rather than their ability to spell the markers correctly, which is typically assessed in relation to "vocabulary" in L2 writing. Secondly, inclusion of misspelled words would increase variation for the algorithms to accurately classify use of MDMs.

RQ1 Coding Procedures

The tagged scripts were then reviewed and coded manually for the use of MDMs by two researchers, following these procedures:

1) Adding any words and expressions to the list of MDMs that are suitable for the genre of email writing. As Hyland's (2005) list was devised based on journal articles, it does not include the full range of MDMs that were found in emails in the current study. For example, among frame markers in Hyland's list, examples expressions include 'here, we try to…' for announcing the goal of the piece of a text. However, this expression is unlikely to be used in emails; instead, we frequently observed 'I am writing this email to…' at the earlier part in emails, which need to be added to the list for this study. There were also more varied attitude markers such as 'disappointing/ disappointed' and 'happy' in the scripts than would be in journal articles. The list of additional MDMs can be found in Appendix 2.

2) Indicating any words or expressions tagged according to Hyland's list that do not serve as MDMs in the current data set. Related to the above point, there are some words and expressions that qualify as MDMs in journal articles, but not in emails. For example, the word 'next' is tagged as a MDM according to Hyland (2005), which signals the sequencing of texts in journal articles (e.g. 'Next, we examine…'). However, in the scripts in this study, 'next' is often used to say 'next week', which does not serve as a MDM in the simulated email texts. These non-applicable tags were identified and removed during coding.

3) Code dichotomously for the identification and accurate or inaccurate use of correctly-tagged (by Text Inspector) and newly identified MDMs (see Figure 1). Specifically, the two coders make decisions on two questions:

Q1: Is this a MDM? (1: yes, 0: no)

Q2: If it is a MDM, is it correctly used ? (1: yes, 0: no)

The coded data was used to address RQ1 (i.e. the frequency, range and accuracy of MDM use across proficiency levels) as well as serving a labelled set for training algorithms for RQs 2.1 and 2.2.

Due to the exploratory nature of this first study to develop AESS models to assess the use of MDMs by L2 writers, we sought a dichotomous instead of polytomous coding scheme regarding the accuracy of MDM because the latter would require a more complex model for machine learning (more will be discussed regarding the procedures for RQ2). Because of the nature of the dichotomous coding scheme, the inaccurate MDM use needed to be undoubtedly inaccurate, see examples below. In this study, the 0 codes (for inaccurate use) therefore largely represented grammatical errors that surround MDMs use (see Examples 1-3)

Example 1: when you return of <u>you</u> holiday ….(A1 script, relational marker)

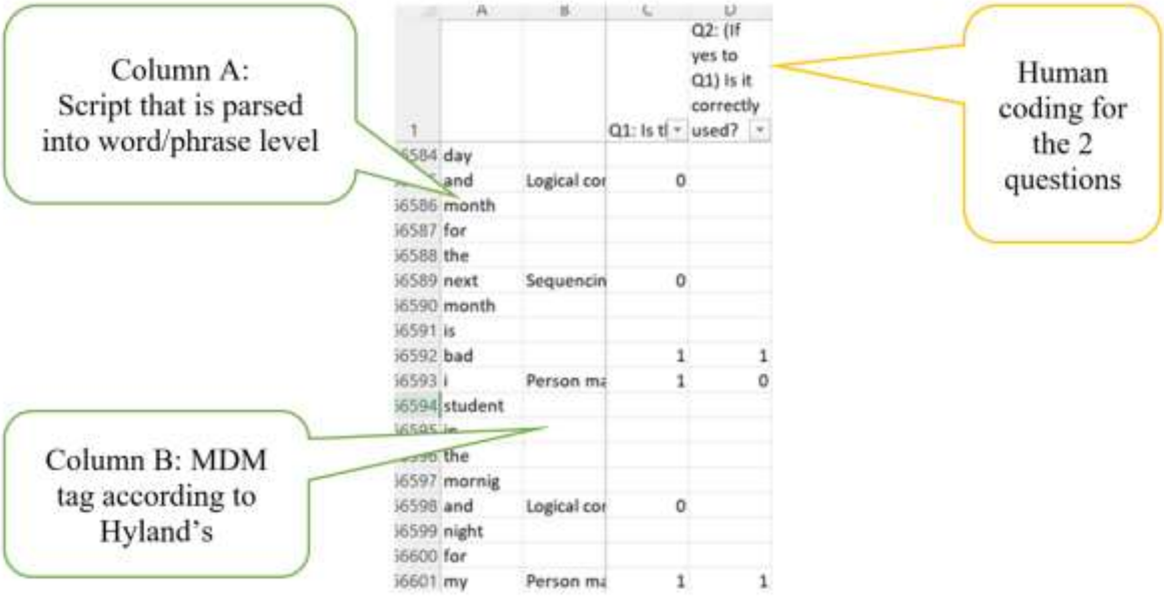
Example 2: 20th of the next moth fo <u>my</u> is the most ….(A2 script, relational marker)

[1] When a potential spelling error was identified and the spelt word exists in English, it was not corrected (for example, in the case of 'Thank your', "your" could have been misspelt (instead of "you"). But since "your" in itself is not misspelt, it was not corrected.

Example 3: Please let's know on the status … (A2 script, relational marker)

Figure 1

A screenshot of example human coding



	A	B	C	D
1			Q1: Is tl	Q2: (If yes to Q1) Is it correctly used?
6584	day			
	and	Logical cor	0	
6586	month			
6587	for			
6588	the			
6589	next	Sequencin	0	
6590	month			
6591	is			
6592	bad		1	1
6593	i	Person ma	1	0
6594	student			
6595	in			
	the			
6597	mornig			
6598	and	Logical cor	0	
6599	night			
6600	for			
6601	my	Person ma	1	1

Column A: Script that is parsed into word/phrase level

Human coding for the 2 questions

Column B: MDM tag according to Hyland's

As a result, the range of inaccurate MDM use is narrower in this study than what might usually be regarded as inaccurate use. Less appropriate use of MDMs (such as using a formal label stage when writing to a friend in informal email task (see Example 4) and using emphatics instead hedges when writing to a manager in the formal email task – see Example 5) was not coded as 0 (inaccurate).

Example 4: in conclusion, … (B1 script, label stage)

Example 5: I am feel really disappointed …(B1 script, emphatics)

To differentiate these developmental features of the discourse competence, a polytomous coding scheme, which was deemed inappropriate for this first study, would be required. The two coders double-coded 123 scripts per task, which makes up 10% of the data. After several rounds of discussions and re-coding, the (working) list of additional MDM for email writing (Appendix 2) was agreed and the exact agreement rate reached over 90% for both tasks (Task 1, Q1[MDM or not]: 96.4%, Task 1: Q2 [accurate use or not]: 96.1%; Task 2, Q1 (98.5%, Task 2, Q2: 98.3%). The coding reliability was deemed sufficiently high, and thus the two coders continued on to code two different sets of scripts (each batch containing 45% of the scripts) independently.

We report a descriptive summary of human coding in relation to the ratio of scripts where at least one MDM (irrespective of accuracy) in each MDM category across proficiency levels appeared for each task to show the general trend. Kruskal-Wallis tests were then run for the average ratio of accurately used MDMs across levels.

Procedures for RQ2 (Transformer-based model to classify use and accuracy of MDM)

To remind the reader, RQ2 aims to explore how a machine learnt automated scoring model could be applied to evaluate test taker's MDM use. This involved four stages: the experimental setup, production of word embeddings, automated classification using word embeddings, and finally improvement of the classifiers used in the classification task.

(1) Experimental Setup

The human coded scripts were used as a labelled dataset for this part of the study. For the purpose of this project, we considered each of the research questions, i.e., RQ2.1 and RQ2.2, as an individual binary classification task. The premise was that each word in our dataset can be labelled as a 1 or 0.

a. 1 – a word is a MDM or 0 – a word is not a MDM [RQ2.1]

b. 1 – it is accurately used or 0 – it is not accurately used [RQ2.2]

In order to select a suitable machine learning methodology to assess the MDM use of the test takers, we considered several algorithms that were capable of producing word embeddings. These included Recurrent Neural Networks (RNNs), Long-Short Term Memory (LSTMs) (Hochreiter & Schmiduber, 1997), and Transformers (Vaswani et al., 2017). Given their success in various downstream natural language processing (NLP) tasks in the literature, Transformers were chosen for this task. Additionally, they offer vastly reduced training times due to its ability to process entire sequences in parallel, through the use of ‘attention mechanisms’ that allows for tracking the relations between words across long text sequences in both forward and backward directions simultaneously.

The following classifiers/classification algorithms were selected to evaluate the performance of the appropriacy of MDM use by test takers:

- AdaBoost (Freund & Schapire, 1999)
- Decision Tree
- k-nearest neighbors classifier (kNN) (Zhang, Z., 2016)
- Multi-layer Perceptron (MLP)
- Naive Bayes (Zhang, 2004)
- Quadratic Discriminant Analysis (QDA)
- Random Forest Classifier (RFC) (Breiman, 2001)

(2) Embeddings

Each script was passed through a given embedding method in order to obtain word embeddings for all words contained in that script. For example, a Transformer mathematically encodes the words in context in the labelled dataset. A word is expressed in the form of a vector input (i.e., a string of numbers) which is called a word embedding.

A simple binary classifier requires a vector input for each data point (i.e., a string of numbers representing a given word in our dataset) in order to predict an output class. A vector representation of our textual data must be derived.

A given word in a sentence taken in isolation has little interpretability. The meaning of a word is dependent on its context and, as such, we must be able to encode information about a sequence of words in a single vector. Word embeddings give us a way to represent each word as an individual vector, whilst maintaining varying levels of contextual information in each embedding.

The majority of NLP tasks use Transformers to obtain these embeddings, given its state-of-the-art performance (SOTA) on benchmark NLP tasks as well as faster training times than conventional machine learning methods designed for sequential data, such as RNNs and LSTMs. Increasingly larger datasets are being used for training which has given rise to generalizable pre-trained models, such as Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018). The application of a pre-trained Transformer enabled us to make use of a model that has been trained on very large datasets compared to the size of the dataset used in this report. As a result, the models provided a bootstrap mechanism for the work in this report.

(3) Automated classification by classifiers

The word embeddings were then used to perform the binary classification tasks, using the labelled dataset in order to train the classifier. Word embeddings serve as features that allow a classifier to group words with similar properties together. The classifier outputs 1 or 0 for each word (as coded in the labelled dataset).

The initial automated classification shows that in our labelled dataset, only 13% of words were labelled by the classifiers as MDMs (RQ2.1) and of these only 5.9% are labelled as not appropriately used (RQ2.2).

(4) Improvement of Classifiers

Based on the results of (3), measures were used to improve the performance of the classifiers. Any given algorithm has a number of parameters affecting the way it is able to learn from data, often significantly affecting classifier performance. To refine classifier performance, we also performed two fine-tuning measures:

a. Resampling methods are usually used to alter the composition of the dataset used for training such that the percentage of data belonging to each class is closer to 50%, generally improving classifier performance. Both undersampling of the majority class (the most frequently occurring class) and oversampling of the minority class (the least frequently occurring class) were trialed to observe the effects of class imbalance on the classifier. SMOTE (Chawla et al., 2002), ADASYN (He et al., 2008), SMOTEENN (SMOTE combined with edited-nearest-neighbours) and SMOTETomek (SMOTE combined with the use of Tomek Links) are resampling methods that have been used to create the resampled datasets.

b. Fine-tuning studies were conducted to find optimal learning parameters for our classifiers.

Results

RQ1: The Use of MDMs at Different CEFR Levels

Overall use of MDMs

The summary of human coding is presented in the form of descriptive statistics in Table 3, showing the ratio of scripts where at least one MDM (irrespective of accuracy) in each category appeared. Figure 2 presents the same information visually.

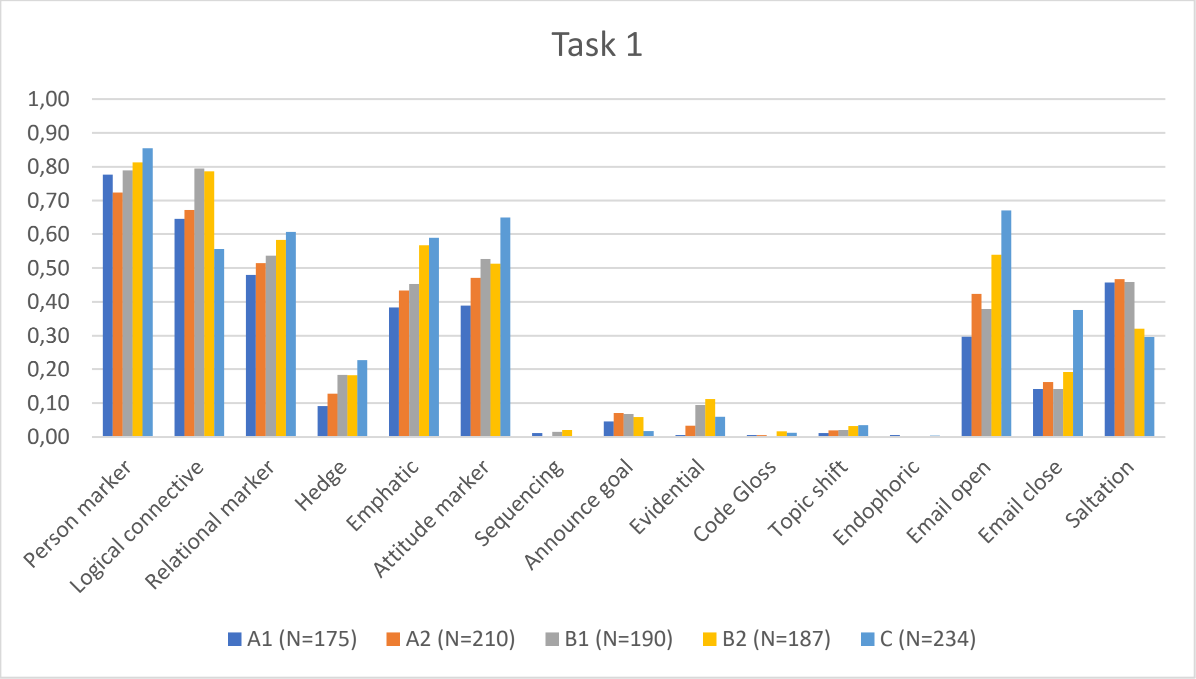
Table 3

Ratio of scripts with at least 1 MDM use (Task 1)

	A1 (N=175)	A2 (N=210)	B1 (N=190)	B2 (N=187)	C (N=234)	Whole (N=996)
Person marker	0.78	0.72	0.79	0.81	0.85	0.79
Logical connective	0.65	0.67	0.79	0.79	0.56	0.68
Relational marker	0.48	0.51	0.54	0.58	0.61	0.55
Hedge	0.09	0.13	0.18	0.18	0.23	0.17
Emphatic	0.38	0.43	0.45	0.57	0.59	0.49
Attitude marker	0.39	0.47	0.53	0.51	0.65	0.52
Sequencing	0.01	0.00	0.02	0.02	0.00	0.01
Announce goal	0.05	0.07	0.07	0.06	0.02	0.05
Evidential	0.01	0.03	0.09	0.11	0.06	0.06
Code Gloss	0.01	0.00	0.00	0.02	0.01	0.01
Topic shift	0.01	0.02	0.02	0.03	0.03	0.02
Endophoric	0.01	0.00	0.00	0.00	0.00	0.00
Email open	0.30	0.42	0.38	0.54	0.67	0.47
Email close	0.14	0.16	0.14	0.19	0.38	0.21
Saltation	0.46	0.47	0.46	0.32	0.29	0.40

Figure 2

Ratio of scripts with at least 1 MDM use (Task 1)



From Table 3 (and Figure 2), Task 1 (i.e., an informal email to a friend) elicited five interpersonal MDM groups (i.e., person marker, relational marker, hedge, emphatic and attitude marker), one textual MDM (i.e., logical connective) and three genre-specific MDM groups (i.e., saltation, email open and email close). The ratio of scripts that had at least one MDM in these categories tended to increase as the levels went up, except for logical connective and saltation marker. Specifically, the ratio of scripts containing at least one logical connective in C scripts (0.56) were lower than B1 and B2 scripts (both 0.79), and for saltation markers, the ratio was lower at B2 (0.32) and C levels (0.29) than the B1 and below (0.46 and 0.47). This echoes with Carlsen (2010)'s finding that higher-level writers tend to rely less on logical connectives to establish discourse structure. Some MDM categories were hardly used in Task 1; namely, sequencing, announce goal, evidential, code gloss, topic shift, and endophoric MDMs. Different from the general perception that discourse competence develops at higher-proficiency levels, most of the lowest-proficiency A1 and A2 writers in this study used person markers and logical connectives, half used relational markers and saltation, and over one-third used emphatic and attitude markers.

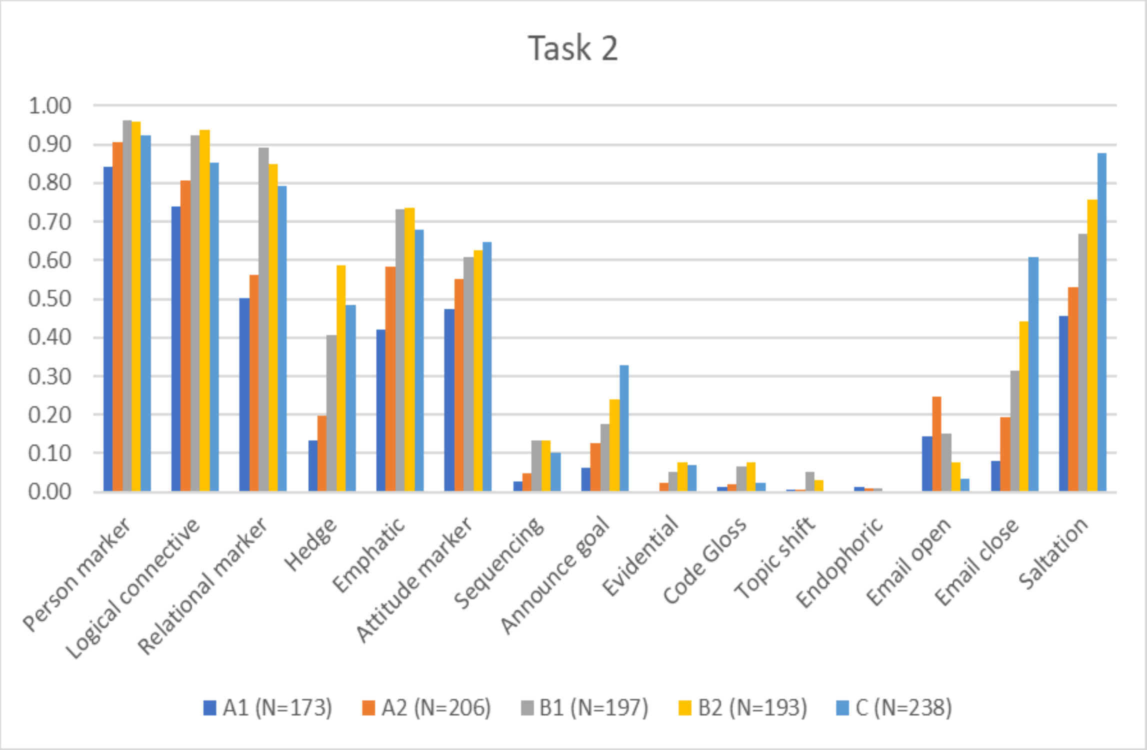
Table 4

Ratio of scripts with at least 1 MDM use (Task 2)

	A1 (N=173)	A2 (N=206)	B1 (N=197)	B2 (N=193)	C (N=238)	Whole (N=1007)
Person marker	0.84	0.91	0.96	0.96	0.92	0.92
Logical connective	0.74	0.81	0.92	0.94	0.85	0.85
Relational marker	0.50	0.56	0.89	0.85	0.79	0.73
Hedge	0.13	0.20	0.41	0.59	0.48	0.37
Emphatic	0.42	0.58	0.73	0.74	0.68	0.64
Attitude marker	0.47	0.55	0.61	0.63	0.65	0.59
Sequencing	0.03	0.05	0.13	0.13	0.10	0.09
Announce goal	0.06	0.13	0.18	0.24	0.33	0.19
Evidential	0.00	0.02	0.05	0.08	0.07	0.05
Code Gloss	0.01	0.02	0.07	0.08	0.03	0.04
Topic shift	0.01	0.00	0.05	0.03	0.00	0.02
Endophoric	0.01	0.01	0.01	0.00	0.00	0.01
Email open	0.14	0.25	0.15	0.08	0.03	0.13
Email close	0.08	0.19	0.31	0.44	0.61	0.34
Saltation	0.46	0.53	0.67	0.76	0.88	0.67

Figure 3

Ratio of scripts with at least 1 MDM use (Task 2)



In comparison to Task 1, Task 2 (i.e., a formal complaint email to a manager) elicited a wider range of MDM groups (see Table 4 and Figure 3). Test takers used five interpersonal MDM groups (i.e., person marker. Relational marker, hedge, emphatic and attitude marker), three textual MDM (i.e., logical connective, sequencing and announce goal) and three genre-specific MDM groups (i.e., saltation, email open and email close). On Task 2, it was not always C level candidates who revealed the highest ratio, although it is generally observed that more candidates use the MDMs at higher levels. It is notable that C level candidates produced more MDMs for saltation (0.88), to announce goals (0.33), and to close the email message (0.61) than the candidates at lower levels. This suggests that C level candidates might be more aware of the structure of a formal complaint email, in which they addressed and stated more clearly why they are writing to the person of power (e.g., school manager) while expressing their feelings (e.g., I am deeply disappointed that…) as well as closing the email often asking for a prompt response. The lowest-proficiency A1 and A2 writers, again, showed evidence of discourse competence through use of MDMs. A vast majority of A1 and A2 writers used person markers and logical connectives, half used relational, emphatic, attitude markers and saltation, and 20% of A2 writers used hedges. It is worth noting their different choices of MDMs between the two tasks, even though the difference was more subtle than that shown by the higher-level writers.

Accurate use of MDM at different CEFR Levels

Table 5 and Figure 4 present the average ratio of accurately used MDMs at different CEFR levels for Task 1. Table 6 and Figure 5 are for those for Task 2. It is clear that, in both tasks, the ratios of accurately used MDM are very similar across the CEFR levels—around 0.90—for most types of MDM. This means that when MDMs were used, candidates used them accurately regardless of their proficiency levels. The exceptions are the slightly lower ratios for announcing goals and email closing for both tasks. While, as aforementioned, lower-proficiency writers used these makers, higher-proficiency writers were more able to use them accurately. In comparison to the other used MDM categories (such as person markers and relational markers), there are multiple ways to achieving announcing goals and email closing and often involve more than a single word. We can also see ‘jagged’ ratios for sequencing and endophoric MDMs in Task 1, but given the very small number of cases in these MDM (as shown in Table 5), this may not be a representative picture.

Table 5

Average ratio of accurately used MDM across CEFR levels (Task 1)

	A1 (N=175)	A2 (N=210)	B1 (N=190)	B2 (N=187)	C (N=234)
Person marker	0.97	0.98	0.99	1.00	1.00
Logical connective	0.99	0.99	0.99	0.99	0.99
Relational marker	0.98	0.98	0.99	0.98	1.00
Hedge	0.84	0.89	0.97	0.99	0.98
Emphatic	0.98	0.97	0.99	0.98	0.99
Attitude marker	0.90	0.97	0.97	0.99	1.00
Sequencing	0.50	-	1.00	0.75	-
Announce goal	0.67	0.65	0.70	0.72	0.78
Evidential	0.00	0.71	1.00	0.90	0.93
Code Gloss	1.00	1.00	-	1.00	1.00
Topic shift	1.00	1.00	1.00	1.00	1.00
Endophoric	1.00	-	-	-	1.00
Email open	0.96	0.93	0.92	0.96	0.96
Email close	0.75	0.62	0.49	0.60	0.76
Saltation	0.97	0.98	1.00	1.00	1.00

Figure 4

Average ratio of accurately used MDM (Task 1)

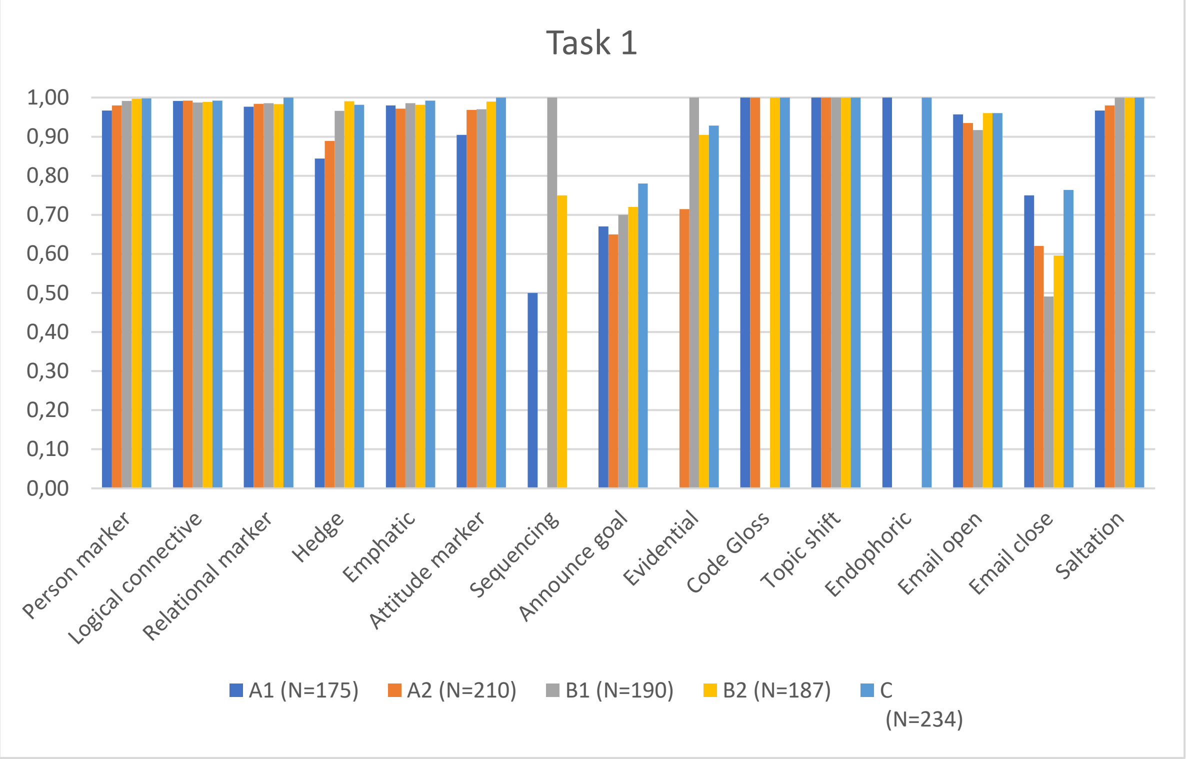


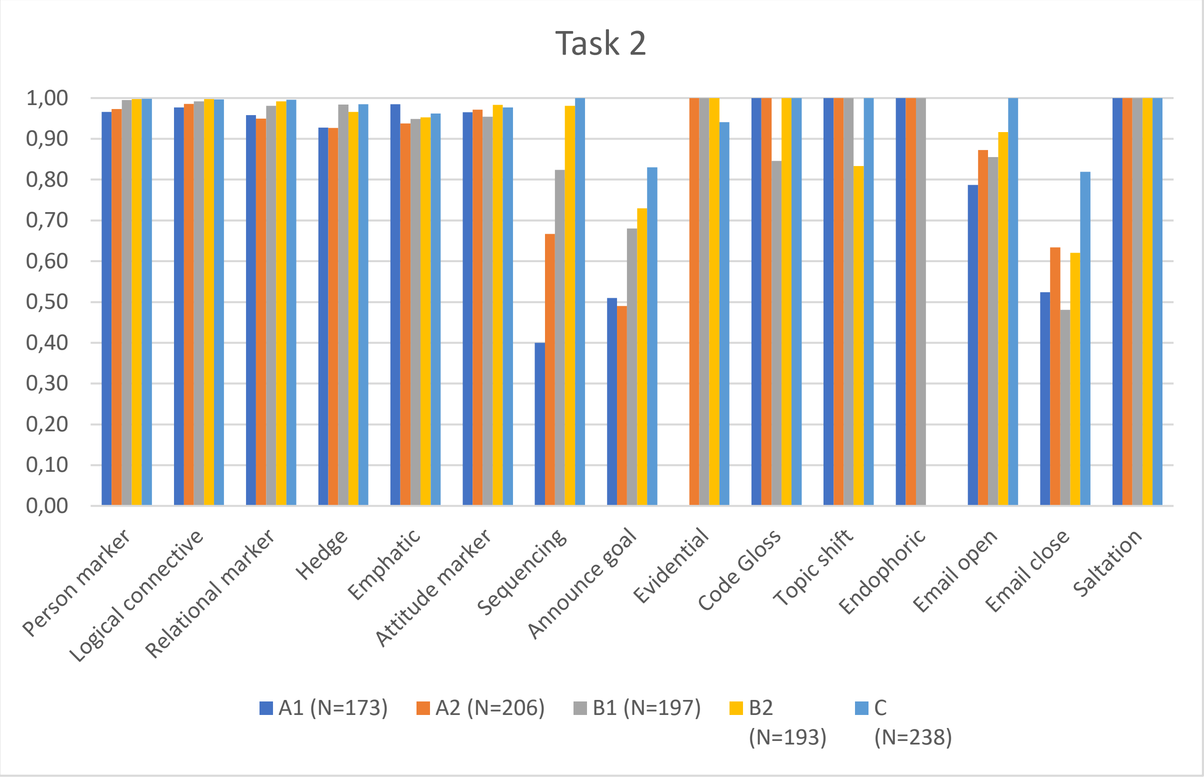
Table 6

Average ratio of accurately used MDM across CEFR levels (Task 2)

	A1 (N=173)	A2 (N=206)	B1 (N=197)	B2 (N=193)	C (N=238)
Person marker	0.97	0.97	1.00	1.00	1.00
Logical connective	0.98	0.99	0.99	1.00	1.00
Relational marker	0.96	0.95	0.98	0.99	1.00
Hedge	0.93	0.93	0.98	0.97	0.98
Emphatic	0.99	0.94	0.95	0.95	0.96
Attitude marker	0.97	0.97	0.95	0.98	0.98
Sequencing	0.40	0.67	0.82	0.98	1.00
Announce goal	0.51	0.49	0.68	0.73	0.83
Evidential	-	1.00	1.00	1.00	0.94
Code Gloss	1.00	1.00	0.85	1.00	1.00
Topic shift	1.00	1.00	1.00	0.83	1.00
Endophoric	1.00	1.00	1.00	-	-
Email open	0.79	0.87	0.86	0.92	1.00
Email close	0.52	0.63	0.48	0.62	0.82
Saltation	1.00	1.00	1.00	1.00	1.00

Figure 5

Average ratio of accurately used MDM (Task 2)



Kruskal-Wallis tests were then conducted to identify the differences in the ratio of accurately used MDMs (see Table 7). The results show significant differences in the accuracy of MDM use. Post-hoc pairwise comparisons identified some combinations of CEFR levels in which significant differences were found. However, the levels tended to be far apart, such as between A1 (beginner) and C (proficient learner).

Table 7

Results of Kruskal-Wallis tests (Task 1)

MDM Type	N	H	df	Sig.		Significant differences found between
Person marker	790	27.168	4	0.000	**	A1 and B2 (mean rank difference = 41.984, SE = 10.472, adjusted p=.001) A1 and C (mean rank difference = 45.640, SE = 9.86, adjusted p =.000) A2 and C (mean rank difference = 27.708, SE = 9.547, adjusted p=.037)
Logical connective	682	0.47	4	0.976		
Relational marker	545	5.119	4	0.275		
Hedge	165	8.08	4	0.089		
Emphatic	488	1.107	4	0.893		
Attitude marker	515	18.061	4	0.001	**	A1 and B2 (mean rank difference = 24.046, SE = 7.504, adjusted p=.014) A1 and C (mean rank difference = 26.765, SE = 6.907, adjusted p=.001)

Table 7 (Continued)

Results of Kruskal-Wallis tests (Task 1)

MDM Type	N	H	df	Sig.		Significant differences found between
Sequencing	9	1.571	2	0.456		
Announce goal	51	0.973	4	0.914		
Evidential	61	13.788	4	0.008	**	A1 and B1 (mean rank difference = 44.703, SE = 15.747, adjusted p=.045) B1 and C (mean rank difference = 44.525, SE = 12.481, adjusted p=.004)
Code Gloss	8	0	3	1.000		
Topic shift	24	0	4	1.000		
Endophoric	2	0	1	1.000		
Email open	471	3.938	4	0.414		
Email close	210	17.565	4	0.002	**	A1 and B2 (mean rank difference = 41.984, SE = 10.472, adjusted p=.001) A1 and C (mean rank difference = 45.640, SE = 9.86, adjusted p =.000) A2 and C (mean rank difference = 27.708, SE = 9.547, adjusted p=.037)
Saltation	394	7.14	4	0.129		

Table 8 presents the results of Kruskal-Wallis tests for Task 2. Like Task 1, the accuracy of use was found to be significantly different in some MDMs. The post-hoc pairwise comparisons identified differences between CEFR levels that are closer for some MDMs (e.g., B2 and C in email closing) in Task 2 than in Task 1. This is probably due to the nature of the email tasks as the MDMs used in formal emails (I'm writing to) tend to be more formulaic than those used in informal emails (e.g., do you know that?, I want to tell you …). The variation between the two tasks will be addressed again in RQ2.

Table 8

Results of Kruskal-Wallis tests (Task 2)

MDM Type	N	H	df	Sig.		Significant differences found between
Person marker	928	37.473	4	0.000	**	A1 and B1 (mean rank difference = 38.137, SE = 13.236, adjusted p=.000) A1 and B2 (mean rank difference = 49.816, SE = 13.314, adjusted p =.000) A1 and C (mean rank difference = 55.994, SE = 12.838, adjusted p =.000) A2 and B1 (mean rank difference = 38.913, SE = 12.389, adjusted p=..002) A2 and B2 (mean rank difference = 50.636, SE = 12.471, adjusted p=.000) A2 and C (mean rank difference = 56.769, SE = 11.962, adjusted p=.000)
Logical connective	860	8.018	4	0.091		
Relational marker	732	19.111	4	0.001	**	A2 and B2 (mean rank difference = 24.046, SE = 7.504, adjusted p=.014) A2 and C (mean rank difference = 26.765, SE = 6.907, adjusted p=.001)

Table 8 (Continued)

Results of Kruskal-Wallis tests (Task 2)

MDM Type	N	H	df	Sig.		Significant differences found between
Hedge	372	4.58	4	0.333		
Emphatic	641	4.104	4	0.392		
Attitude marker	591	5.096	4	0.278		
Sequencing	91	21.206	4	0.000	**	A1 and B2 (mean rank difference = 26.885, SE = 8.085, adjusted p=.009) A1 and C (mean rank difference = 28.500, SE = 8.139, adjusted p=.005) A2 and C (mean rank difference = 18.300, SE = 6.231, adjusted p=.033)
Announce goal	196	18.204	4	0.001	**	A1 and B2 (mean rank difference = 71.870, SE = 18.292, adjusted p=.001) A1 and C (mean rank difference =59.288, SE = 17.553, adjusted p=.007)
Evidential	47	1.765	3	0.623		
Code Gloss	40	4.263	4	0.372		
Topic shift	19	2.167	4	0.705		
Endophoric	6	0	2	1.000		
Email open	129	5.432	4	0.246		
Email close	346	56.427	4	0.000	**	A1 and C (mean rank difference = 75.880, SE = 25.959, adjusted p=.035) A2 and C (mean rank difference = 49.578, SE = 16.566, adjusted p =.028) B1 and B2 (mean rank difference = 42.128, SE = 15.491, adjusted p=.065) B1 and C (mean rank difference = 98.182, SE = 14.075, adjusted p=.000) B2 and C (mean rank difference = 56.054, SE = 12.671, adjusted p=.000)
Saltation	675	0	4	1.000		

We have so far reported the results of RQ1 regarding the use and accuracy of MDMs between the informal and formal email tasks by the L2 writers across the proficiency spectrum. In the next section and onwards, we present the results of RQ2 regarding the extent to which outcomes of RQ1 can be used to build AI models in relation to the classifier performance for whether an MDM or Not, the classifier performance for accurately or inaccurately used MDM, and the impact of task dependency.

RQ2: Classifications using a transformer-based AI Model

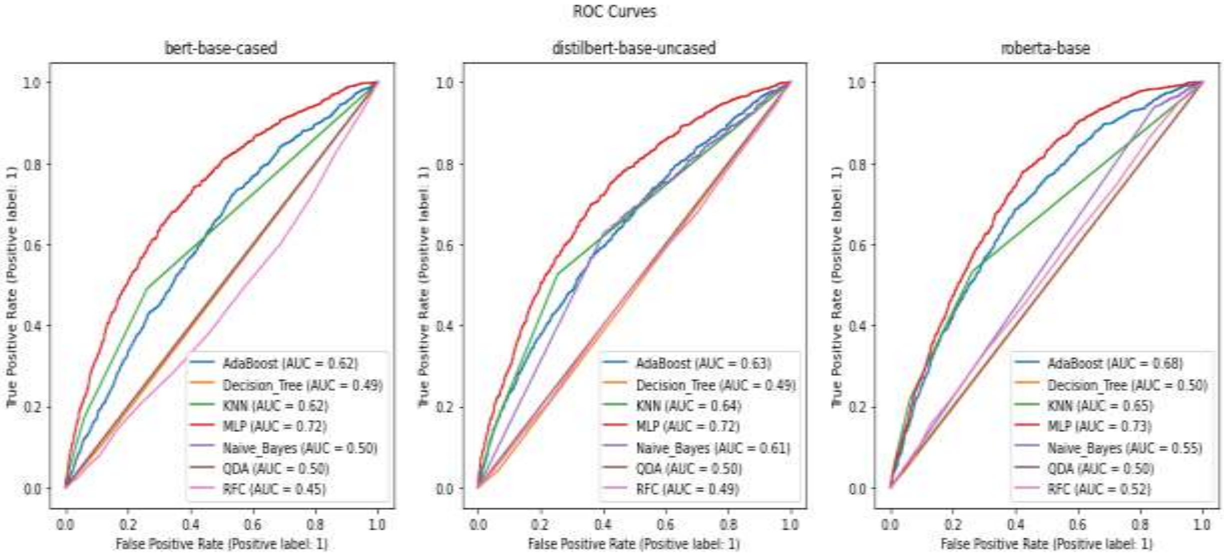
Word Embeddings

For all experimentation, our dataset was split into a train, validation and test set with 60%, 20% and 20% of the dataset belonging to each set respectively.

After consideration of benchmark performance and training times, the performance of the three shortlisted transformers BERT, RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) were evaluated and compared. An initial 10% of the overall dataset was used in order to reduce training times at this stage. Multiple out-of-the-box classifiers (i.e., using default learning parameters) were used together with the transformer architectures, to avoid the need to fine-tune classifier parameters. The resulting receiver operating characteristic (ROC) Curves are shown in Figure 6.

Figure 6

ROC curves of the 10% dataset of different transformer architectures evaluated on a range of out-of-the-box classifiers. Area under curve (AUC) scores are also shown



The distribution of resulting AUC scores[2] were relatively similar across the embedding methods. Among them, the MLP and the boosting algorithm, AdaBoost, showed the best out-of-the-box performance (ROC curves with points closer to the upper left of the graph show better performance due to their lower False Positive Rate for a given True Positive Rate).

Due to the limited variation in performance between embedding methods, BERT embeddings (far left in Figure 6) were selected for use. Owing to faster training times, a variation of the boosting algorithm, LightGBM (Ke et al., 2017), was used to evaluate performance. From this point forward, the entirety of the labelled dataset was used for experimentation unless otherwise specified.

Classifier Performance for Whether an MDM or Not

When evaluated on the 20% test set, with the other lines on the graph showing either classifiers trained on resampled datasets or fine-tuned classifiers. From the Precision-Recall curve[3] and the ROC, we can see an apparent trade-off between the opposing classes as the classification threshold is varied. However, even with the introduction of both under-sampling and oversampling techniques (i.e., ADASYN and SMOTE), we see limited changes in the metrics.

[2] AUC stands for "Area under the ROC Curve", which measures the area underneath the ROC curve from (0,0) to (1,1). The AUC score can also be thought of as the probability that a randomly chosen positively labelled prediction ranks higher than a negatively labelled prediction. An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate (TPR) and False Positive Rate (FPR).

[3] According to Shafi (2022), precision-recall is a useful measure of success of prediction when the classes are very imbalanced. Precision is calculated by dividing the true positives by anything that was predicted as a positive. Recall (or True Positive Rate) is calculated by dividing the true positives by anything that should have been predicted as positive. The precision-recall curve shows the trade-off between precision and recall for different thresholds. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

We now describe the evaluation of classifier performance for RQ2.1. The blue lines (LGB) in Figure 7 show the performance of the trained LightGBM classifier on the original dataset.

Figure 7

Precision-Recall curves and ROC curves of the baseline classifier, classifiers trained on resampled datasets and fine-tuned classifiers. Average precision (AP) and AUC scores are shown.

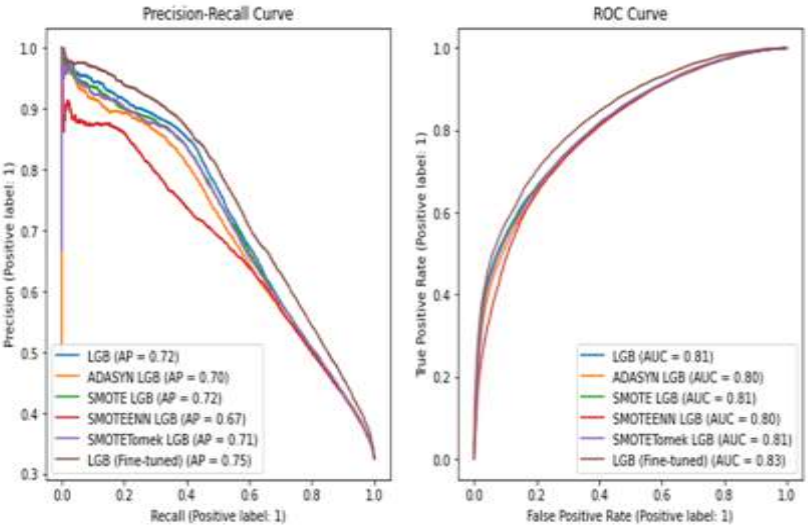


Table 9 shows the metrics for evaluating the accuracy of predictions by different resampling methods (i.e. Precision for class 0 and 1 (Pre0 and Pre1), Recall for class 0 and 1 (Rec0, Rec1), G-Mean and F1-Macro). For all the values, the closer to 1, the better the predictions are.

The addition of fine-tuning showed significant improvements in the recall of class 1, however, this came at the cost of a lower precision. According to the classification probability histogram of class 1 in Figure 8 between our baseline classifier and our fine-tuned model, LGB (Fine-tuned), we can see a definitive shift in the overlap between the classification of our two opposing classes. In the baseline model, class 1 exhibited a bimodal distribution. In our fine-tuned model, we see a rise in the bias of the classifier towards class 1, which results in higher confidence of correctly identifying a MDM at the cost of a rise in data from class 0 having a higher probability of being predicted as belonging to class 1 (as demonstrated by the larger tail of the blue line in the fine-tuned model). Whilst the confidence of our model has improved through fine-tuning, a tradeoff still exists and there is room for improvement.

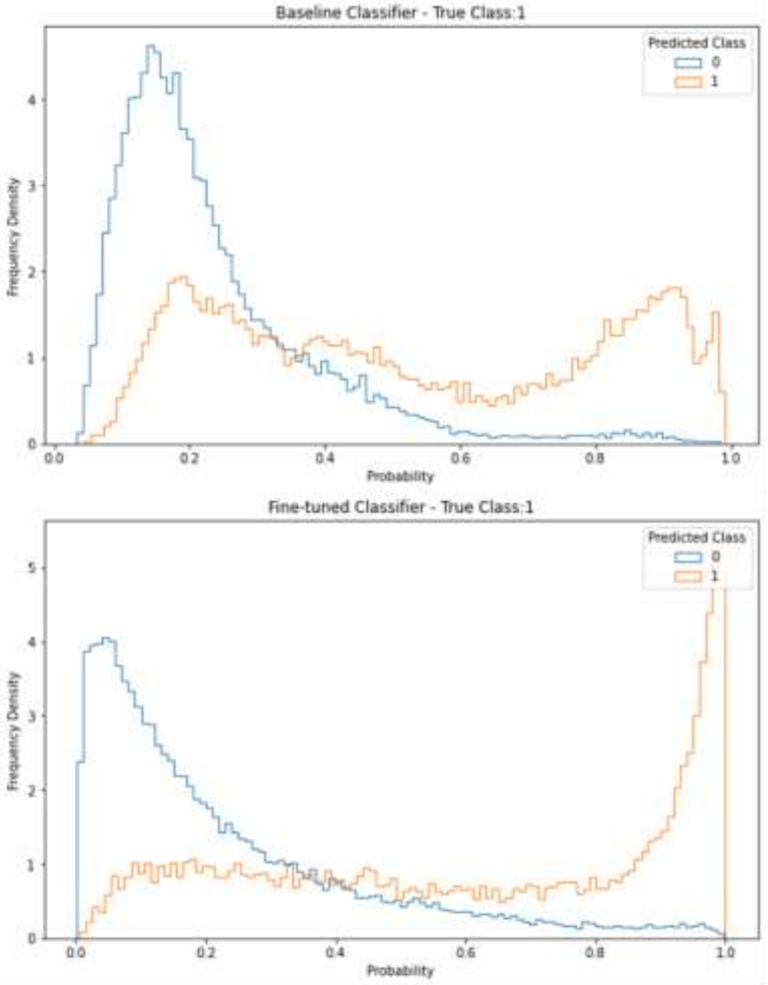
Table 9

Metrics for evaluating the accuracy of predictions (RQ2.1)

Classifier	Pre$_0$	Rec$_0$	Pre$_1$	Rec$_1$	G-Mean	F1-Macro
LGB	0.79	0.79	0.94	0.48	0.67	0.73
ADASYN LGB	0.83	0.60	0.79	0.66	0.72	0.72
SMOTE LGB	0.82	0.63	0.82	0.64	0.72	0.73
SMOTEENN LGB	0.88	0.47	0.55	0.84	0.68	0.64
SMOTETomek LGB	0.82	0.63	0.83	0.63	0.72	0.73
LGB (Fine-tuned)	0.82	0.70	0.88	0.61	0.73	0.75

Figure 8

Frequency density of the probability of prediction of the baseline and fine-tuned classifier for class 1

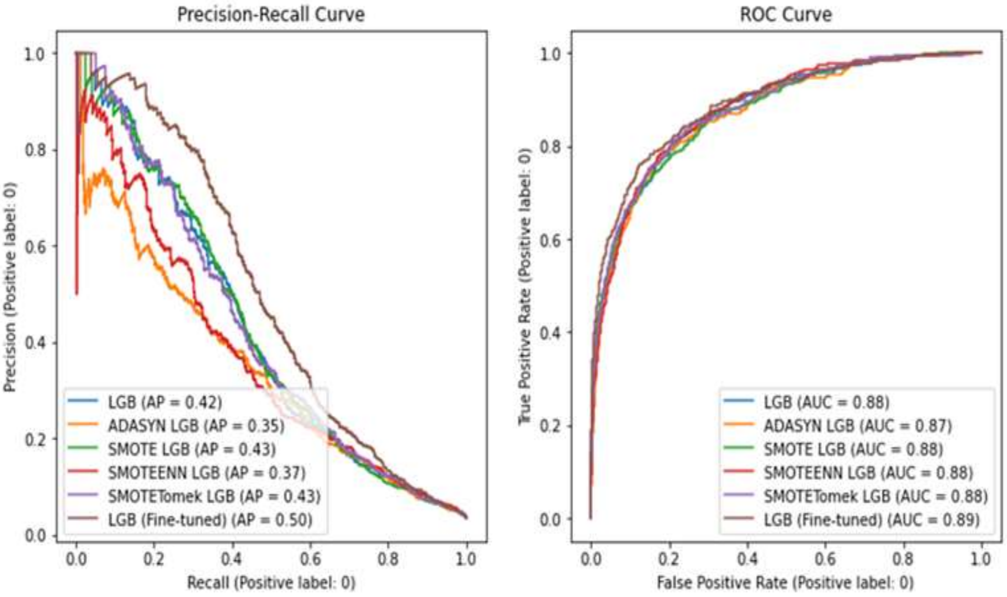


Classifier Performance for Accurately or Inaccurately Used MDM

This section describes the evaluation of classifier performance for RQ2.2. Once again, the blue lines in Figure 9 shows the performance of the trained LightGBM classifier on the original dataset, with the other lines on the graph involving LGB classifiers trained on resampled datasets or otherwise represent fine-tuned classifiers.

Figure 9

Precision-Recall curves and ROC curves of the baseline classifier, resampled datasets and fine-tuned classifiers. Average precision (AP) and AUC scores are shown.



From Table 10, we see all classifiers performing exceedingly well in predicting accurate MDM (class: 1), however they had very little success in confidently predicting inaccurate use cases (class: 0). The introduction of resampling techniques increased the recall of the classifier at heavy cost to the precision when compared with the baseline classifier. This trade-off is made even more apparent when looking at the Precision-Recall curve and AP scores shown in Figure 8, with an average AP score 0.40 amongst all classifiers.

Table 10

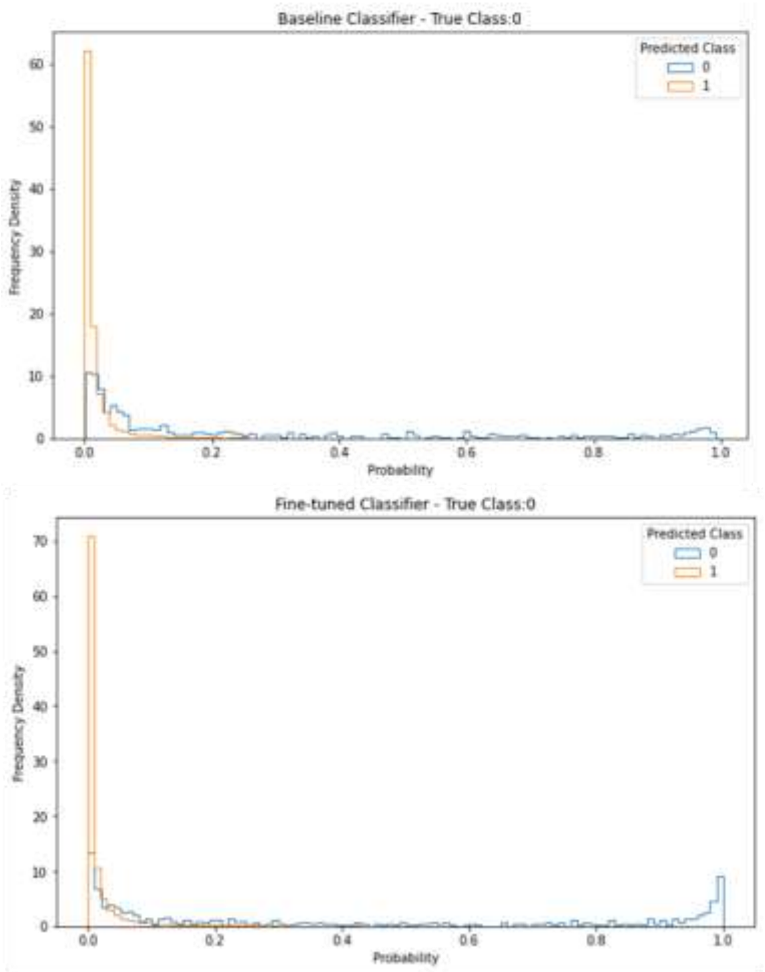
Metrics for evaluating the accuracy of predictions (RQ2.2)

Classifier	Pre$_0$	Rec$_0$	Pre$_1$	Rec$_1$	G-Mean	F1-Macro
LGB	0.74	0.24	0.97	1.00	0.49	0.67
ADASYN LGB	0.27	0.56	0.98	0.94	0.73	0.66
SMOTE LGB	0.29	0.55	0.98	0.95	0.72	0.67
SMOTEENN LGB	0.19	0.68	0.99	0.89	0.78	0.61
SMOTETomek LGB	0.30	0.55	0.98	0.95	0.72	0.68
LGB (Fine-tuned)	0.67	0.38	0.98	0.99	0.62	0.74

Furthermore, looking at the classification probability histograms for class 0 shown in Figure 7, we see improved performance in the discriminative ability of our fine-tuned classifier, LGB (Fine-tuned), when compared to baseline performance. However, a significant portion of our inaccurate MDM use cases were predicted as having a very low probability of belonging to class 0. Due to the aforementioned overwhelming imbalance in our dataset (roughly 24 accurate use cases for every inaccurate use case), the consequences of adjusting the classification threshold were significant. Whilst only a small percentage of either class was affected by changes to the classification threshold, the class imbalance resulted in a much larger absolute value of accurate use cases being misclassified as the threshold decreases. As such, attempting to include these low confidence occurrences is not feasible and our classifier, as a result, is only able to predict a portion of inaccurate use cases well.

Figure 10

Frequency density of the probability of prediction of the baseline and fine-tuned classifier for class 0



Task Dependency

In addition, we tested the classifiers dependence on a given task by training classifiers on one task exclusively, whilst using data from the other task to test its performance. We show results from both research questions RQ2.1 and RQ2.2 on classifiers trained with all the training split data (the baseline classifier) alongside classifiers trained solely on either Task 1 or Task 2 training data whilst using the unused task for testing.

For RQ2.1, our results from Figure 11 and Table 11 show that a classifier trained solely on Task 2 data, classifier 2, is better capable of generalizing on an unseen task than a classifier trained solely on Task 1 data, classifier 1. Classifier 2 outperforms the baseline classifier in several areas as shown by our established evaluation metrics.

Figure 11

Precision-Recall curves and ROC curves of the baseline classifier and classifiers trained and tested on differing classes

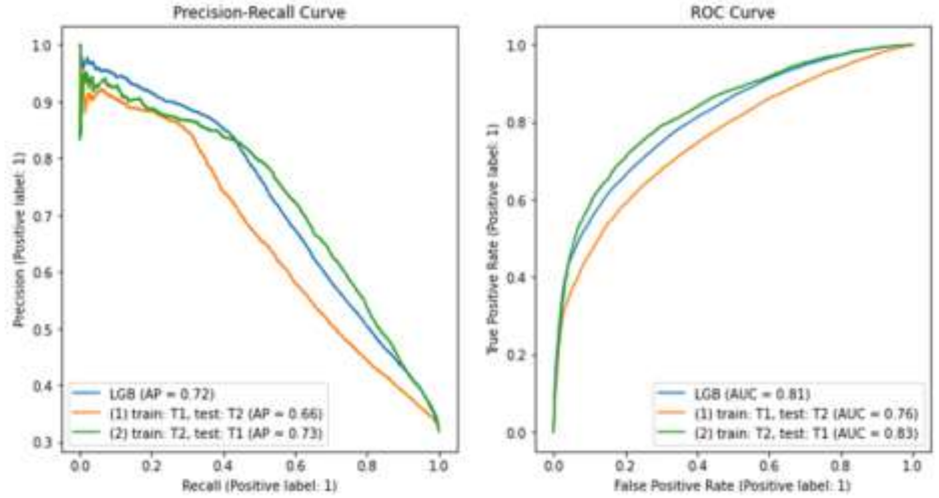


Table 11

Metrics for evaluating the dependence of the classifier on a given task (RQ2.1)

Classifier	Pre_0	Rec_0	Pre_1	Rec_1	G-Mean	F1-Macro
LGB	0.79	0.79	0.94	0.48	0.67	0.73
(1) train: T1, test: T2	0.77	0.71	0.91	0.44	0.63	0.69
(2) train: T2, test: T1	0.81	0.76	0.92	0.55	0.71	0.75

However, for RQ2.2, the inverse is true; with classifier 1 giving better performance than classifier 2, as shown in Figure 12 and Table 12. Classifier 2 severely underperforms in all areas pertinent to the classification of class 0 when compared with our baseline and classifier 1.

Figure 12

Precision-Recall curves and ROC curves of the baseline classifier and classifiers trained and tested on differing classes

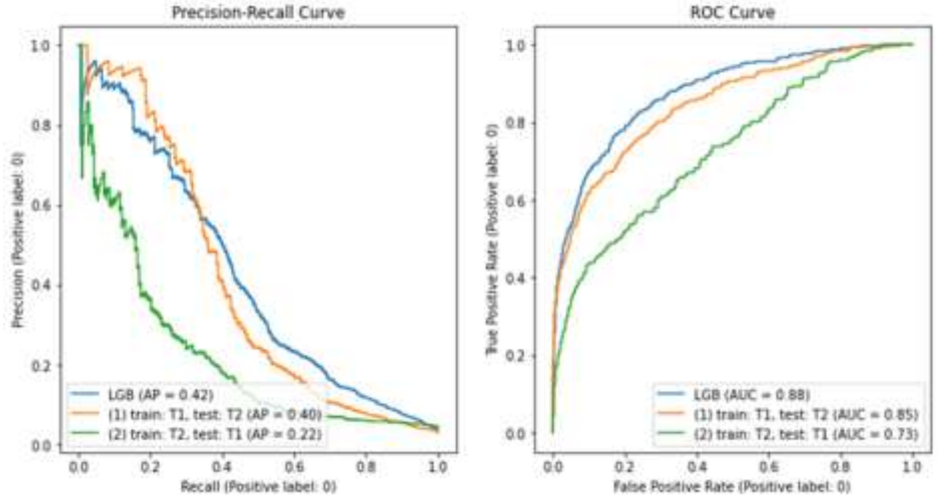


Table 13

Metrics for evaluating the dependence of the classifier on a given task (RQ2.2)

Classifier	Pre_0	Rec_0	Pre_1	Rec_1	G-Mean	F1-Macro
LGB	0.74	0.24	0.97	1.00	0.49	0.67
train: Task 1, test: Task 2	0.80	0.23	0.98	1.00	0.48	0.67
train: Task 2, test: Task 1	0.61	0.10	0.96	1.00	0.31	0.57

To summarise, the detection of MDMs (RQ2.1) is much less dependent on a given task than detecting the appropriacy of use (RQ2.2). However, with only two tasks to test dependency, a definitive conclusion cannot be drawn for either of our research questions as to whether our classifiers can be said to be task-agnostic.

Discussion

Questions and Limitations

Findings of RQ1 show that, different from the general notion that discourse competence emerges at the B2/C1 threshold, L2 writers start to display discourse competence (or at least in terms of the use of MDM) even at CEFR A1, A2 and B1 levels. However, C level candidates are clearly most aware of the difference between the informal and formal essay tasks and most able to adjust their use of MDM accordingly. Previous studies (e.g., Knoch et al., 2014) found that lower-level writers used proportionally more MDMs overall than more proficient writers, or that more proficient writers used a significantly wider range of MDMs than lower-level writers (e.g., Bax et al., 2019). However, this is not the case in the current study. This indicates the importance of considering genre variation in research of MDMs (or discourse competence) in L2 writing. For example, the task investigated in Knoch et al., (2004) was an essay task whereas informal and formal emails were investigated in the current study. Emails are served as communication tools whereas essays are often used to displace understanding of a certain topic and to introduce one’s (new) perspectives. It can be argued that the main goal of emails is to communicate with a specific (usually known) audience, whereas essays are often expository and argumentative aiming to address a wider unknown audience. Discourse competence in email writing is essential to ensure effective communication and to conform to the real-world expectation of audience awareness and appropriateness. For example, as shown in this study, an informal email to a friend allows for a more relaxed, friendly style through the use of interpersonal markers (e.g., person marker, relational marker, emphatic an attitude marker). Textual and genre-specific MDMs are less important, especially in short informal emails. Formal emails, even shorter ones, required more careful thought on tone and structure. As demonstrated by the L2 writers in this study, this can be achieved by the use of interpersonal markers (e.g., hedge), textual markers (e.g., sequencing and achieving goals). Genre-specific MDMs in formal emails (e.g., saltation, email open and email close) are deemed necessary. The findings show that even the lowest-proficiency L2 writers demonstrated some evidence of discourse competence in email writing through the use of MDMs. Nevertheless, they may struggle to demonstrate it in essay writing where textual and genre-specific MDMs might play a larger role. Perhaps there needs to be a different framework for evaluating the accuracy of MDM use according to the genres and levels of formality required. This also indicates the benefits of evaluating discourse competence through a detailed analysis of the use of individual MDMs in different categories in order to capture the nuanced evidence of the development of discourse competence in L2 writers.

In terms of accuracy of use, the findings show that L2 writers across the CEFR levels seem to use most of the MDMs accurately, more so in the formal email when use of MDM was more formulaic than the informal email. Nevertheless, several MDM types appeared to be exceptions, showing an upward trend as the levels went up: hedge and announce goals on Task 1 and sequencing, announce goals and email close on Task 2. These MDM categories tend to involve more than a single word and thus there is a higher chance for lower-proficiency writers to make mistakes. Higher-proficiency writers, on the other hand, can demonstrate their discourse competence through accurate use of these phrasal/clausal MDMs.

Nevertheless, the overall uniform pattern in terms of accuracy of MDM use led to extremely small numbers of inaccurate MDM use. This further led to imbalanced data distribution for automated classification which led to attempting various resampling methodologies in RQ2.1 and RQ2.2, since a heavily imbalanced binary classification task can lead to the classifiers exhibiting heavy bias to the majority class. The tested resampling methods all performed relatively similarly, increasing the precision of the minority class as well as the recall of the majority class. Fine-tuning of the basic classifier was also attempted to improve performance and, whilst not consistently resulting in the highest individual class precision or recall scores, resulted in the highest F1-macro score. One might wonder: would focusing on MDMs be useful at all for classifying (and predicting discourse competence in L2 writing further in the future) and overall proficiency levels of the candidates? The simple answer is that measures of MDM would not be very useful alone in building automated essay scoring systems; they need to be incorporated with other features.

Moreover, the results for RQ2.1 and 2.2 clearly showed that there is a tradeoff between the classification and misclassification of 1 and 0. Of course, the ideal scenario is that we do not have false positives and false negatives, but it was not possible to achieve with the current dataset. We therefore need to consider what should be prioritized by asking the following questions:

- Do we worry more about misclassifying accurate use cases as inaccurate or the other way around?
- Do we worry more about identifying MDMs at the tradeoff of misclassifying?

Would judging accurately used MDM to be inaccurate be more damaging to the candidate’s scores? Or would judging inaccurately used MDM to be accurate be worse for candidates? Misclassifying accurate and inaccurate MDM use can have significant consequences for learner assessment. When accurate MDM usage is misclassified as inaccurate, learners may receive unfairly low scores, leading to a loss of confidence in their writing abilities. For instance, a learner who effectively uses transition markers like "however" or "therefore" might be penalized if the system misinterprets their usage as incorrect, discouraging them from experimenting with more advanced language structures. Conversely, misclassifying inaccurate MDM usage as accurate can result in inflated scores, giving learners a false sense of mastery. Ultimately, the relative importance of these two errors depends on the candidates’ learning stage. For beginners, it may be crucial to minimize errors where correct usage is judged as incorrect, because this type of error can discourage those who are making genuine progress and may lead them to question their understanding of the writing skills they are developing. On the other hand, for more advanced learners, the focus should perhaps shift to identifying and addressing incorrect usage to help students achieve greater accuracy in their writing. With a larger number of data points for our minority class (inaccurate use cases), weighting techniques could be explored to minimize errors of a given type depending on the writer's CEFR level.

One issue to bear in mind is that, although MDMs are found vitally important in articles, journals and newspapers (e.g., Hyland, 2005; Dafouz-Milne, 2008), they may carry slightly different ‘weight’ in email messages–more personal, shorter pieces of writing that are addressed to one person. Therefore, in order to explore answers to this set of difficult trade-off questions, we will need to scrutinize the construct of discourse competence that is being measured by the commonly used email tasks in L2 writing assessments and how the ability to use MDM is considered to contribute to the construct. For example, would having a specific MDM accuracy framework for informal writing than formal genres be appropriate or viable? It would also be important to consider the balance between accuracy, complexity and appropriateness.

Another question is how different the results might be if we are to expand the coding scheme to tap into pragmatic appropriacy. As described in the methods section, we employed a dichotomous coding scheme in this feasibility study, which required making unambiguous judgements on the use of MDM, narrowing the range of errors that we coded for. They were mostly grammatical errors that surround the MDM but considering that the main role of MDM is to involve the message recipients, signpost, and communicate the writer's stance effectively, grammatical accuracy may only play a small part in it. Coding for pragmatic appropriacy will require polytomous coding (for example, 3: 'appropriate', 2: 'acceptable', 1: 'not appropriate') in order to capture other important aspects of MDM use. Furthermore, it might contribute to differentiate better between tasks with recipients with differential social status (i.e. a friend/classmate or a school manager), which could not be achieved with the dichotomous coding scheme in this study. Since the two email tasks used in the current study are designed to tap into the ability to use language according to different situations and recipients, coding for pragmatic appropriacy appears to be the clear next step forward. This can form one of the additional features to incorporate in automated classification, although it will be a more resource-intensive study which requires a bigger labelled dataset.

Conclusion and Future Work

In summary, the contributions of this study are three-fold. Firstly, it offers valuable insights within the context of Explainable AI. Transparency and explainability in automated scoring models are critical for ensuring fairness and stakeholder understanding in language tests. By integrating MDM usage and accuracy into the scoring framework, this research moves beyond frequency-based evaluation. The finding that the range of MDM use and accuracy are highly task-dependent highlights the need for task-specific tuning rather than relying on generalized models, which contribute to the design of AI-based systems that are both practical and explainable in educational applications.

Moreover, given that MDM usage reflects discourse competence, this study also makes significant contributions to the current understanding of L2 writing development. While previous research has often underestimated the discourse competence of lower-proficiency learners, this study demonstrates that even these learners exhibit evidence of discourse competence through their accurate use of MDMs as well as their choice of MDMs in response to genre. This finding suggests that L2 learners may develop aspects of discourse competence earlier than traditionally assumed, which offers a new perspective on how discourse analysis can be integrated as a core element in AEEs.

From the perspective of expanding the construct representation in automated scoring systems, this study provides a critical examination of the limitations of many AEE models, which have heavily relied on vocabulary and grammar features. By exploring the feasibility of incorporating MDMs as predictive features, this research demonstrates the potential for construct expansion. However, the task dependency of accuracy classification and the data imbalance—caused by the predominance of correct MDM usage—present challenges that need to be addressed.

There have been many AEE studies utilizing both handcrafted features as well as features obtained through the use of a deep neural network, however no definitive answer exists as to whether either will perform better on any given dataset. Several papers (e.g., Liu et al., 2019; Lin et al., 2020) suggest improved performance by training classifiers on a combination of both such feature types. One of the handcrafted features in the future could incorporate 'whether or not certain MDM are used' in the scripts. All the codes assigned in this study by the human coders indicated if MDM was used; their absence was not coded and therefore was not taken into account in automated classification. However, given accuracy use was consistently high across levels, it is possible that the absence of certain MDM could be good indicators that differentiate between the levels of proficiency. For example, the accurate use of MDM for announcing goals increased with the levels especially in the formal email task (RQ1). The absence of this MDM type might also differentiate effectively between CEFR levels, contributing to improving classifier performance. Another handcrafted feature could be formed using the genre of the text that we are attempting to analyze. Whilst transformer-based embeddings are context aware, explicitly introducing the genre of the text into the set of features fed into a classifier could allow for better performance over the detection of certain MDMs within said genres.

Another potential limitation of the current methodology surrounds the idea of phrasal MDMs. These are phrases consisting of words that in isolation are not MDMs, but when combined together are classed as an MDM. Currently, our transformer-based architecture returns a word embedding for each individual word that was input into it and these individual word embeddings are then classified. Whilst these embeddings are context dependent, labelling each individual word within that phrase as an MDM in its own right and training a binary classifier based on this data (RQ2.1) may be misleading and lead to the model underperforming when generalizing to the larger dataset. Future work could involve creating an extension of our word embedding methodology in which we form phrasal embeddings to represent our phrasal MDMs to assess the impact of this decision within our current framework. Phrasal embeddings could be formed using the word embeddings associated with words in the phrase we wish to obtain phrasal embeddings for (e.g., an average of all relevant word embeddings to obtain a singular phrasal embedding).

Finally, as previously mentioned, a dichotomous coding scheme was chosen for this particular study. Extending this coding scheme to a polytomous coding scheme may better allow a machine-learning model to discern the nuances associated with MDM usage. For example, as mentioned in Procedures for RQ1, less accurate MDM usage was not counted as inaccurate usage for our binary classification task. A polytomous coding scheme could allow a classifier to better distinguish such levels of MDM use since we currently have both less accurate use cases and extremely accurate use cases both having the same label. Additionally, this would somewhat reduce the class imbalance since data points would be moved from the current majority class (accurate use cases) to any intermediary labels included in the new coding scheme. However, it is still likely that collecting more data involving inaccurate use cases would allow the classifier to better establish class boundaries. Whilst the basis of the methodology would remain the same, using transformer-based architectures to extract word embeddings, our binary classification task could be extended to a multi-class classification task. Given the lack of research directly concerning MDMs in machine learning literature, we cannot definitively say how any of these factors would affect performance for our given tasks; however, all are valid areas that should be investigated in further work. Another area of research is how our results would impact skilled labour across various sectors, domains and contexts from construction, manufacturing up to medicine.

Declarations

Gen-AI use : The authors of this article declare (Declaration Form #: 212241332) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Ethical Approval: Ethical approval for the study was received from University of Bedfordshire, CRELLA Research Institute Ethics Committee on 14.04.2022.

Author Contribution: Sathena Chan: funding application, project lead, conceptualization, methodology, data analysis, writing, revision and editing. Manoranjan Sathyamurthy: methodology (RQ2), data analysis (RQ2), writing and revision. Chihiro Inoue: funding application, conceptualization, methodology (RQ1), data analysis (RQ1), writing, revision and supervision. Michael Bax: funding application, conceptualization, review and editing. Johnathan Jones: data analysis (RQ1). John Oyekan: funding application, conceptualization, methodology (RQ2), data analysis (RQ2), writing, revision and supervision.

Funding: The research was funded by the British Council Aptis Research Grants 2021.

Consent to Publish: Written consent was sought from each author to publish the manuscript.

Competing Interests: The authors have no competing interests to disclose.

References

Adel, A. (2006). Metadiscourse in L1 and L2 English. John Benjamins Publishing. https://doi.org/10.1075/scl.24

Bachman, L. and Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Barkaoui, K. (2016). What changes and what doesn't? An examination of changes in the linguistic characteristics of IELTS repeaters' Writing Task 2 scripts. *IELTS Research Reports Online Series*, vol. 2016/3, 1–55.

Bax, S., D. Waller and Nakatsuhara, F. (2019). Researching L2 writers' use of MDM at intermediate and advanced levels, *System*, *83*, 79-95. https://doi.org/10.1016/j.system.2019.02.010

Breiman (2001). Random Forests, *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brezina V. & Gablasova, D. (2015) Is There a Core General Vocabulary? Introducing the *New General Service List*, *Applied Linguistics*, 36(1), 1-22, https://doi.org/10.1093/applin/amt018

Burneikaitė, N. (2008) "Metadiscourse in Linguistics Master's Theses in English L1 and L2", Kalbotyra, 59, pp. 38–47. doi:10.15388/Klbt.2008.7591.

Camiciottoli, B. C. (2003). Metadiscourse and ESP reading comprehension. *Reading In A Foreign Language*, *15*(1), 28–44. https://nflrc.hawaii.edu/rfl/item/69

Carlsen, C. (2010). Discourse connectives across CEFR-levels: A corpus based study. In I. Bartning, M. Maatin, & I. Vedder (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and Language testing research* (pp. 191-210). European Second Language Association.

Chapelle, C. A. and Chung, Y-R. (2010). The Promise of NLP and speech processing technologies in language assessment. *Language Testing*, 27(3), 301–315. https://doi.org/10.1177/026553221036440

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321-357. https://doi.org/10.1613/jair.953

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press. https://doi.org/10.1017/CHOL9780521221283

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, *14* (5), 771-780. http://www.yorku.ca/gisweb/eats4400/boost.pdf

Crompton, P. (2012). Characterising hedging in undergraduate essays by Middle-Eastern students. *Asian ESP Journal*, 8(2), 55-78. http://asian-esp-journal.com/wp-content/uploads/2013/11/Volume-8-2.pdf

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2018). *BERT: Pre-training of deep Bidirectional Transformers for language understanding*. https://arxiv.org/pdf/1810.04805.pdf

Jarvis, S. (2013). 'Defining and measuring lexical diversity.' in S. Jarvis and M.H. Daller (eds.) *Vocabulary knowledge: human ratings and automated measures.* John Benjamins, pp. 13-44.

He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)* (pp. 1322-1328). IEEE. https//doi.org/10.1109/IJCNN.2008.4633969

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyland, K. (2005). *Metadiscourse*: Exploring Interaction in Writing. Bloomsbury Publishing.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems*, *30* (NIPS 2017). https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An Investigation of the Effect of Task Type on the Discourse Produced by Students at Various Score Levels in the TOEFL iBT ® Writing Test. *ETS Research Report Series*, *23*, 14–43. https://doi.org/10.1002/ets2.12038

Lee, J. J., & Deakin, L. (2016). Interactions in L1 and L2 undergraduate student writing: Interactional metadiscourse in successful and less-successful argumentative essays. *Journal of Second Language Writing*, *33*, 21-34. https://doi.org/10.1016/j.jslw.2016.06.004

Lialin, V., Deshpande, V. & Rumshisky, A. (2023). Scaling down to scale up: A guide to parameter-efficient fine-tuning. arXiv preprint arXiv:2303.15647. https://doi.org/10.48550/arXiv.2303.15647

Lin, W., Hasenstab, K., Chnha, G. M. & Schwartzman, A. (2020) *Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment*, *Scientific Reports*, *10*, 20336. https://doi.org/10.1038/s41598-020-77264-y

Lin, T., Wang, Y., Liu, X. and Qiu, X. (2022). A survey of transformers. *AI open*, 3, 111-132. https://doi.org/10.48550/arXiv.2106.04554

Liu, J., Xu, Y. and Zhu, Y. (2019) ‘Automated Essay Scoring based on Two-Stage Learning’, *arXiv [cs.CL]*. https://doi.org/10.48550/arXiv.1901.07744

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692

O’Loughlin, K. (2013). Investigating lexical validity in the Pearson Test of English Academic. Pearson Research Reports, p.1-21. https://www.pearsonpte.com/ctf-assets/yqwtwibiobs4/6iHE8HxuGJT3OMAgFoXxwV/88d7be6a2d43a9274d12fe7de838fef0/Investigating_lexical_validity_in_the_Pearson_Test_of_English_Academic-_Kieran_O___Loughlin.pdf

O'Sullivan, B., Dunlea, J., Spiby, R., Westbrook, C., and Dunn, K. (2020). Aptis General Technical Manual Version 2.2. https://www.britishcouncil.org/sites/default/files/aptis_technical_manual_v_2.2_final.pdf

Owen, N., Shrestha, P. and Bax, S. (2021). Researching lexical thresholds and lexical profiles across the Common European Framework of Reference for Language (CEFR) levels assessed in the Aptis test. *ARAGs Research Reports Online, AR- G/2021/1.*

Sanford, S. (2012). *A comparison of metadiscourse markers and writing quality in adolescent written narratives.* Missoula: Unpublished MSc thesis. The University of Montana.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108. https://doi.org/10.48550/arXiv.1910.01108

Schiffrin, D., Tannen, D., & Hamilton, H. (2001). *The handbook of discourse analysis.* Blackwell Publishers Ltd.

Zhang, H. (2004). The optimality of Naive Bayes. https://typeset.io/papers/the-optimality-of-naive-bayes-r4zge3fp91

Zhang, Z. (2016). Introduction to machine learning: k-nearest neighbors. *Annals of Translational Medicine, 4*(11), 218. https://doi.org/10.21037/atm.2016.03.37

Appendix

Appendix 1

Full list of MDMs analysed

Announce Goals (Frame marker)			
here I will	my purpose	the aim	I intend
I seek	I wish	I argue	I propose
I suggest	I discuss	I would like to	I will focus on
we will focus on	I will emphasise	we will emphasise	my goal is
in this section	in this chapter	here I do this	here I will
Code glosses			
put another way	for example	for instance	e.g.
i.e.	that is	that is to say	namely
in other words	this means	which means	in fact
Viz.	specifically	such as	
known as	defined as	called	
Endophorics			
see	noted	discussed below	discussed above
discussed earlier	discussed later	discussed before	section
chapter	fig	figure	table
example	page		
Hedges			
apparently	appear to be	approximately	assume
believed	certain extent	certain level	certain amount
could	couldn’t	doubt	essentially
estimate	frequently	generally	in general
indicate	largely	likely	mainly
may	maybe	might	mostly
often	perhaps	plausible	possible
possibly	presumably	probable	probably
relatively	seems	sometimes	somewhat
suggest	suspect	unlikely	uncertain
unclear	usually	would	wouldn’t
little	not understood	almost	
Logical connectives			
but	therefore	thereby	so
so as to	in addition	similarly	equally
likewise	moreover	furthermore	in contrast
by contrast	as a result	the result is	result in
since	because	consequently	as a consequence
accordingly	on the other hand	on the contrary	however
besides	also	whereas	while
although	even though	though	yet
nevertheless	nonetheless	hence	thus
leads to	or	and	
Relational markers			
incidentally	determine	consider	imagine
by the way	let us	let’s	lets
let	notice	our	recall
note	us	we	you
our	one’s	assume	think about
your			
Attitude markers			
admittedly	I agree	amazingly	unusually
accurately	correctly	curiously	disappointing
disagree	even	fortunately	have to

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2024; 15(Special issue); 348-360

Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness

Mo ZHANG*   Matthew JOHNSON**   Chunyi RUAN***

Abstract

AI scoring capabilities are commonly implemented in educational assessments as a supplement or replacement to human scoring, with significant interest in leveraging large language models for scoring. In order to use AI scoring capability responsibly, the AI scores should be accurate and fair. In this study, we explored one approach to potentially mitigate bias in AI scoring by using equal-allocation stratified sampling for AI model training. The data set included 13 open-ended short-response items in a K-12 state science assessment. Empirical results suggested that stratification did not improve or worsen fairness evaluations on the AI models. BERT based AI scoring models resulting from the stratified sampling method but trained on much less data performed comparably to models resulting from simple random sampling in terms of overall prediction accuracy and fairness on the subgroup level. Limitations and future research are also discussed.

Keywords: AI scoring, educational assessment, large language model, sampling, prediction accuracy, fairness

Introduction

AI scoring capabilities are commonly implemented in educational assessments as a supplement or replacement to human scoring. For example, AI scoring has been used to score open-ended text responses in various content domains (e.g., math, reading, writing, science, speaking) and assessments with varying levels of scale and stakes, including PTE English, TOEFL iBT, GMAT, GRE, LSAT, and certification/licensure tests such as Praxis, as well as many K-12 state-level assessments (e.g., Kentucky Summative Assessment). The literature on AI scoring has grown substantially in the past 10 to 20 years. Bennett and Zhang (2016) considered AI (or automated) scoring as "machine grading of constructed responses that are generally not amenable to exact-matching approaches because the specific form(s) and/or content of the correct answer(s) are not known in advance." An AI scoring algorithm is a computational procedure used in educational testing to predict or determine scores for test items or responses automatically. These algorithms typically use natural language processing and statistical or machine learning techniques to generate the predicted scores based on patterns or associations found in the data.

In early examples of AI scoring such as automated essay scoring, the AI score is usually a weighted combination of a small set of well-defined linguistic features, such as grammatical accuracy, vocabulary sophistication, sentence structure, and so forth, and these features are carefully evaluated by content experts to closely align to and cover the construct of measurement. The scoring algorithms tend to be white-box or gray-box models such as decision trees, linear regressions, and k-means. For these earlier approaches to AI scoring, the features used in the model are construct-relevant, the weights given to

* Senior Research Scientist, Educational Testing Service, New Jersey-USA, mzhang@ets.org, ORCID ID: 0000-0003-2689-2089
** Principal Research Director, Educational Testing Service, New Jersey-USA, msjohnson@ets.org, ORCID ID: 0000-0003-3157-4165
*** Principal Research Data Analyst, Educational Testing Service, New Jersey-USA, cruan@ets.org, ORCID ID: 0009-0009-3073-229X

To cite this article:
Zhang, M., Johnson, M. & Ruan , C. (2024). Investigating Sampling Impacts on an LLM-Based AI Scoring Approach: Prediction Accuracy and Fairness*, Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 348-360. https://doi.org/10.21031/epod.1561580

Received: 4.10.2024
Accepted: 12.11.2024

each feature can be extracted, and the reasoning from the features can be tracked. In this case, the scores are highly explainable and interpretable.

As generative AI has surged in popularity and revolutionized various sectors in the society, interest has increased in leveraging large language models (LLMs) for scoring (Chamieh, Zesch, & Giebermann, 2024; Lee, Latif, Wu, Liu, & Zhai, 2024; Lubis, Putri, et al., 2021; Kortemeyer, 2024; Oka, Kusumi, & Utsumi, 2024; Whitmer et al., 2021). Using LLMs for scoring is particularly relevant to assessing content and reasoning in areas in which traditional approaches have fallen short. Even though white- or gray-box models have great interpretability, their prediction accuracy is usually lower compared to black-box models such as transformer-based models (e.g., GPT, BART), deep learning, and neural networks (Ali, Abuhmed, El-Sappagh, et al., 2023; Kumar, Dikshit, & de Albuquerque, 2021). However, as models increase in complexity, interpretability diminishes substantially because millions of parameters are estimated to generate a score. For example, LlaMa 3.1 (released on 06/23/2024 by Meta AI) has 405 billion parameters. Although significantly smaller, the BERT$_{\text{BASE}}$ model (by Google AI) used in this study still has about 110 millions parameters.

In order to use AI scoring capability responsibly, the scores and the scoring process should follow standards in educational testing. There are several entries in the testing standards jointly published by APA, AREA, and NCME that are specifically about AI scoring. For example, Standard 3.8 states that "(AI) scoring algorithms need to be reviewed for potential sources of bias. The precision of scores and validity of score interpretations resulting from automated scoring should be evaluated for all relevant subgroups of the intended population" (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). This standard highlights two core principles in responsible use of AI in educational assessment: AI scores should be accurate and AI scores should be fair. Most of published research to date on AI scoring using LLMs has emphasized prediction accuracy of the models with little discussion of fairness. In Johnson and Zhang (2024), the authors argued that the accuracy of AI is only one component of its responsible use in education and demonstrated that there may be inherent or implicit biases in LLMs that will lead to unfairness in AI scoring. In this study, we conducted an exploratory analysis to investigate whether choices of sampling methods can help mitigate biases in LLM-based AI models.

Statement of Research Problem

Experts from various disciplines have identified, examined, and discussed social, cultural, and gender biases present in pretrained LLMs; see Ayoub et al. (2024); Ma, Scheible, Wang, and Veeramachaneni (2023); Manvi, Khanna, Burke, Lobell, and Ermon (2024); Navigli, Conia, and Ross (2023); Bai, Wang, Sucholutsky, and Griffiths (2024), and Caton and Haas (2024), to name a few. Inherent biases in LLM models are deeply rooted in the data used for their training. These models absorb, internalize, and propagate any biases and stereotypes present in their training data sets, thereby making this issue rather complex. In their recent work, Johnson and Zhang (2024) found that GPT-4o can predict the racial/ethnic group membership of a writer of an essay response better than GPT-4o can score using a zero-shot approach. In order to improve prediction accuracy, a common practice is to fine-tune pretrained LLMs for downstream tasks. The fine-tuning process involves a selection of a pretrained model, preparation of the data, (iterative) model training, and evaluation of operational deployment in which preprocessing of the data is a critical step. Chu, Wang, and Zhang (2024) summarized four stages in the AI model development process that can be adjusted to mitigate inherent bias: (a) preprocessing, (b) in-training, (c) intraprocessing, and (d) postprocessing (in which the authors suggested "data augmentation" as one way to mitigate bias in the preprocessing stage). The goal of data augmentation is to ensure a balanced representation of training data across various subgroups (social, cultural, gender, age, religion, etc.) in the target population. In the field of machine learning, data augmentation, artificially increasing the size of a data set by applying transformations to the train data (Chhabra, Singla, & Mohapatra, 2022), is a common technique. In image recognition and computer vision, transformation techniques include rotating, flipping, or changing the contrast or brightness of images. In text classification, transformation techniques include random deletion or insertion (of words or characters), sentence shuffling, synonym replacement, and so forth. Another approach to achieve a balanced training data is the equal-allocation stratified sampling technique, which effectively down-weights the larger subgroups by oversampling smaller subgroups in the population. Specifically, given a population P that can be divided into G nonoverlapping subpopulations or strata G_1, $G_{2\dots}G_g$, a sample s_g of size n_g is taken within each stratum g independently from one stratum to another. Let $n = \sum_{g=1}^{G} n_g$ be the total sample size. In equal-allocation stratified random sampling, n_g is constant for each stratum, that is, $\forall h \quad n_g^{eq} = \frac{n}{G}$. In this study, we examined this equal-allocation sampling approach in fine-tuning LLMs for scoring. Our premise is that if an AI model training data set is imbalanced, meaning a subgroup of test takers is underrepresented, the model may struggle to make accurate predictions for the underrepresented subgroup. In survey sampling, proportional stratification leads to mean estimators (which may be thought of as human mean scores) that are more accurate than those obtained under simple random sampling given the same sample size, while equal-allocation stratified sampling ensures a minimum level of precision in each stratum but does not lead to the best global mean estimates, particularly when the variabilities (or human-score standard deviations) are different between strata (Lohr, 2021). In our current AI scoring scenario, we are not only interested in a model's overall performance, but also in its performance within specific subgroups to ensure fairness. Therefore we still prioritized the equal-allocation stratified random sampling technique and compared it to simple random sampling when constructing the AI model training samples. Given there were implicit biases in pretraining LLMs that we fine-tuned for our scoring tasks, equal-allocation sampling was arguably one method to strike a balance between prediction accuracy and fairness in the case of AI scoring. Finally, we note the lack of systematic analysis of the impact of sampling when applying an LLM-based AI scoring approach in the field of educational assessment. For instance, earlier work on sample-size requirements for automated scoring were mostly conducted prior to the era of LLMs. The amount of data required to fine-tune a pretrained LLM sufficiently for scoring purposes remains uncertain, and, to our best knowledge, there are no published studies addressing this issue. Generally speaking, the literature has indicated that effectiveness of fine-tuning is highly task-specific and is dependent on the model size and data quality. However, we believe it is still worthwhile to fill the gap in the literature and explore this aspect by using the same data source, which includes the same assessment task, test-taker population, and pretrained LLM. Specifically, we addressed two research questions in this study:

1. How well do AI models resulting from different sampling methods predict human scores?
2. To what extent are the AI models resulting from different sampling methods fair? Does stratified sampling help improve fairness?

Methods

Data Set

We used a data set collected from a standardized state science assessment in the United States between 2020 and 2021. There are 13 open-ended questions (or prompts) included in this analysis. All the prompts were graded by trained human raters on a 2-point integer scale: 0, 1, and 2. About 30% of the responses in each prompt were graded by a second human rater to monitor the reliability of the human scores. The response length in characters across prompts are shown in Table 1. By design, the responses are relatively short; on average, the number of characters are between 120 to 200 characters across prompts (about 20 to 40 words). The total number of responses in each prompt ranged from 2,458 to 2,531.

Table 1

Response Length by Item (Character Count Means and Standard Errors)

	Item												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Mean	179.7	196.5	190.6	137.1	150.9	174.8	172.9	122.6	119.7	100.9	132.2	144.8	187.1
S.E.	3.20	3.02	3.72	6.53	2.47	8.27	2.69	2.24	2.29	1.72	2.26	8.62	3.42

To investigate fairness, we focused on race/ethnicity in this study because previous research mostly raised concerns about AI models’ performance across different racial/ethnic groups. Primarily due to the geographic location of the state assessment, the test takers were predominately identified in one of the following three race/ethnicity groups: White, Asian, or Hispanic/Latino, accounting for about 25%, 10%, and 50%, respectively, of the test-taker population. The remaining racial/ethnic groups (including Black/African American, American Indian or Alaskan Native, Native Hawaiian or other Pacific Islander, two or more races, or other) each accounted for less than 4% of the test-taker population; altogether they accounted for around 15% of the test-taker population. Due to the sample size of the smaller racial/ethnic groups, they were combined into a single group for sampling purposes. As seen in Table 2, the sample size distribution of the racial/ethnic groups was similar across prompts.

Table 2 also highlights the difference in performance across the groups. The test takers identified as Asian (denoted as Subgroup 3) had, on average, higher human mean scores than the test takers identified as White (denoted as Subgroup 1). The Hispanic/Latino test takers (denoted as Subgroup 2) received, on average, much lower human mean scores. Subgroup 4, which consisted of a mix of test takers from many racial/ethnic groups, had similar human mean scores, on average, as Subgroup 1 across prompts. This difference in performance might be due to differences in writing style, use of vocabulary, or even test-taking strategy and cultural background, among other factors. In the stratified sampling approach, which is described in the next section, the AI models were trained using samples with equal representation from all racial/ethnic groups.

Table 2

Human Mean Scores and Standard Deviations by Subgroup (Test Set)

	Subgroup 1		Subgroup 2		Subgroup 3		Subgroup 4	
Item	*N*	Mean(S.D.)	*N*	Mean(S.D.)	*N*	Mean(S.D.)	*N*	Mean(S.D.)
1	241	0.81(0.81)	477	0.74(0.75)	115	1.24(0.82)	119	0.92(0.78)
2	233	1.00(0.92)	501	0.72(0.86)	94	1.03(0.90)	113	0.98(0.91)
3	251	0.57(0.69)	477	0.33(0.58)	116	0.86(0.81)	110	0.62(0.75)
4	260	0.53(0.76)	490	0.42(0.70)	97	0.86(0.85)	101	0.60(0.79)
5	251	0.86(0.85)	502	0.60(0.76)	110	1.17(0.83)	103	0.75(0.85)
6	264	0.84(0.83)	491	0.69(0.73)	100	1.26(0.77)	128	0.87(0.85)
7	267	0.65(0.75)	488	0.56(0.66)	91	0.93(0.81)	125	0.77(0.80)
8	251	0.59(0.74)	475	0.39(0.63)	114	0.96(0.80)	132	0.71(0.80)
9	253	0.52(0.75)	466	0.26(0.57)	97	0.82(0.88)	140	0.58(0.78)
10	233	0.47(0.69)	504	0.27(0.55)	110	0.65(0.72)	124	0.52(0.73)
11	254	0.42(0.62)	521	0.33(0.58)	75	0.84(0.84)	127	0.41(0.67)
12	244	0.71(0.81)	483	0.42(0.65)	109	0.97(0.87)	131	0.74(0.85)
13	233	0.70(0.82)	496	0.39(0.66)	102	1.01(0.92)	128	0.69(0.88)

Sampling and AI Model Building

Due to the content-specific nature of the items (that is, one item may be about global warming and another item may be about playing poker game), we built and evaluated AI models on an item basis (i.e., item-specific models). For each item or prompt, we first randomly selected and put aside 40% of the responses as the test set. The percentage of responses for the test set was meant to strike a balance so that even the smallest subgroup under investigation would have at least 100 responses independent from the model-building process for model evaluation. The test-set responses were untouched until the final model evaluation. The remaining 60% of the responses were used for model building and were further split into a training sample and a validation sample. Based on our research question, we compared two sampling approaches to construct the training sample: (a) simple random sampling and (b) equal-allocation stratified random sampling by race/ethnicity (each racial/ethnic group contributed equally to model training). For each prompt, we then used the training and validation samples to fine-tune a pretrained uncased BERT$_{\text{BASE}}$ model – one of the transformer-based pretrained LLMs – to predict human scores using deep learning neural networks (NN). AdamW was used as the optimizer in fine-tuning the hyperparameters of the NN models, with a learning rate set at *1e-5*. The batch size was set at 128 and training epoch was set at 25. The script was written in Python and was run on Amazon Web Services (AWS). The statistical analyses of the model performance were conducted on the author's local machine using Python. The model performance resulting from all sampling methods was compared and evaluated using the same test set.

Specifically, for the simple random sampling (denoted as "r" in the paper), two-thirds (66.7%) of the model-building data were used for training and the rest for validation. Of note is that the situation for model validation was slightly complex under stratified sampling due to the fact that (a) the sizes of the racial/ethnic groups were quite unbalanced and (b) after selecting the same number of responses from each racial/ethnic category for model training, the distribution of both human score and race/ethnicity in the remaining validation sample became rather different from the original sample. Therefore we investigated two variations on the validation sample: one simply using what was left after stratification (denoted as s1), knowing that this validation sample drastically differed from both the training sample and the original data, and the other resampling after stratification to match the subgroup (hence also score) distribution to the total sample (denoted as s2). As a result, under the s2 condition, the validation sample would have essentially the same score and subgroup distributions as the test set (which, as a reminder, is 40% of the whole sample). Specifically, we set a total sample of 560, or 140 per subgroup, in constructing the training sample in the s1 method to ensure that there were some responses left for validation in each subgroup. In implementing the s2 method, the (equal) sample size for each racial/ethnic group in the training sample was determined by 90% of the smallest racial/ethnic group. To construct the validation samples, all the remaining responses from the smallest racial/ethnic group were used while the sample sizes for other subgroups were determined according to their proportions in the test set. Because we forced the validation sample to emulate the test set, the larger subgroups for any prompt in the s2 method could be inevitably down-sampled quite a bit, resulting in a much smaller validation set overall.

To provide a full picture of the sampling result, Table 3 lists the final sample size for the training, validation, and test sets in each prompt. A few observations are worth noting. The sample size for the r training set was nearly twice the size of the training sets under the s1 and s2 methods. The training set sample size was similar between the s1 and s2 methods, but the validation sample was drastically reduced under the s2 method, ranging from only 121 to 182 across prompts, compared to 906 to 949 across prompts for the s1 method.

Table 3

Number of Responses in Training, Validation, and Test Sets

	r		s1		s2		
Item	training	validation	training	validation	training	validation	test
1	985	486	560	911	560	140	981
2	984	485	560	909	560	132	980
3	1,011	498	560	949	592	181	1,006
4	1,002	495	560	937	584	137	998
5	1,002	494	560	936	596	167	998
6	1,007	496	560	943	564	141	1,002
7	1,010	498	560	948	516	182	1,006
8	1,003	495	560	938	552	141	1,000

Table 3

Number of Responses in Training, Validation, and Test Sets (Continued)

	r		s1		s2		
Item	training	validation	training	validation	training	validation	test
9	1,007	496	560	943	528	133	1,003
10	978	482	554	906	484	130	974
11	988	488	557	919	540	121	985
12	981	484	555	910	488	128	978
13	1,006	496	560	942	568	157	1,002

Model Evaluation Metrics

To evaluate the accuracy and fairness of the AI scoring model, we followed the best practice suggested by ETS (McCaffrey et al., 2022). Specifically, for scoring accuracy, we examined quadratically weighted kappa (Cohen, 1968), disattenuated correlation, and standardized mean score differences (SMD) between human and AI scores on the test set. Additionally, we examined how well AI could predict the human true score using the proportional reduction in mean squared error (PRMSE) metric (Haberman, 2019; Loukina et al., 2020). The PRMSE is calculated as follows: $PRMSE = 1 - \frac{E(T-M)^2}{V(T)}$, where T is the human true score and M is the AI score. In the case of human scoring, true scores involve expected human ratings given the responses observed. But the variance of human true score cannot be directly estimated. But according to classical test theory, $V(T) = V(O) - V(e)$, where O is the observed score and e is the measurement error. Assuming measurement errors of the human ratings on the same essay are uncorrelated, we can use the agreement samples (responses with two human ratings) to estimate the variance of the measurement error of each prompt: $\hat{V}(e_k) = \sum_{i=1}^{r_k}(X_{ik} - X'_{ik})^2 / 2r_k$, where k is the prompt and r is the number of raters. Disattenuated correlations are calculated as: $d.R = R_{H,M}/\sqrt{R_{H,H}}$, where the numerator is the correlation of human score H and AI score M and the denominator is the correlation of the two human scores. Similar to PRMSE, disattenuated correlation attempts to evaluate prediction accuracy after removing noise in human ratings. Worth noting is that there is a fine distinction between prediction accuracy and agreement: According to Haberman (2019), kappa or QWK is a form of agreement metric and PRMSE is a metric of prediction accuracy. In the context of this study, we evaluated AI model performance on both metrics. The SMD is calculated as $SMD = (\bar{H} - \bar{M})/\sqrt{(s_H^2 + s_M^2)/2}$, where the mean differences between human score H and AI score M is divided by the pooled standard deviation of H and M. While SMD has been commonly suggested in the literature for evaluating the bias of AI models, one issue with SMD is that it can be sensitive to the differences in scales between human and AI scores. For fairness evaluation on the subgroup level, we used the mean difference in standardized scores (MDSS) metric: $MDSS = \bar{H}' - \bar{M}'$, where H' and M' are standardized scores. The MDSS metric compares the human and AI mean scores by first removing their scales differences. MDSS is also the metric that we operationally use in practice for subgroup evaluation at the authors' organization. Additionally, for subgroup evaluation, we computed an adjusted mean score difference that was conditioned on human true score for each subgroup. The concept of this more recently developed metric is closely aligned to the concept of differential item functioning in psychometrics. That is, people with the same latent ability should have equal probability of getting a machine score, regardless of their group membership. Furthermore, this concept of predicted score M being conditionally independent of group membership G given the human true score $T(M \perp G \mid T)$ is termed "separation" fairness in the machine learning community. Hence for brevity, we denote this metric as the separation metric. The larger the separation is, the more the potential bias is for a given subgroup. The technical details of this metric can be found in Johnson, Liu, and McCaffrey (2022) and Johnson and McCaffrey (2023).

Results

Prediction Accuracy Results

Table 4 gives the means and standard deviations of the human score and AI score resulting from different sampling methods on the same test set. On the raw mean differences between human and AI scores, all differences are within a magnitude of 0.15. It is obvious that all the AI scores resulting from any sampling method have a slightly smaller standard deviation than the human scores. This minor scale shrinkage, however, does not appear to affect systematically the scoring accuracy and fairness of the AI scoring models.

Table 4

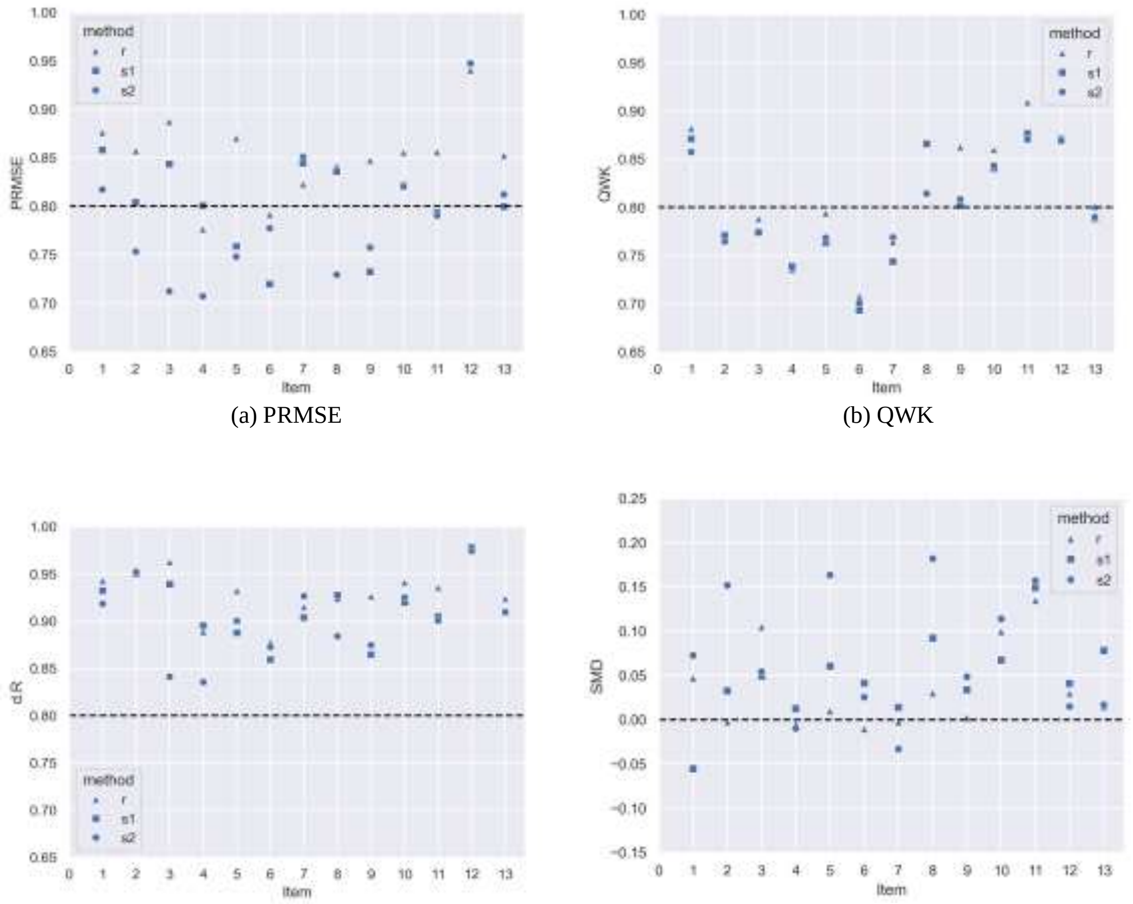
Human and AI Mean and Standard Deviations by Sampling Method (Test Set)

			Sampling method		
Item	Test set *N*	Human score	r	s1	s2
1	981	0.74(0.82)	0.78(0.82)	0.69(0.80)	0.80(0.82)
2	980	0.48(0.68)	0.48(0.59)	0.50(0.65)	0.58(0.68)
3	1,006	0.62(0.72)	0.70(0.66)	0.66(0.67)	0.66(0.47)
4	998	0.54(0.72)	0.54(0.64)	0.55(0.63)	0.53(0.47)
5	998	0.42(0.70)	0.42(0.57)	0.46(0.61)	0.53(0.64)
6	1,002	0.38(0.64)	0.37(0.57)	0.41(0.58)	0.40(0.56)
7	1,006	0.39(0.63)	0.39(0.58)	0.40(0.52)	0.37(0.55)
8	1,000	0.55(0.78)	0.57(0.75)	0.62(0.77)	0.69(0.75)
9	1,003	0.58(0.76)	0.58(0.71)	0.60(0.68)	0.61(0.67)
10	974	0.83(0.89)	0.91(0.87)	0.89(0.82)	0.93(0.85)
11	985	0.82(0.79)	0.93(0.79)	0.94(0.77)	0.94(0.75)
12	978	0.50(0.75)	0.52(0.68)	0.53(0.68)	0.51(0.69)
13	1,002	0.80(0.80)	0.81(0.70)	0.86(0.72)	0.81(0.71)

The standardized mean score differences (SMD) between human and AI scores, shown in Figure 1(d), suggest that all SMDs resulting from r and s1 methods are within the magnitude of 0.15 – a threshold value suggested in the literature (McCaffrey et al., 2022; Williamson, Xi, & Breyer, 2012). However, the s2 method showed slightly larger SMDs on four items (i.e., Items 2 and 11 on the borderline of 0.15 and Items 5 and 8 in between 0.15 and 0.20), where the AI scores have overall higher means than human scores. This result indicates that the smaller validation sample in the implementation of the s2 method, even though "matched" to the subgroup distributions in the test set, seemed to have an impact on model performance, in particular the AI score means. Even though the validation samples for the s1 method are "not matched" to the test set, there are a much greater number of responses representing each racial/ethnic group. In other words, prioritizing a larger validation sample may be more crucial than to achieve a distributional "match" by sacrificing sample size to AI model performance.

Figure 1

Results of Prediction Accuracy



(a) PRMSE

(b) QWK

(c) Disattenuated Correlation

(d) SMD

Included in Figure 1 are the results for other evaluations on prediction accuracy, that is, PRMSE, QWK, and disattenuated correlation (denoted as "d.R"), between human and AI scores resulting from different sampling methods. All PRMSE statistics were greater than 0.7, which is considered a minimum performance threshold for AI scoring models (McCaffrey et al., 2022). Among the lower PRMSEs, such as those in between 0.7 and 0.8, most resulted from the s2 method and some resulted from s1. Previous research suggested $QWK >= 0.7$ when evaluating automated scoring models (Williamson et al., 2012). In this analysis, all QWK values were greater than 0.7 with the exception of one instance: the $QWK = 0.694$ on Item 6 resulting from the s2 method. The QWKs were also on the borderline of 0.7 for the other two sampling methods. One speculation for why this happened is that the standard deviation of the human scores on this item is small (S.D. = 0.64), which could have an impact on the AI model building and evaluation. The d.Rs were all above 0.83, with $d.R^2$ all larger than the threshold of 0.7 suggested in McCaffrey et al. (2022).

Overall, the AI scoring models based on all three sampling methods demonstrated reasonably good performance. While the intention to match the validation sample to the test set in the s2 method was to enhance the AI model performance, empirical evidence did not support that decision. AI models based on the simple random sampling (s1) showed the best performance in many cases. Interestingly, in s1, it worked well to use a much smaller (i.e., about half of the size) model training sample with equal subgroup representations but with a much larger validation sample that was different from the test set in terms of score and subgroup distributions. Even though the s2 method did not outperform the r method, the model performance, in general, was in fact quite acceptable.

Fairness Results

Figures 2 shows the results of the separation metric, which evaluated the human-AI mean score differences conditional on true score for each racial/ethnic group. All of the values in Figure 5 were within 0.2, with the majority of the values within 0.1. This result means that, for a given subgroup conditional on the true score, the AI mean scores only differed from human mean scores by less than one tenth of a score point on the 2-point scale. These differences could be considered negligible. The only notably larger separation between human and AI scores was for Item 7, which all the AI models underscored for Subgroup 3 (the Asian test-taker group) on average. In this case, the s1 method notably outperformed the r and s2 methods by better predicting the means of Subgroup 3.

Figure 2

Separation Results by Subgroup

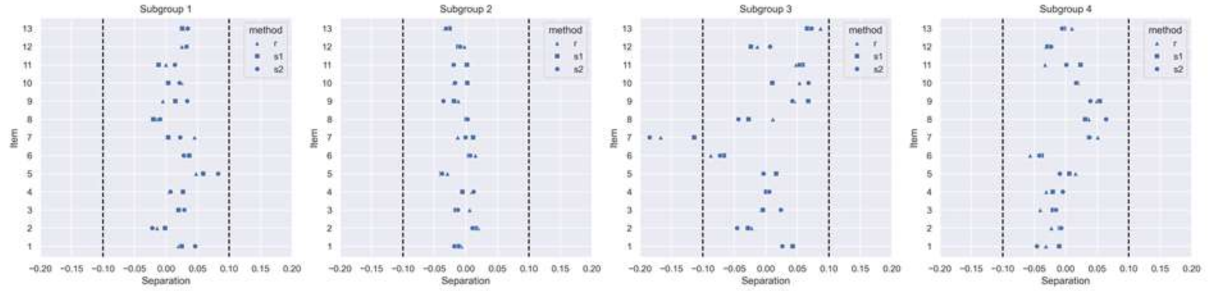
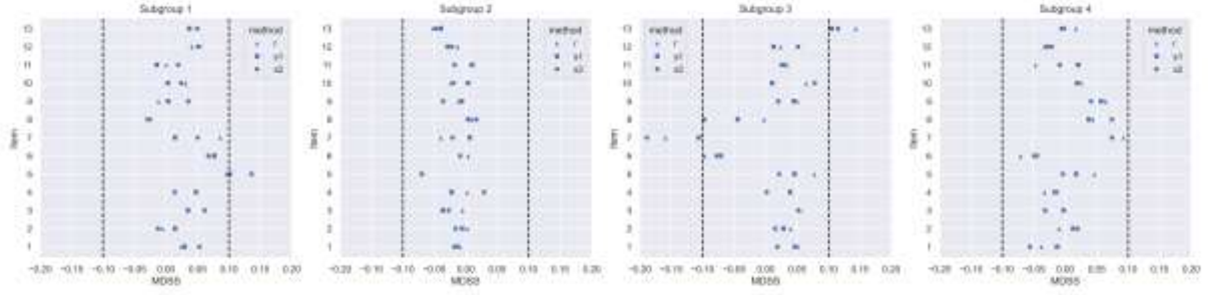


Figure 3

Mean Differences in Standardized Scores (MDSS) by Subgroup



The results on the MDSS for the racial/ethnic groups are shown in Figure 3. The findings are similar to the separation metric. All the mean differences were within the magnitude of 0.2. While there is no established or recommended evaluation threshold for this metric, we applied 0.20 as a common in-house threshold for test operations. As with the separation metric results, the largest MDSS values were associated with Subgroup 3 on Item 7. It is interesting to observe that the s1 method tended to outperform simple random sampling, especially on cases that had larger MDSS values (e.g., Subgroup 1 on Item 5, Subgroup 3 on Item 7, and Subgroup 3 on Item 13).

Discussion

In this exploratory study, we empirically evaluated the impact of sampling on AI scoring model performance. Simple random sampling (r) and equal-allocation stratification random sampling were compared in constructing the AI model training samples. Two variations of the sampling strategy (s1 and s2) in constructing the validation sample in the AI modeling building process were further examined under the equal-allocation stratification random sampling. Fine-tuned LLMs were trained and evaluated, and all the AI models were prompt-specific for the 13 items included in this analysis. We summarized the characteristics of the samples under each sampling method in Table 5.

Table 5

Sample Characteristics

Method	Training	Validation	Test (same across methods)
r	*N*≈1000; representative of whole population; dominated by large subgroups.	*N*≈500; representative of whole population; dominated by large subgroups.	*N*≈1000; representative of whole population; dominated by large subgroups.

s1	*N*≈550;
	smaller in size;
	equal contribution from each subgroup
		N≈900;
		very large in size;
		very different from whole population
		even more dominated by large subgroups.
			N≈1000;
			representative of whole population;
			dominated by large subgroups.

s2	*N*≈550;
	smaller in size;
	equal contribution from each subgroup
		N≈140;
		very small in size;
		representative of whole population;
		dominated by large subgroups.
			N≈1000;
			representative of whole population;
			dominated by large subgroups.

In response to the research questions, the models were evaluated from two perspectives: overall prediction accuracy and fairness. For RQ1, we found that, in general, the AI scoring models predicted human scores reasonably well regardless of the sampling method. Even when the training sample size was relatively small as in s1 and in s2 compared to r, or when the validation sample was extremely small (as in s2) or relatively large (as in s1), the model performance was marginally affected and was comparable across methods. Even though the AI models appeared to perform slightly worse using the s2 method, the observation was only on the SMD index for three out of the 13 prompts while the evaluations did not reveal other obvious issues for the s2 method. In addressing RQ2, we found that using model training samples with equal representation from subgroups of test takers (s1 and s2) did not systematically improve the fairness of the AI scoring models. In almost all cases, models based on simple random sampling were fair across the different racial/ethnic groups. In a couple of rare cases where models resulting from the r method did not work as well (Subgroup 3 on Items 7 and 13), stratification appeared to have improved fairness. From a big picture point of view, however, equal-allocation stratification did not improve, or worsen, fairness. One could argue, though, that stratification is a critical and early treatment to mitigate bias in the data preprocessing step during model development process (Chu et al., 2024) in the sense that all subgroups of interest contributed equally in the model training process. The model is not dominated by inherent biases associated with any specific subgroup of interest. The results may also arguably favor stratification given that both the training and validation samples can be relatively small and the validation sample does necessarily need to resemble the test-taker population (s1). These potential advantages on sample requirements seem especially useful when only small data set is available for AI model building.

There is relatively little prior research that specifically focused on sample size requirements for AI scoring with few exceptions such as Haberman and Sinharay (2008), Zhang (2013), and Heilman and Madnani (2015). To our best knowledge, most of the former work was conducted with the earlier generation of AI scoring practice (pre-LLM era) when well-defined features were used in less complex, but more explainable AI scoring models such as logistic regression or multiple linear regression. The authors in investigated the training sample sizes in AI scoring of short-response items using support vector regression models where the predictors included various word n-grams and a proxy of response length. Their findings (Figure 1 in the refereed article) showed that from small training sample size of 100 to larger training sample sizes of 200, 400, 800, 1600, and up to 3200, the scoring model performance as evaluated by QWK steadily and considerably improved. This study found different results related to sample sizes from the prior work, mostly likely due to the use of fine-tuned LLMs. In Heilman and Madnani (2015), about half of the items achieved a human-AI QWK of 0.7 or greater and required at least 800 model training responses. Even when the training sample sizes were as large as 1600 or 3200, it appeared difficult to achieve a QWK of 0.8 and above. In the current study, LLM-based AI scoring models achieved QWK of 0.75 or above on 11 of 13 prompts, regardless of the sampling method. This result aligns well with the literature in the AI community in that complex AI models such as NN tended to achieve greater accuracy in prediction tasks than simpler models such as SVM or decision trees.

Even though the most of the LLM-based AI scoring models demonstrated high prediction accuracy and an acceptable degree of fairness, there still seems to be room for improvement. The top performing models reported in Whitmer et al. (2021) achieved average human-AI QWK ranging from 0.860 to 0.888 across NAEP Reading items, about 0.05 points higher on average than the QWKs reported in this study. Most of the top performing models in the NAEP study were either ensembles of multiple models or leveraged information in the prompt and source text. The total samples in the NAEP study ranged from 19,934 to 28,307 across items (Whitmer et al., 2021), which are much larger than the samples per item available in this study. So it is highly likely that we can further improve the current AI model performance on these items when we collect more responses in test operations. By improving the overall model prediction accuracy, fairness will likely be improved accordingly. In this analysis, we did not customize the model fine-tuning process for each item; instead, we applied the same setting for all items. Customizing the fine-tuning process will most likely improve model performance on the item level as well.

Overall, this study offers some empirical evidence on the choice of sampling methods in building LLM-based AI scoring models for short-response assessment items. For the items investigated in this study, a training sample size of 1000 from simple random sampling was generally sufficient. We found the models based on stratified samples performed comparably to models based on simple random samples. However, it is worth noting the stratified training samples were only half of the size. For testing programs that intend to prioritize fairness in the AI model training process, stratified sampling can be seriously considered.

This study has several limitations. One is that the $\text{BERT}_{\text{BASE}}$ LLMs were fine-tuned with minimum effort. The same settings were using across prompts. It is possible that differences in prediction accuracy and fairness may emerge along with more optimal fine-tuning such as adjusting the learning rate in each item. The results may not generalize to LLMs beyond $\text{BERT}_{\text{BASE}}$, making a comparative analysis worthwhile in the future. We also did not consider intersections of demographic variables (e.g., gender by race/ethnicity, language skill by gender), which future research is encouraged to explore. Additionally, as a follow-up of the current analysis, stratified sampling by race/ethnicity *and* score levels can provide more useful results on improving fairness in the model-training process. For example, selecting the equal number of responses from each racial/ethnic group at each score level essentially makes the (human) scores orthogonal to race/ethnicity and, as a result, any detected biases in the machine scores would be due to other reasons than one's race or ethnicity alone. This is a natural next step once more responses are collected. Due to the limited responses in some racial/ethnic groups, equal-allocation stratified random sampling on student-written responses could only utilize a relatively small sample. Future research can consider augmenting the data set with synthetic data (e.g., using GPT, for underrepresented subgroups). Alternatively, future research may also apply techniques that algorithmically mitigate bias in the training data, such as sample reweighting and adopting fairness-aware machine learning models (Ferrara, Sellitto, Ferrucci, et al., 2024; Haberman, 1984). Finally explaining detected biases is challenging with complex AI scoring models. Johnson and McCaffrey (2023) proposed one method to weight AI features differently to reduce subgroup biases in simpler models; future research is encouraged to generalize the method in Johnson and McCaffrey (2023) to LLM-based AI scoring.

Declarations

Gen-AI Use : The authors of this article declare (Declaration Form #: 2611241949) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Author Contribution: M. Zhang and M. Johnson conceptulized the study and wrote the manuscript. M. Zhang and C. Ruan conducted the modeling and statistical analysis.

Conflict of Interest: None

Ethical Approval: Not applicable.

References

Ali, S., Abuhmed, T., El-Sappagh, S., et al. (2023). Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, *99*(C). Retrieved from https://doi.org/10.1016/j.inffus.2023.101805

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Ayoub, N. F., Balakrishnan, K., Ayoub, M. S., Barrett, T. F., David, A. P., & Gray, S. T. (2024). Inherent bias in large language models: A random sampling analysis. *Mayo Clinic Proceedings: Digital Health*, *2*, 186–191. Retrieved from https://doi.org/10.1016/j.mcpdig.2024.03.003

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). *Measuring implicit bias in explicitly unbiased large language models*. arXiv. Retrieved from https://arxiv.org/pdf/2402.04105

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology in testing: Measurement issues* (pp. 142–173). Taylor & Francis.

Caton, S., & Haas, C. (2024). Fairness in machine learning: A survey. *ACM Computing Surveys*, *56*(7), Article 166. Retrieved from https://doi.org/10.1145/3616865

Chamieh, I., Zesch, T., & Giebermann, K. (2024). LLMs in short answer scoring: Limitations and promise of zero-shot and few-shot approaches. In E. Kochmar et al. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 309–315). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.bea-1.25.pdf

Chhabra, A., Singla, A., & Mohapatra, P. (2022). Fair clustering using antidote data. In J. Schrouff, A. Dieng, M. Rateike, K. Kwegyir-Aggrey, & G. Farnadi (Eds.), *Proceedings of the algorithmic fairness through the lens of causality and robustness* (Vol. 171, pp. 19–39). PMLR. Retrieved from https://proceedings.mlr.press/v171/chhabra22a.html

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: A taxonomic survey. *ACM SIGKDD Explorations Newsletter*, *26*(1), 34–48. Retrieved from https://doi.org/10.1145/3682112.3682117

Cohen, J. (1968). Weighted kappa: Normal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213-220. http://dx.doi.org/10.1037/h0026256

Ferrara, C., Sellitto, G., Ferrucci, F., et al. (2024). Fairness-aware machine learning engineering: How far are we? *Empirical Software Engineering*, *29*(9). Retrieved from https://doi.org/10.1007/s10664-023-10402-y

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *Annals of Statistics*, *12*(3), 971–988. Retrieved from https://www.jstor.org/stable/2240973

Haberman, S. J. (2019). *Measures of agreement versus measures of prediction accuracy* (Research Report No. RR-19-20). Retrieved from https://doi.org/10.1002/ets2.12258

Haberman, S. J., & Sinharay, S. (2008). *Sample-size requirements for automated essay scoring* (Research Report No. RR-08-32). Retrieved from https://doi.org/10.1002/j.2333-8504.2008.tb02118.x

Heilman, M., & Madnani, N. (2015). The impact of training data on automated short answer scoring performance. In J. Tetreault, J. Burstein, & C. Leacock (Eds.), *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 81–85). Retrieved from https://doi.org/10.3115/v1/W15-0610

Johnson, M. S., Liu, X., & McCaffrey, D. F. (2022). Psychometric methods to evaluate measurement and algorithmic bias in automated scoring. *Journal of Educational Measurement*, *59*, 338–361. Retrieved from https://doi.org/10.1111/jedm.12335

Johnson, M. S., & McCaffrey, D. F. (2023). Evaluating fairness of automated scoring in educational measurement. In V. Yaneva & M. von Davier (Eds.), *Advancing natural language processing in educational assessment.* Routledge.

Johnson, M. S., & Zhang, M. (2024). *Examining the responsible use of zero-shot AI approaches to scoring essays.* Manuscript submitted for publication.

Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, *4*(47). Retrieved from https://doi.org/10.1007/s44163-024-00147-y

Kumar, A., Dikshit, S., & de Albuquerque, V. (2021). Explainable artificial intelligence for sarcasm detection in dialogues. *Wireless Communications and Mobile Computing*, *1*, 1–13. Retrieved from https://doi.org/10.1155/2021/2939334

Lee, G.-G., Latif, E., Wu, X., Liu, N., & Zhai, X. (2024). Applying large language models and chain-of-thought for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100213. https://doi.org/10.1016/j.caeai.2024.100213

Lohr, S. L. (2021). *Sampling: Design and analysis* (3rd ed.). Chapman and Hall/CRC. Retrieved from https://doi.org/10.1201/9780429298899

Loukina, A., Madnani, N., Cahill, A., Yao, L., Johnson, M. S., Riordan, B., & McCaffrey, D. F. (2020). Using PRMSEs to evaluate automated scoring systems in the presence of label noise. In J. Burstein, E. Kochmar, C. Leacock, N. Madnani, H. Y. Ildikó Pilán, & T. Zesch (Eds.), *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 18–29). Retrieved from https://doi.org/10.18653/v1/2020.bea-1.2

Lubis, F. F. M., Putri, A. W. D., et al. (2021). Automated short-answer grading using semantic similarity based on word embedding. *International Journal of Technology*, *12*(3), 571–581. Retrieved from https://doi.org/10.14716/ijtech.v12i3.4651

Ma, W., Scheible, H., Wang, B., & Veeramachaneni, G. (2023). Deciphering stereotypes in pre-trained language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 11328–11345). Association for Computational Linguistics. Retrieved from https://doi.org/10.18653/v1/2023.emnlp-main.697

Manvi, R., Khanna, S., Burke, M., Lobell, D., & Ermon, S. (2024). *Large language models are geographically biased*. arXiv. Retrieved from https://arxiv.org/abs/2402.02680

McCaffrey, D. F., Casabianca, J., Ricker-Pedley, K. L., Lawless, R., & Wendler, C. (2022). *Best practices for constructed-response scoring* (Research Report No. RR-22-17). Retrieved from https://doi.org/10.1002/ets2.12358

Navigli, R., Conia, S., & Ross, B. (2023). Biases in large language models: Origins, inventory, and discussion. *Journal of Data and Information Quality*, *15*(2), 1–21. Retrieved from https://doi.org/10.1145/3597307

Oka, R., Kusumi, T., & Utsumi, A. (2024). Performance evaluation of automated scoring for the descriptive similarity response task. *Nature Scientific Reports*, *14*, Article 6228. Retrieved from https://doi.org/10.1038/s41598-024-56743-6

Whitmer, J., Deng, E. Y., Blankenship, C., Beiting-Parrish, M., Zhang, T., & Bailey, P. (2021). *Results of NAEP reading item automated scoring data challenge (fall 2021)*. EdArXiv. Retrieved from https://osf.io/preprints/edarxiv/2hevq

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. Retrieved from https://doi.org/10.1111/j.1745-3992.2011.00223.x

Zhang, M. (2013). *The impact of sampling approach on population invariance in automated scoring of essays* (Research Report No. RR-13-18). https://doi.org/10.1002/j.2333-8504.2013.tb02325.x

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2024; 15(Special issue); 361-377

Generative AI in K12: Analytics from Early Adoption

Brad BOLENDER* Sara VİSPOEL** Geoff CONVERSE ***
Nick KOPROWİCZ**** Dan SONG***** Sarah OSARO******

Abstract

The integration of generative AI in K12 education and assessment development holds the potential to revolutionize instructional practices, assessment development, and content alignment. This article presents analytical insights and findings from early adoption studies utilizing AI-powered tools developed by Finetune—Generate and Catalog. Generate enhances the efficiency of assessment item development through customized natural language generation, producing high-quality, psychometrically valid items. Catalog intelligently tags and aligns educational content to various standards and frameworks, improving precision and reducing subjectivity. Through three comprehensive case studies, we explore the practical applications, benefits, and lessons learned from employing these AI systems in real-world educational settings. The purpose of this series of studies was to investigate the ways generative AI is currently being used in practical applications in test development to improve processes and products. The studies demonstrate significant reductions in time and costs, enhanced accuracy, and consistency in content alignment, and improved quality of educational and assessment materials. The findings underscore the substantial benefits and critical importance of customized AI systems, rigorous training for both AI models and users, and adopting appropriate evaluation metrics. With the use of off-the-shelf generative AI models expanding rapidly, it is vital that the effectiveness of AI systems that are highly customized through collaborations with measurement experts be presented, in order to maximize benefits and uphold the fundamental principles and best practices of test development.

Keywords: Generative AI, Assessment Development, Content Alignment, Educational Measurement

Introduction

Over the past decade natural language processing (NLP), machine learning, and artificial intelligence (AI) have grown steadily as tools to increase efficiency across industries, including education and assessment. The uses for these tools have been wide-ranging, from developing tests (Gierl & Haladyna, 2013) through automated scoring (Yan & Rupp, 2020) of short-answer constructed-responses (Burrows et al., 2015) and essays. These technologies have spread through the industry at a steady pace, as applications have been coded and refined (Attali & Burstein, 2006), validity arguments for their use have been developed and defended (Bennett & Zhang, 2015), and data analyses have been executed and presented that support their judicious implementation into testing organization’s pipelines. However, in the past few years Generative AI has exploded onto the scene, and it has the power to dramatically alter educational instruction, test development, content analysis, and curriculum alignment methods.

Generative AI represents a transformative field in artificial intelligence focused on the autonomous generation of data. Central to this innovation are Large Language Models (LLMs), inherently complex neural networks optimized to understand, generate, and manipulate human language. These models,

* Principal Measurement Scientist, Finetune Learning, Iowa-City, US, bbolender@finetunelearning.com, ORCID ID: 0009-0000-1521-6136
** Chief Assessment and Learning Officer, Finetune Learning, Iowa-City, US, sara@finetunelearning.com, ORCID ID: 0009-0002-8159-8210
*** AI Scientist, Finetune Learning, Iowa-City, US, geoff@finetunelearning.com, ORCID ID: 0000-0001-8764-9950
**** Senior AI and Machine Learning Applied Scientist, Finetune Learning, Iowa-City, US, ORCID ID: 0009-0005-1548-6116
***** Graduate Research Assistant, University of Iowa, Iowa-City, US, dan-song@uiowa.edu, ORCID ID: 0000-0002-7466-6150
****** Research Assistant at University of Iowa, Iowa-City, US, ORCID ID: 0009-0008-9264-9072

To cite this article:
Bolender et al. (2024). Generative AI in K12: Analytics from Early Adoption. *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special Issue), 361-377. https://doi.org/10.21031/epod.1539710

Received: 28.08.2024
Accepted: 25.11.2024

often based on transformer architectures, such as GPT (Generative Pre-trained Transformer) and its derivatives, leverage vast amounts of text data to learn linguistic patterns including syntax, semantics, and context. Despite their remarkable achievements, LLMs are not without challenges, such as considerable computational requirements and the propensity for generating contextually inappropriate or biased content, reflecting biases present in training data. Ongoing research addresses these issues through model distillation, ethical frameworks, and improved dataset curation, improving generative AI's alignment with human values, fairness, and inclusiveness.

With the rapid development of use-cases for Generative AI, concerns have also been raised about potential implications for education and assessment, including protecting the validity of assessments and avoiding the introduction of fairness and bias issues. As a result, several groups composed of researchers and testing organizations have convened and published guidelines for responsible use of the technology (Bolender et al., 2023; Hao et al., 2024; Ho, 2024). The focus of these guidelines has been to protect the validity and reliability of educational assessments, to ensure fair testing practices for test takers from all demographic backgrounds, and to specify methods for protecting the privacy and security of all test data including individually identifiable data from test takers. The guidelines also provide recommendations for ensuring transparency and accountability surrounding the use of Generative AI in the test development process, so stakeholders will be fully informed of the ways in which AI was used to aid in development of the test instruments, but also what measures were taken to protect validity.

Finetune has developed two AI-supported systems to assist with tasks related to K12 education and assessment, called Generate and Catalog. Due to the novelty of generative AI, not many studies exist of incorporation into real-world processes, especially those that focus on assessment. This paper will serve as an additional contribution to the Finetune research agenda (Khan et al., 2021a; Khan et al., 2021b). Since Finetune AI scientists and psychometricians were given early access to generative AI models as far back as 2018, multiple years of real-world data from use, as well as lessons learned on what generative AI does well, what it does not, and what is required for efficacy, have been integrated into these test development tools.

Finetune’s applications and research contribution differs from most AI research using LLMs, due to a direct emphasis on customizing these AI tools specifically for assessment purposes. Incorporating best practices for AI and assessment and custom-building to customer requirements is significantly different than simply submitting a query to an LLM API. In this paper, we will share data from real users across multiple content areas and contexts to share critical information about how SMEs are using AI-assisted technologies and the degree to which best practices are being followed when using generative AI.

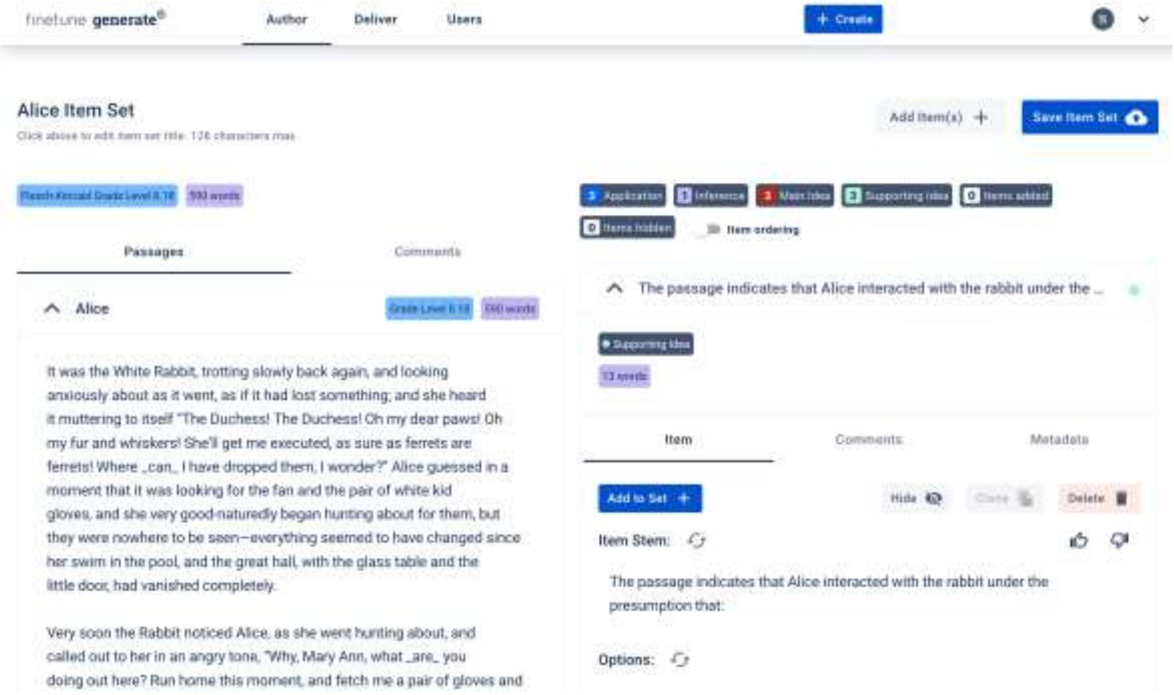
In this paper data will also be shared on using customized AI features in Finetune Catalog to automatically tag and align educational content such as items and learning materials, to standards, learning objectives, and cognitive complexity levels across content areas. This process involves harnessing both generative AI as well as more conventional machine learning techniques. Additionally, a natural language rationale can be generated explaining why an item is tagged with a particular standard. Outcome data will also be shared on how using this AI application can decrease the amount of subjectivity in tagging.

Generate

Finetune Generate (Khan et al., 2021b) is an AI-assisted system designed to enhance the efficiency and scalability of assessment content generation for educational purposes. The system leverages state-of-the-art natural language generation (NLG) methodologies in conjunction with the domain-specific expertise of assessment developers to facilitate the creation of a large volume of highly customized and psychometrically valid assessment items. Central to Finetune Generate is the Transformer architecture, which, through extensive pretraining on diverse text corpora, is adept at producing sophisticated, context-sensitive text that serves as a foundation for item generation.

Figure 1

The Finetune Generate AI-powered test development system.



Alice Item Set

Click above to edit item set title. 128 characters max.

Alice

It was the White Rabbit, trotting slowly back again, and looking anxiously about as it went, as if it had lost something; and she heard it muttering to itself "The Duchess! The Duchess! Oh my dear paws! Oh my fur and whiskers! She'll get me executed, as sure as ferrets are ferrets! Where _can_ I have dropped them, I wonder?" Alice guessed in a moment that it was looking for the fan and the pair of white kid gloves, and she very good-naturedly began hunting about for them, but they were nowhere to be seen—everything seemed to have changed since her swim in the pool, and the great hall, with the glass table and the little door, had vanished completely.

Very soon the Rabbit noticed Alice, as she went hunting about, and called out to her in an angry tone, "Why, Mary Ann, what _are_ you doing out here? Run home this moment, and fetch me a pair of gloves and

The passage indicates that Alice interacted with the rabbit under the ...

The passage indicates that Alice interacted with the rabbit under the presumption that:

When building a Generate model, items are intentionally tailored to meet requirements determined uniquely by test developer needs. Various sources of content are acquired from the user, including test blueprints, learning objectives, and cognitive complexity frameworks to build selectable sub-models that target constructs of interest in accordance with test specifications. Additionally, implicit reinforcement is provided relative to the user's style of item writing, as well as influence how an item aligns with a construct, by utilizing exemplar items provided by the user. With LLMs, there is often a trade-off between creativity—or randomness—and factual correctness—or determinism. To strike a balance, multiple sources of randomization are introduced to give a sense of variety in items, while still rooting the core content in the user-provided data. Additional features are integrated that connect AI-generated content to factual source material. This can be done in either a pre-generation manner, through Generate using a textbook passage as inspiration, or at post-generation where, given some generated content, a search is executed for a relevant textbook passage to serve as reference.

The Generate user-experience is customizable to different use-cases. At a baseline, selectable sub-models are provided that align to relevant constructs that drive the AI-generated content. Users may also input key words / key phrases to guide the AI, input a custom passage to use as a fact-base or inspiration, or input a reading comprehension stimulus passage to create item sets. Aside from content generation, AI solutions have been built into other post-hoc features, such as identifying a correct answer to an MCQ item, finding a citation for an item, and creating a rationale for correctness of the key option(s) and rationales for falsifiability of distractors.

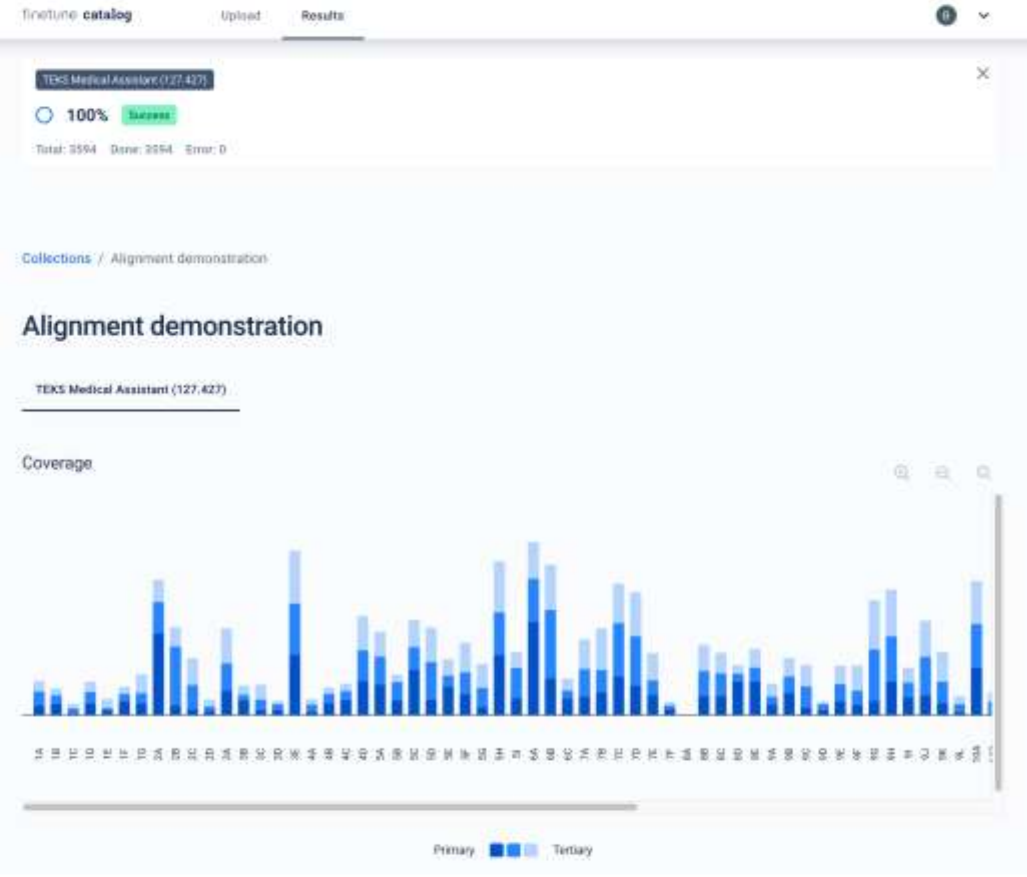
This approach is unique in how AI scientists and measurement scientists work together to integrate the information in specifications and guidelines to develop a customized generative AI model. The first process of AI-enhanced item development involves partners providing details about their assessments including test purposes, test specifications, descriptions of constructs, test blueprints, item types, cognitive complexity requirements, references they want to include, and item writing guidelines. The resulting model is deployed in the Generate application so high-quality item drafts that meet requirements are produced.

A noteworthy unique feature about this generative AI application is that the capability of interacting with the customized AI model persists throughout the entire item development process. SMEs develop the items within the application both by editing stimuli, stems, and answer options directly, and also by regenerating portions of the items with additional requests to the customized AI model. If users have reference materials that they want to use, we upload those materials into the application so SMEs can employ features like using the AI-assisted references to find citations that provide evidence for a key. In addition, content-, bias-, and committee-review processes can be completed within the application. Reviewers are able to share comments on items and can access the AI for assistance in generating possible fixes as they make subsequent revisions until the content is considered to be in its final state. At that point, the full set or a subset of items can be exported in multiple formats including plain-text, csv, and QTI-compliant XML.

Catalog

The second AI system involves applying generative AI to the task of associating learning materials to frameworks. In K-12 especially, in order for any learning material to be used flexibly and adaptively, associated metadata must be accurate, including tagging information to frameworks describing learning objectives, competencies, and cognitive complexity levels. For this task, a different application was developed: Finetune Catalog (Khan et al., 2021a). This system has been used to complete projects that entail tagging hundreds of items to larger projects tagging more than 50,000 items to various frameworks. Additionally, the Catalog engine has been used to provide AI-authored rationale statements for all tags assigned.

Figure 2
The Finetune Catalog AI tagging and alignment system.


finetune catalog
Upload
Results
TEKS Medical Assistant (127.427)
100%
Success
Total: 3594 Done: 3594 Error: 0
Collections / Alignment demonstration
Alignment demonstration
TEKS Medical Assistant (127.427)
Coverage
Primary
Tertiary

Catalog employs a comprehensive AI-driven methodology to intelligently match educational content with relevant tags, serving as an expert across diverse educational domains. The process begins with the conversion of different types of educational materials into a format that is suitable for advanced analysis. An innovative framework is utilized to identify deep semantic relationships between content and tags, ensuring that diverse educational content and standards are aligned into a cohesive semantic space. This approach addresses a critical industry challenge and enhances the precision of our tagging process.

Emulating the expertise of subject matter experts, Catalog deploys a multi-level analytical approach tailored to interpret the educational intent behind each piece of content. The system navigates through multiple stages to determine the most relevant tags, incorporating mechanisms to validate its decisions at each step. Additionally, the process is customized based on user input and iterative refinement, allowing the AI system to adapt and align closely with the user's needs. This rigorous yet flexible methodology ensures that Catalog delivers highly accurate and contextually appropriate tagging.

Catalog uses a pipeline of varied techniques including embeddings with similarity measures, LLM prompting techniques such as few-shot prompting, chain-of-thought (CoT) prompting, self-reflection, and multi-turn interactions, along with hierarchical search to most effectively recognize correct associations between content.

The purpose of this article is to share three case studies illuminating how real users are currently interacting with AI-powered systems designed to streamline education and assessment processes including assessment item development, tagging of assessment items, and gap analysis to determine how well curricular materials cover learning objectives measured on summative assessments.

Case Study 1: Generate To Support Assessment Item Development

Research Questions

In what manner, and to what degree, do SMEs edit AI-generated item drafts before moving them to the next phase of item review?

A customized instance of Generate was built to assist with the development of a high school English Language Arts test. The Generate instance was a typical item set model that enables SMEs to paste a reading stimulus passage into a text box to use as the basis for a set of items testing various reading comprehension constructs, such as detail retrieval, inference making, overall text understanding, and understanding of language features. Additionally, the model was designed to support paired-text stimuli, so multiple texts could be entered, and synthesis items could be developed that require test takers to draw conclusions based on information in both texts.

The item writer guidelines document, the test blueprint, and example items were used by the Finetune AI team to develop a custom model that would generate many item drafts resembling the user’s existing content, but that would not be based on any templates or copy any items already on their exams. A secure, unique instance of the Generate application was provided to the users – a team of 7 SMEs who would be using the model in their development cycle to help write roughly 60 items for an upcoming set of test forms. Training was provided to the user team to show them how to import stimulus passage, generate item drafts, revise and edit the items, and save them in Generate’s interface to folders identifying the item sets as ready for review in the next phase of the project. Users were instructed that Generate is intended to be a human-in-the-loop AI-supported system for item development, so they as SMEs were expected to treat the outputs as item drafts, and their expertise was necessary to refine the items into their final forms.

Methods

After Generate training was done SMEs engaged in a typical development process, except rather than starting from a blank slate and having to come up with ideas for items, they used the customized Generate model to create item drafts. SMEs would then browse the item drafts and make decisions about which item drafts to work further on and which items to reject. SMEs would then consider the items in relation to the reading stimulus and make any minor edits necessary to the stem to ensure best measurement of the construct of interest. If needed, the key was edited to make sure there was one clear correct answer, and distractors were edited as needed to ensure they were clearly incorrect. SMEs interacted with the custom AI model throughout this procedure, by editing stimulus materials and using AI to regenerate stems, and by further editing the stem and regenerating answer options. Additionally, the custom model underwent continuous improvement through SMEs using a “thumbs-up” button to identify the best generated items, which were then integrated into examples for the model to consider in subsequent item generation, and a “thumbs-down” button to identify poor items to discourage the model from generating more like them. All of these efforts contribute to the custom generative model getting better and better as it is used by the SME in the process. When SMEs were satisfied with the items, then they saved them to a folder for further review.

To investigate the research question for this data set, the database underlying the application was accessed, as each item generated is assigned a unique ID that stays with the item through all versions as it is edited or revised during the test development process. The unique IDs for all 58 items that were saved for further review were queried, and the original AI-generated versions of the items were exported from the database so they could be compared to the versions that were saved in the review folder.

To compare the original AI-generated item drafts to the versions SMEs moved into the review folder, the Levenshtein edit distance (Levenshtein, 1966) was calculated between the two versions. Levenshtein edit distance (from here on simply referred to as edit distance) is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change the original version of an item into the saved version. This is an established method to make quantitative comparisons between language strings. Although character edits may not be especially intuitive in terms of imagining the extent to which an item has been changed, it may help to consider that some testing organizations use a concept called “standard word count” to refer to the length of content, which is total number of characters (including punctuation and spaces) divided by 6. So an edit distance of 48 could be imagined to correspond roughly to 8 words being changed.

After edit distances were recorded for both stems and options, results were summarized by grouping the items by the SME who worked with them, and the mean distances were calculated. This gives an idea of the typical amount individual SMEs edited the AI-generated items, both stems and options, prior to saving them for review, and it also gives an idea of the variation in editing behavior between SMEs on this project.

Additionally, it was determined how many out of the 58 saved items had 0 stem edits and 0 answer option edits, in order to understand how often portions of the AI-generated item drafts were satisfactory in their original states to move forward to the review phase of development.

Case Study 1: Results And Discussion

Table 1 shows mean edit distances between AI-generated item stems and review versions of item stems by SMEs who worked on them. Table 2 shows mean edit distances between AI-generated item options and review versions of item options by SMEs who worked on them. Table 3 shows the frequency of edit distance of 0 between AI-generated item stems and review versions of item stems overall, indicating items where the stems were satisfactory to be moved forward to the review process without additional editing. Table 4 shows the frequency of edit distance of 0 between AI-generated answer options and review versions of answer options overall, indicating items where the answer options were satisfactory to be moved forward to the review process without additional editing.

Table 1
Mean Edit Distance Between AI-Generated Stems and SME-Revised Stems

Subject Matter Expert	Mean Edit Distance Stem
SME_1	29.7
SME_2	32.0
SME_3	51.4
SME_4	59.5
SME_5	65.0
SME_6	6.1
SME_7	41.2

Table 2
Mean Edit Distance Between AI-Generated Answer Options and SME-Revised Answer Options

Subject Matter Expert	Mean Edit Distance Options
SME_1	54.2
SME_2	66.0
SME_3	90.3
SME_4	129.0
SME_5	75.0
SME_6	31.4
SME_7	167.0

Table 3
Edit Distance Frequency – Stem

Edit Distance	Freq.
0	5
>0	53

Table 4
Edit Distance Frequency – Options

Edit Distance	Freq.
0	9
>0	49

These results show that SMEs were active in working with the AI-generated item drafts. Although there was some variation between SMEs, most of them moved items into the review phase with at least 29 edit distance or more between drafts and review versions. The exception is SME 6 whose review items had a mean edit distance of 6.1 from the AI-generated items, which was quite a bit lower than the other SMEs. However, referring to Table 2 we see that the answer options for SME 6's review items had a mean edit distance of 31.4 from the AI-generated answer options. So it is possible that SME 6 was satisfied with the AI-generated stems, and they spent relatively more time working on refining the answer options.

On the whole, SMEs worked extensively with the answer options of AI-generated items, producing mean edit distances between 31.4 and 167. A future line of research could involve investigating the amount and type of editing that was done to keys, in order to enhance construct measurement or to make correctness certain, versus editing done to distractors, which may have been done to make items easier or harder or to introduce common errors and misconceptions. Results from that research could be used to inform further advancement in AI model and system development in terms of continuing to integrate best measurement practices into AI systems.

Out of the 58 items, there were 5 instances in which no edits were made to AI-generated stems, and 9 instances in which no edits were made to AI-generated answer options. Again, this is evidence that SMEs were generally active in working with the draft materials, but that in some cases the AI-generated materials were of sufficient quality to move them forward to the peer review phase of development.

Case Study 2: Catalog To Support Tagging

Another use case for customized AI systems specifically for pedagogical and assessment-related insights to execute alignments. That is, using customized AI systems to align assessment materials such as items, instructional materials, and texts from across all content areas to frameworks such as competencies, CEFR, national standards, learning objectives, Bloom's taxonomy, and assessment blueprints. This is particularly exciting due to current methods being used in the field to conduct alignments and the well-known problems that each brings. At present, alignments of educational material are typically done using one of four methods: manually, by keyword, by semantic similarity, or by crosswalk.

The manual process involves having subject matter experts (SMEs) read the original materials and make a personal decision about which aspect(s) or standard(s) of a framework the item content is aligned with, and then manually match these up and tag them as being associated. Unfortunately, when using this approach, the alignment results from individuals inevitably vary even though the materials and framework do not. Whether due to differences of opinion in expert judgment, inconsistent interpretation of a framework, or lack of attention, alignments done by multiple or even single SMEs do not tend to provide repeatable results.

Another conventional approach for executing alignments is to use Natural Language Processing (NLP) technology and applying keyword searches. This method involves identifying specific words to search for in the content, and establishing rules for assigning specific tags based on the search results. Using a keyword approach typically results in overly focusing on the content or topic and fails to consider other critical process or behavioral aspects of an item that elicit what an examinee should know and be able to do. The keyword approach also requires a considerable upfront investment of SME resources to identify and cross-check potential keywords that are representative of every standard without triggering too many false positive matches with the wrong standards.

Another alignment approach involves a computational linguistic strategy of calculating the semantic similarity between the object being aligned (assessment item, learning content) and framework elements (state standards, learning objectives, etc.). Using text embeddings, an individual piece of content such as an assessment item can be compared to every framework element such as learning standards, and values can be computed that represent the similarity in meaning between all pairs. Afterward, the standard with the top similarity value can be assigned, or a SME could select the best standard from the top several options.

For some domains semantic similarity may be a useful approach, specifically when items and standards are expected to be highly semantically similar and similar in content, such as science standards that focus on core science ideas and recall-type items (e.g., both refer to "phases of the moon"). However, this method does not work for other domains or situations where standards and items are not expected to be semantically similar. For example, consider a reading standard that says, "Read closely to determine what the text says explicitly/implicitly and make logical inferences from it," and an item that asks, "Why did the narrator choose a particular course of action?" Semantic similarity is not a strong approach when working with rich assessment and learning tasks that go beyond knowledge of content topics.

A fourth alignment/tagging approach involves using crosswalks of frameworks (e.g., state standards, test blueprints, learning objectives, cognitive complexities). The crosswalk approach focuses on relating all elements within one framework to a different framework, then using the transitive property to infer resultant mappings. This process only involves the frameworks and does not directly involve the material being aligned. Step one is to associate each element in Framework 1 to the most similar element that can be found in Framework 2 (e.g., a state standard in one state associated as nearly the same as a similar state standard in a different state). Then, the reasoning is that any assessment task or lesson that had been associated with that element/statement in Framework 1 should now be considered aligned with the statement in Framework 2 that had been identified as aligning with Framework 1.

Assessment and educational experts use the crosswalk approach in order to try to save time and resources. Without carefully considering the assessment task, the crosswalk approach enables alignment based solely on SME ideas on relationships among the statements of the frameworks. The crosswalk approach is particularly notorious due largely to the limitation of not using the primary source of text of the materials that are being aligned. Without looking at the actual tasks and lessons, subtleties for why they had been aligned to a particular element of a framework may be missed. Additionally, since frameworks rarely align directly or use the same language in the same way, associations must be interpolated at best, rather than interpreted from direct evidence. Another problem with the crosswalk approach is that errors are cascaded and proliferated throughout the project. If one element is not truly similar to another it had been associated with, then everything coming in and out of that relationship contributes to errors.

Given the heavy lift required, sometimes this alignment work is outsourced to an external group who uses one or more of these methods. Inevitably, external SMEs lack the insight of internal SMEs about the materials themselves, which hinders the ability to align accurately. Also, the external group may or may not have the requisite experience with the desired framework to infer necessary interpretations of how the framework is intended to be operationalized in the relevant assessment or learning context. Ironically, outsourced alignments must be audited and reviewed by the very SMEs whose time was meant to be protected by outsourcing the work in the first place.

Regardless of method, any error or inconsistency in tagging introduces significant consequential error when assembling an assessment or when providing remediation recommendations for improvement. Tagging tasks must be as error-free as possible. In addition, given how quickly things are changing in educational settings educators and assessment developers should be assuring that all tagging is consistent and up to date thus enabling instruction and assessment to be more consistent and accurate across all materials.

As discussed previously, current alignment strategies (manual, keywords, semantic similarity, and crosswalks) are fraught with known problems. This specific study features multiple aspects of using customized AI to perform alignments.

Research Questions

First is answering the question: how much time is saved. For that inference, we will look at a use case of a K-12 educational service center in Texas that had to maintain and align an item bank comprising 90,000 test questions to state standards amidst evolving educational trends and the introduction of technology-enhanced items (TEIs). The next question is, can a customized AI model align items to multiple frameworks and provide evidence-based justifications all at one time. Finally, the third question is answering the most common question that is asked about the customized AI tagging technology: namely, how good is it? In this case, hundreds of assessment tasks had each been aligned with multiple frameworks so preexisting tags for each item and each framework were available for comparison to assess quality.

Methods

Each of these studies involved developing a customized AI model to align materials and provide evidence-based rationales justifying the application of each tag. The first use case involved 90,000 items. The model was designed to align items to Webb’s DOK framework which gives inferences about the cognitive complexity of each assessment task. The next case study involved developing a customized AI model to be able to align and provide evidence-based rationales for nearly 600 assessment items to six different frameworks simultaneously.

Each framework in this study focused on a different aspect of the construct. One framework consisted of task descriptions that were highly technical in nature, focusing primarily on the content of the assessment task. Another framework focused on competencies that could be measured by executing the task. A framework focused on inferences that could be made about the examinee’s social-emotional or foundational/durable skills. Another framework required inferences about the examinee’s proficiency with respect to different process skills. The final framework required inferences about the level of cognitive complexity executed by the examinee during the task. SMEs were then asked to review the tags and provide feedback about accuracy.

Results And Discussion

Regarding the first case study of 90,000 items, results included an 88% reduction in item alignment time, and 85% cost savings over manual methods. Additional quality metrics included a 96% accuracy rate in content alignment. In the second use case, 600 items were aligned successfully to 6 frameworks resulting in roughly 3600 alignment decisions each with a customized evidence-based justification for each tagging decision. Regarding the effectiveness, the assessment items and their tags were provided back to SMEs along with previous tags. Initial agreement of the AI-assigned tags compared to the previous tags can be seen in Table 5.

Table 5

Initial Agreement Between AI-assigned and SME-assigned Tags

Framework	Initial Agreement
Technical content	41.4%
Competencies	64.0%
Discipline	59.1%
Foundational Skills	60.8%
Process Skills	65.6%
Thinking Skills	72.6%

SMEs were then told to review the quality of the newly assigned AI tags and provide any corrections. Table 6 shows the proportion of mismatches where the SME agreed with the AI-assigned tag over the original tag assigned by an SME.

Table 6

Mismatched Tags: SME Agreement With AI-assigned Tags Over Original SME-Assigned Tags

Framework	Agreement
Technical content	88.1%
Competencies	75.1%
Discipline	74.1%
Foundational Skills	71.8%
Process Skills	64.3%
Thinking Skills	49.3%

Table 7 shows the level of SME agreement with the AI-assigned tags. Note that this is before this feedback was taken into account and used to recalibrate the custom model. Each SME provided a rationale for the disagreement.

Table 7

SME Agreement With AI-assigned Tags After Updates but Before Model Recalibration

Framework	Agreement
Technical content	93.0%
Competencies	89.8%
Discipline	89.8%
Foundational Skills	89.8%
Process Skills	87.7%
Thinking Skills	86.1%

These results demonstrate SME agreement across frameworks ranging from 86% to 93% agreement before calibration. That is, the feedback provided for those areas of nonagreement was used to recalibrate the AI model, therefore improving tagging performance and increasing agreement even further in the next round.

All of these results suggest that using customized AI systems could significantly decrease time for alignment tasks, increase the consistency of tags and evidence supporting each tag, as well as demonstrating very high levels of accuracy according to SMEs across content areas and frameworks.

Case Study 3: Catalog For Gap Analysis

Another disruptive use case features applying customized AI systems to execute gap analyses about how well current items and assessment materials cover and align with test blueprints and requirements. Typically, the same basic manual and keyword approach described above are also used to conduct gap analyses. Once again, executing this manually takes a great deal of time and ends up being inconsistent due to the role of opinions in alignment decisions. Additionally, executing by keyword often results in an overreliance on content topic and subsequent undervaluing of the skills described by requirements.

Importantly, given the inordinate amount of time and resources the current typical process takes, stakeholders must be judicious with how often such an analysis can be performed. Using customized AI for this purpose is not only faster and more repeatable, but analyses can also be run to provide additional levels of insight. For example, in a typical gap analysis, an item bank can be queried for coverage. However, queries are lacking sufficient insight into how a particular item truly measures a particular construct. Outputs of analyses will be shared regarding whether the material is considered to be a direct match or less-direct match to the framework elements. Additionally, if something is missing, customized AI can provide specific insight into what is covered and what is not covered.

Typically, a gap analysis is used as a summative activity. The result is used to evaluate the final item pool against the framework or test blueprint after assessment development is largely or entirely complete. The goal of the evaluation is, as with most summative assessments, to get a passing grade—i.e., to have the assessment judged as ‘covered’ with the standards it is intended to assess and to serve as one piece of evidence in a validity argument for use in decision-making.

In this approach, robust information about the alignment will be provided to the SME reviewers and provided much earlier in the assessment development process. For example, an initial set of items may be drafted to cover only one aspect of the framework. That set can be submitted to the evaluation system immediately, to see whether the system agrees with the coverage estimation, whether it uncovers other aspects of the framework also assessed in the set, and whether it adequately covers the selected part of a given learning standard. Here, the alignment system can provide information about what learning standards are covered—and what parts are not—for each item to be routed back to the assessment developer. The AI system can provide evidence in a narrative format, explaining why the item was associated with a particular framework element. Feedback on this analytic evidence set from the SMEs may be given back to the AI scientists and psychometricians so that updates and refinements to the AI model can be made, making future iterations of the evaluation steadily more precise. As more items are added to the set and more framework coverage is assumed, they can be rapidly verified by the alignment system.

Using AI to execute these analyses can be repeated as many times as desired with consistent results, where repetition with subject-matter experts is time-consuming and costly, in addition to the likelihood of disagreement between SMEs. The ability to check coverage repeatedly, rapidly, accurately, and easily will ensure that the final product is fully aligned to the relevant learning objectives, with no gaps or weak points of coverage. Routinely reviewing accuracy and breadth of coverage should improve the assessment development process, while also making it faster and more efficient. This partnership optimizes the combination of strengths from human expertise with automated system consistency, speed, and accuracy.

This third and final use case comes from the need in primary through upper secondary educational settings to understand systematic and rigorous coverage of educational concepts vertically as well as horizontally. Documenting where and when prerequisite skills are taught offers insights and should provide scaffolding for learner pathways. Currently, this is typically approached by coordinating teacher panels and collecting their professional opinions about scope and sequence. The challenge comes from how much time these analyses take and, again, how much opinion may vary among teachers. Additionally, the world is changing so fast, desired skills and learning standards are updated frequently, and educators must keep pace in order to make sure students are well prepared for college and beyond.

Unfortunately, any significant change in curriculum immediately evokes the need for a new analysis. Given how long and how much time scope and sequence documents take to develop, teachers are restricted from making changes at the risk of introducing problems into educational progression. Instead, examination of results from applying AI customized for assessment insights related to multiple scope and sequence situations, and it will discuss what insights were able to be uncovered with respect to covering material in optimal pathways, identifying gaps in instruction, and identifying whether prerequisite skills were indeed covered adequately. This customized approach goes beyond text embeddings to employ the latest techniques to encode content in a context-attentive fashion, thus enabling valid and repeatable capture of deep conceptual and contextual relations in item content as well as in the educational/workforce frameworks, facilitating alignment of those materials to each other.

This study features a use case of applying customized AI technology in order to obtain K-12-specific inferences with respect to identifying potential instructional issues and opportunities to improve student performance. The study involves one large U.S. public K-12 school district made up of over 40,000 students and over 2000 teachers, distributed across more than 30 different schools.

School district leadership had analyzed student test data across grades and subjects from previous school years. The content area identified as having the biggest deviation from desired performance was in a specific Algebra course taught at multiple schools within the district. One restriction of the study was that student performance data would not be available. Therefore, the decision was made to execute AI-assisted analysis of the instructional materials with a particular focus on how well these materials covered the requisite knowledge, skills and abilities described in detail in the benchmark statements and descriptors found in the state standards.

Materials eligible for analysis included primarily instructional artifacts. These materials included the state specific scope and sequence documentation, state standards and benchmarks, assessment blueprints and unit tests, lessons and student-facing instructional materials. Benchmark level descriptions and evidence found in the state standards totaled around 70 different statements. The student-facing instructional materials consisted of 74 lessons. The assessment blueprints and sample assessments were at the lesson level for the Algebra course.

Research Questions

The research questions guiding this study were, could we use customized AI tools to help make evidence-based inferences on how well the current instructional materials covered desired topic areas? And could any instructional gaps be identified that might potentially account for lower student performance?

Methods

The first step of the study involved developing the customized AI system (Catalog) such that it would provide multiple insights. The most critical inferences were having the customized model provide primary tags relating to the benchmark level of specificity and secondary tags when appropriate, along with evidence-based rationales for those tagging decisions. The customized AI model also needed to provide prerequisite skills in terms of the benchmarks from lower grades. Having prerequisites identified for the instruction enables educational leaders to see any places in the current curriculum that students might struggle if they are not at sufficient proficiency at the start of the instruction. In those places, adding specific remediation strategies at the start of these lessons might increase the student access to the current instruction and increase engagement.

Once the AI model was developed, next steps included executing multiple analyses of the various instructional materials. The first analysis was a unit-level analysis of over 70 instructional units of student-facing instruction, problems, activities, and practice problems. All of the various student-facing materials and teacher plans were provided to the customized AI model. The model analyzed each complete unit separately and provided primary and secondary tags at the benchmark level along with evidence-based justifications for each of the tagging decisions. Each instructional unit had been previously tagged to benchmark level by an unidentified source, but that information was not used in the analysis. In addition, prerequisite skills, as articulated by the benchmarks from previous grades, were provided for each lesson. The rationale for having this information again is that if students were lacking sufficient proficiency in prerequisite skills and knowledge, they might not be able to fully engage and benefit from the instruction.

The second analysis was at the sub-unit level of the curricular materials. That is, each of the Algebra units were broken into 15-20 subunits, totaling 874 subunits. In this analysis, the customized AI provided primary and secondary tags at the benchmark level of specificity for each of the subunits. Note that this more specific level of analysis had not been completed prior to this analysis. Tagging at the sub-unit level provides a more detailed view of the instruction and coverage within each unit, revealing benchmark coverage learning during the days and hours within the entire unit.

Final steps of the study included executing multiple comparisons of information obtained by the AI-enabled analyses to current documentation of instructional and assessment coverage.

Results

Results from tagging at the unit level are presented first. Primary and secondary tags and their evidence-based justifications provided along with prerequisite skills and knowledge required for each of the lessons. The primary and secondary tags from the customized AI tagging were then compared to the preexisting tags provided for the lessons from a different source. Results showed 81% agreement between the AI-tagging and the preexisting tags for the primary benchmarks covered by the course. The remaining 19% of the primary tags were different. Some of the AI-identified primary benchmarks identified were quite similar to the previously identified tags, however, others were quite different. The evidence-based justifications made the task of validating the AI-provided tags easy and direct. Unfortunately, as is commonly the case, the preexisting tags were not accompanied by any information about how they were decided or justification for the tags.

Lessons that had a mismatch of AI-assigned tags to preexisting tags were pointed out for educational leaders to consider. For example, one of those mismatches revealed that a multiple-day instructional unit was actually designed to cover the content for a particular kind of function. In fact, the standards and benchmarks required for this particular course did not require primary coverage of that topic. In this case, those multiple days might be better spent not covering that lesson, but instead covering something else more important.

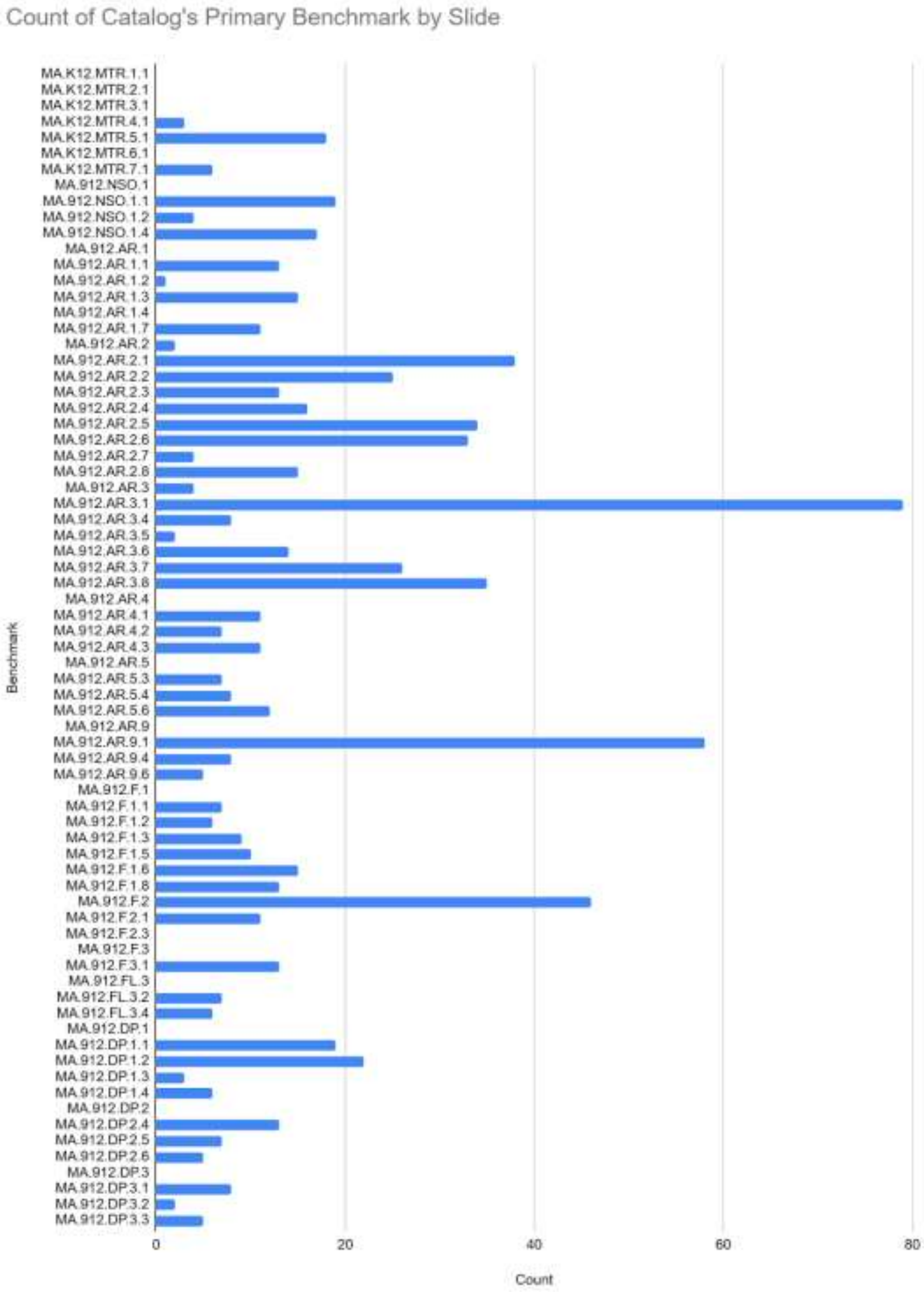
Second, the distributions of primary benchmark coverage at the sub-unit level were analyzed. These results are shared in Figure 1. This analysis provided insight into the hourly and daily coverage within the instructional subunits so that the district could easily make a judgment about whether all benchmarks were being sufficiently covered. The prerequisite skills for each subunit of instruction were also listed in case known issues in previous proficiency could be responsible for impeding efficacy of the instruction.

Some of the most compelling results were the comparisons of the primary content coverage (benchmark tags) as identified by AI to the assessment blueprint for particular instructional lessons. The analysis for the first unit revealed that 33.3% of the blueprint was not covered by primary instruction according to the AI. The analysis for the second unit assessment revealed 50% of the assessment were benchmarks that did not receive primary unit instruction according to the AI. When shared with district leaders, SMEs confirmed that this analysis actually confirmed their suspicion about the instructional misalignment with the assessment, but they had previously lacked the data to support it.

Overall, the customized AI model was able to provide consistent, evidence-based alignments for multiple levels of instruction that districts and teachers lack time and resources to complete manually. This study demonstrated the power of being able to perform multiple levels of analyses efficiently and accurately in to be able to answer various questions about instructional coverage.

Figure 3

Count of Catalog's primary benchmark by slide.



Discussion

Generative AI has great potential in the K-12 space, including for instruction and assessment. The novel, real-world applications presented in these studies have demonstrated great promise as well as shed light on some lessons learned when working with generative AI.

First, all the real-world studies in these studies have used generative AI systems customized for applications in education and assessment by teams of AI scientists, psychometricians, and experts in measurement and education. As mentioned previously, using customized AI systems are not the same as simply prompting a Large Language Model (LLM) that lacks additional information, training, and expertise in assessment and instruction. Therefore, the gains and efficiencies of the customized systems are not expected to be reproducible using a general LLM, unspecific to a particular domain. The generative system is most effective at outputting high quality material when examples, descriptions, knowledge, and elaborations from the domain experts are integrated into the pipeline.

A second lesson learned is the importance of training, both for the AI model and for the SMEs using the customized system. If a customized model is being developed, ensuring high-quality exemplars are featured as the majority of the training set can improve the quality of initial drafts of items coming from the customized model. Many times, test developers will tend to want to use higher numbers of examples of items that they have available to customize their model rather than choosing fewer, but higher-quality questions. The problem with large numbers of items is that the likelihood is greater for those items to be ones that the organization does not deem as high quality. When a model is trained with lower quality inputs, then the greater the likelihood for lower quality drafts being produced by the customized model. The higher quality the training set, the better the customized model will be. Just as important as the training set is the SMEs chosen to interact with the customized system. The best SMEs for working with the AI system are people that are eager to use the technology and have a positive attitude about the potential of doing things slightly differently. They should be the kinds of SMEs that are motivated to use all of the features actively. Taking actions like regenerating and editing stems and options that are not ideal gives very helpful and actionable feedback so that the model improves much more quickly and efficiently.

A third lesson learned is to stress that this is a new way of doing things so therefore the outputs of interest are slightly different. For example, when talking about developing a customized AI model for item generation, the output of interest is not “number of items” as much as a customized AI model that is able to produce a high-quality draft at any level of specificity, cognitive level, of any kind across the entire test blueprint. Similarly, when considering AI-enabled alignment, the output is not just a single alignment as much as a customized model specific to the framework and materials provided that can be validated, calibrated and reused producing extremely reliable and consistent results.

A fourth lesson learned is to be careful when choosing metrics to evaluate the quality of the AI tools. As we have seen, when it comes to alignment, mismatches should be investigated fully and not just presupposed to be due to either the AI or the SME being incorrect. Many organizations will want teachers and SMEs to be the arbiters of quality. Many SMEs have developed their own heuristics and notes to save time when aligning materials. Unfortunately, many times those heuristics may not work as well as taking a fresh look at each item and each framework element as the AI is doing. Similarly, the AI model should not be over- or under-rated. The AI model needs to be checked to be sure inferences are being made appropriately according to evidence.

Conclusion

The research undertaken on the application of generative AI within the K12 educational setting highlights significant potential for these technologies to impact and enhance various educational processes. From assessment item development utilizing Finetune Generate to the intelligent tagging and alignment of content with Finetune Catalog, our findings present robust evidence supporting the efficiency and efficacy of customized AI systems. The case studies underscore the tangible benefits these technologies can offer, such as substantial reductions in time and costs, marked increases in consistency and accuracy, and improvements in the quality of educational content and its alignment with standards and learning objectives.

Key lessons have emerged from these studies, the foremost being the irreplaceable role of customization in achieving high-quality output from AI systems. Off-the-shelf LLMs, while powerful, do not match the efficacy of models tailored specifically to the nuances of educational content and assessment requirements. This customization involves crucial input from domain experts, high-quality training data, and continuous interaction and feedback from SMEs to refine model performance. This underscores a new paradigm in AI application where the fusion of advanced computational techniques and human expertise yields superior results. Therefore, researchers and practitioners, alike, should not settle for using off-the-shelf, general models to generate draft-quality assessment content for further refinement by SMEs, as the results will be lacking in terms of how well generated content adheres to specific style and structure specifications, and also how well it upholds the fundamental principles of assessment. The best results will come from collaborative system-building done through cooperation between content experts, psychometricians, measurement scientists, and AI scientists.

Furthermore, our research highlights the importance of rigorous training for both AI models and human users. The effectiveness of AI systems in generating and refining educational content greatly depends on the quality of the training data fed into the models and the proficiency of SMEs in leveraging these systems. The active engagement of motivated and knowledgeable SMEs in using AI tools ensures that the outputs are continuously improved, and the AI systems evolve to meet the specific needs and standards of educational contexts. This collaborative approach not only enhances the AI's performance but also fosters greater acceptance and utilization of technology among educators. It should not be expected that SMEs who are simply given access to AI-powered tools will figure out the best way to accomplish efficiency and quality gains. Specific training on how to use the AI-powered systems is a must, and providing time for learning the systems is critical.

Finally, it is critical to adopt appropriate metrics for evaluating AI systems. Traditional measures may fall short in capturing the nuanced improvements AI can bring to educational processes. For instance, merely calculating metrics like agreement percentage between the AI and existing tags misses the opportunity to consider evidence for a fresh approach to tagging decisions. Researchers and practitioners should strongly consider that specific instances may occur where AI-assigned tags could be as accurate, or more accurate, than SME-assigned tags—whether due to real advantages of AI analysis of associations between content and tags, or due to possibilities such as fatigue on the part of human SMEs. A better performance measure than agreement with existing tags may be SME agreement with AI-generated rationales explaining why certain tags were assigned. This holistic approach to evaluation will help stakeholders better understand and appreciate the profound impacts of generative AI in education, ultimately driving forward its integration and advancement.

An additional note is warranted for researchers and practitioners who would use generative AI for assessment and education applications: AI models are continuously and rapidly evolving, as well as learning new information, which means previously generated assessments and constructs could be called into question as new versions of models are released and developed (Kaldaras et al., 2024). This is a suggestion against using AI generated assertions and materials directly in production-level applications, and a suggestion for continuing to have SMEs retain final control, to refine and smooth over implicit assertions and choices made by AI models that could be inconstant as new versions are rolled out.

As we look ahead, the future of AI-assisted test development and AI-assisted tagging work is bright. With continuous advancements in AI, particularly in the development of even more sophisticated and contextually aware models, we can anticipate continued enhancements in the precision, speed, and creativity of educational content creation. The capacity for seamless, real-time alignment of educational materials to evolving standards and the personalized adaptation of learning resources to meet individual student needs will revolutionize instructional practices. Our initial studies are promising, and we envision a future where educators are empowered with AI tools that not only relieve them of repetitive tasks but also open up new horizons for innovative teaching strategies, enabling a richer, more responsive educational environment for all learners.

Declarations

Gen-AI use : The authors of this article declare (Declaration Form #: 212241835) that Large Language Models (LLMs), were used for writing and editing text in up to 10% of the article. They further affirm that all content generated by GenAI has been carefully reviewed, and they assume full responsibility for its inclusion.

Conflict of Interest: None

References

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

Bennett, R., & Zhang, M. (2015). Validity and Automated Scoring. *Technology and Testing*, 142–173. Routledge. https://doi.org/10.4324/9781315871493-8

Bolender, B., Foster, C., & Vispoel, S. (2023). The Criticality of Implementing Principled Design When Using AI Technologies in Test Development. *Language Assessment Quarterly*, *20*(4-5), 512–519. https://doi.org/10.1080/15434303.2023.2288266

Burrows, S., Gurevych, I., & Stein, B. (2015). The Eras and Trends of Automatic Short Answer Grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Gierl, M. J., & Haladyna, T. M. (2013). *Automatic Item Generation: Theory and Practice*. https://doi.org/10.4324/9780203803912

Hao, J., Alina, Yaneva, V., Lottridge, S., Matthias von Davier, & Harris, D. J. (2024). Transforming Assessment: The Impacts and Implications of Large Language Models and Generative AI. *Educational Measurement*. https://doi.org/10.1111/emip.12602

Ho, A. D. (2024). Artificial Intelligence and Educational Measurement: Opportunities and Threats. *Journal of Educational and Behavioral Statistics*, *0*(0). https://doi.org/10.3102/10769986241248771

Kaldaras, L., Akaeze, H. O., & Reckase, M. D. (2024). Developing Valid assessments in the Era of Generative Artificial Intelligence. Frontiers in Education (Vol. 9, p. 1399377). https://doi.org/10.3389/feduc.2024.1399377

Khan, S., Rosaler, J., Hamer, J., & Almeida, T. (2021a). Catalog: An educational content tagging system. In Hsiao, I., Sahebi, S., Bouchet, F., Vie, J. (Eds.), *Proceedings of the International Conference on Educational Data Mining*, 736-740. International Educational Data Mining Society.

Khan, S., Hamer, J., & Almeida, T. (2021b). Generate: A NLG system for educational content creation. In Hsiao, I., Sahebi, S., Bouchet, F., Vie, J. (Eds.), *Proceedings of the International Conference on Educational Data Mining*, 741-744. International Educational Data Mining Society.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady 10*(8), 707-710.

Yan, D., Rupp, A. A., & Foltz, P. W. (2020). *Handbook of Automated Scoring*. CRC Press.

ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2024; 15(Special issue); 378-394

Ask NAEP: A Generative AI Assistant for Querying Assessment Information

Ting ZHANG* Luke PATTERSON** Maggie BEITING-PARRISH*** Blue WEBB****
Bhashithe ABEYSINGHE***** Paul BAILEY****** Emmanuel SIKALI*******

Abstract

Ask NAEP, a chatbot built with the Retrieval-Augmented Generation (RAG) technique, aims to provide accurate and comprehensive responses to queries about publicly available information of the National Assessment of Educational Progress (NAEP). This study presents an evaluation of this chatbot's performance in generating high-quality responses. We conducted a series of experiments to explore the impact of incorporating a retrieval component into GPT-3.5 and GPT-4o large language models and evaluated the combined retrieval and generative processes. This work presents a multidimensional evaluation framework using an ordinal scale to assess three dimensions of chatbot performance: correctness, completeness, and communication. Human evaluators assessed the quality of responses across various NAEP subjects. The findings revealed that GPT-4o consistently outperformed GPT-3.5, with statistically significant improvements across all dimensions. Incorporating retrieval into the pipeline further enhanced performance. The RAG approach resulted in high-quality responses. Ask NAEP reduced the occurrence of hallucinations by increasing the correctness measure from 85.5% of questions to 92.7%, a 50% reduction in non-passing responses. The study demonstrates that leveraging large language models (LLMs) like GPT-4o, along with a robust RAG technique, significantly improves the quality of responses generated by the Ask NAEP chatbot. These enhancements can help users to better navigate the extensive NAEP documentation more effectively by providing accurate responses to their queries.

Keywords: *Generative AI, chatbot, NAEP*

Introduction

The purpose of this paper is to introduce an information retrieval chatbot powered by generative artificial intelligence (GAI). This chatbot aims to enhance public access to the National Assessment of Educational Progress (NAEP) publicly available online sources and facilitate knowledge sharing for the National Center for Education Statistics (NCES). The chatbot, Ask NAEP, answers user queries based on publicly accessible information from the NCES website with relevant web links (see Figure 1). By incorporating cutting-edge Gen AI techniques and ensuring a rigorous evaluation, the chatbot strives to deliver timely, accurate, and comprehensive responses.

This paper begins by describing the context and development of the chatbot, including its design philosophy, framework, and the challenges the project faced along with our corresponding solutions. The subsequent sections cover the evaluation methodology and results. The report concludes with a discussion of the findings and outlines future directions for continuing to develop the chatbot.

Context

NAEP is the longest-standing federally funded U.S. assessment. As an assessment arm of NCES, NAEP's mission is to inform policymakers, educators, researchers, and the public about what the nation's students know and can do in various subjects through comprehensive reports and on-demand

* Senior Researcher, Dr., American Institutes for Research, USA, tzhang@air.org, ORCID ID: 0009-0001-1724-6141
** Data Scientist, American Institutes for Research, USA, lpatterson@air.org, ORCID ID: 0009-0000-2612-0375
***Impact Fellow, Dr., Federation of American Scientists, USA, mbeitingparrish@fas.org, ORCID ID: 0000-0002-3998-8672
**** Researcher, American Institutes for Research,USA, bwebb@air.org, ORCID ID: 0009-0004-4080-9864
***** Researcher, Dr., American Institutes for Research, USA, babeysinghe@air.org, ORCID ID: 0009-0006-4107-8615
****** Principal Economist, Dr., American Institutes for Research, USA, pbailey@air.org, ORCID ID: 0000-0003-0989-8729
******* Acting Branch Chief: Reporting & Dissemination, Assessment Division, Dr., National Center for Education Statistics, USA, emmanuel.sikali@ed.gov, ORCID ID: 0009-0007-5325-0475

To cite this article:.Zhang,T., Patterson,L., Beiting-Parrish, M., Webb, B., Abeysinghe,B., Bailey,P., Sikali,E., (2024). Ask NAEP: A Generative AI Assistant for Querying Assessment Information, *Journal of Measurement and Evaluation in Education and Psychology, 15*(Special issue), 378-394. https://doi.org/10.21031/epod.1548128

Received: 10.10.2024
Accepted: 2.12.2024

access to results. NAEP is committed to be transparent about the psychometric, sampling design, instrument design, and other scientific methodologies it uses to produce its assessments, surveys, and estimation procedures. To fulfill the mission, NCES documents the information on two main websites: the main NAEP website under NCES (National Center for Education Statistics, 2024a) and the Nation's Report Card (National Center for Education Statistics, 2024b.

These NAEP websites provide a wealth of publicly available information, including well-documented assessment frameworks, survey and assessment methodologies, data on participating teachers and schools, student questionnaires, and results from decades of assessments. However, locating information on NCES websites can be particularly challenging for NAEP users due to the vast quantity of documents developed over time by different vendors, with older releases rarely removed. Web pages may contain overlapping information (e.g., sampling designs) and inconsistent details (e.g., the number of plausible values in NAEP). Answers to questions often need to be retrieved from multiple documents or resources and verified for their currency. Some example queries include: What content is in the 2018 NAEP Technology and Engineering Literacy assessment? Can I opt my child out of participating in the NAEP assessment? And What is stratification in NAEP sample design? (see more examples in Table 1).

Development of the Generative AI Chatbot

Large language models (LLMs), such as the GPT(Brown et al., 2020), Llama (Touvron et al., 2023), and Gemini (Anil et al., 2024) models, have demonstrated powerful capacities in language understanding and generation. Most can generate responses to users' queries with patterns of speech that closely resemble those of humans (Gao et al., 2023). However, these models are trained on large datasets that may not be curated exclusively for reliability, and their output is not specifically evaluated for accuracy (Abeysinghe & Circi, 2024). Additionally, some models have limitations in providing up-to-date and content-specific information. Although trained on vast amounts of data, they may still miss specific or niche information, and their knowledge is fixed at the time of training and confined to what they encountered during that training (Gao et al., 2023).

Through this work, we sought to develop an information retrieval chatbot, Ask NAEP, to provide responses to users' queries on NAEP information. We do not claim to have perfect accuracy in all responses, as it would be a claim that is unprovable and inflated. However, in this article, we describe how we worked to increase the quality of responses based on three dimensions: correctness, completeness, and communication.

The RAG Framework and Technology

Retrieval-augmented generation (RAG) is a mechanism that combines the strengths of information retrieval and generative models to produce more accurate and contextually relevant responses. The RAG architecture was introduced to address some of the limitations of purely generative models by incorporating an external knowledge retrieval step before generating a response (Gao et al., 2023).

We used a RAG pipeline that retrieves relevant information from a customized knowledge base. This knowledge base aggregates data from the NAEP application programming interfaces (APIs) and content-related text from web pages under the Nation's Report Card (NRC) subsection of the NCES website as well as under the NRC website. The process is shown in Figure 1 and described in this section, with reference to the steps shown in Figure 1.

Figure 1
Architecture of Ask NAEP.

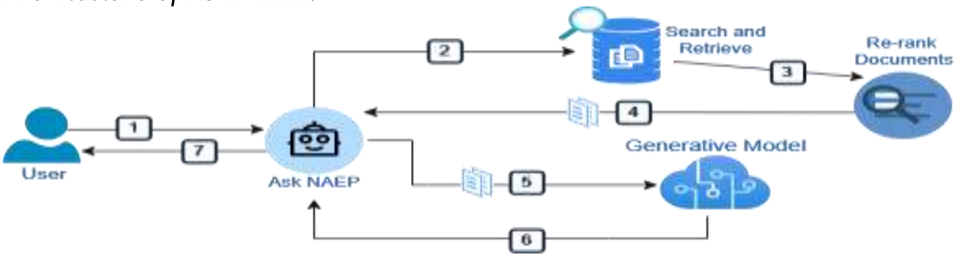


Figure 1 illustrates the workflow of the Ask NAEP Architecture. Upon receiving a user query, relevant documents are retrieved from a vector database and subsequently reranked. The query and documents are then sent to the agent, and an LLM is used to generate the final response, which is then returned to the user.

The information in the knowledge base is projected into numeric vector embeddings using OpenAI's *text-embedding-ada-002* model. When users submit a query, it is converted into a vector embedding using the same model, and documents with the closest vectors to the query (i.e., the most similar vector embeddings) are retrieved (Figure 1, steps 1 and 2). For persistent data storage, we use ChromaDB, which we chose because it is open source, self-hostable, lightweight, and easily integrated into Python applications. It also offers customization options for the parameters used in its search algorithm, Hierarchical Navigable Small Worlds (HNSW) (Malkov & Yashunin, 2018). We measure distance using cosine similarity. Once the top documents have been retrieved using this metric, they are reranked using a cross-encoder model. Cross-encoder models output relevance scores for document query pairs, which are learned through supervised training. Our framework currently uses the *ms-marco-MiniLM-L-6-v2* cross-encoder model from HuggingFace (HuggingFace, 2024).

The query, along with a prompt and the reranked documents, is sent to an OpenAI LLM to generate responses (Figure 1, steps 3, 4, and 5). Our RAG framework offers developers the flexibility to choose from various LLMs, including different versions of GPT models. The overall application was developed in Python, with the front end currently deployed in a preproduction environment using a Flask application.

Finally, the user is shown both the chatbot's response and the top documents associated with the response (Figure 1, step 6 and 7).

Knowledge Base

The NCES NAEP websites are a compendium of assessments and results, information for parents, students, researchers, media, school administrators, teachers, and resources for researchers and educators. It also includes a variety of data tools, state and district profiles, etc. To illustrate the complexity of this website, we unpack a small section of the resources for researchers, specifically the NAEP Technical Documentation Website (TDW, National Center for Education Statistics, 2024c). The TDW is the technical description of all the operations that NAEP has used to conduct and assess students since 2000. Prior to this, technical documentation reports were printed. Altogether, there are about 37,000 static and interactive pages on the NRC. The static pages are on the main NAEP website, while all interactive pages are on the NRC.

We conducted an extensive crawl of the NCES websites using the open-source Scrapy(GitHub, 2024a) and Selenium(GitHub, 2024b) modules in Python for crawling, collecting the raw HTML from about 5,000 pages. The purpose of these scrapes was to collect unprocessed HTML to retain in persistent storage, allowing us to experiment with different approaches to processing and splitting the page contents. From the raw HTML, we extracted items such as paragraph text, alt text for figures, and table titles and contents. We separated the page contents into paragraphs, sections, and titles before creating embeddings and adding documents to our vector database. Sections were identified by programmatically splitting the full-page contents at section headers, which were detected by their use of HTML markup language (e.g., bold text). Within each section, contents were further divided into paragraphs based on the presence of newline characters. As detailed in a subsequent section, the current knowledge base includes documents derived from the paragraph text on these pages, with tables stored as associated metadata in JSON format.

We also sent requests to the data API that powers the state and district profile tools. Each response provided a summary of performance for the specified state or district.

A challenge arose when augmenting our data with content scraped from the Nation's Report Card website, where much of the information is presented in interactive figures or tables that require human interaction to navigate and extract specific results (e.g., the gap between English language learners (ELLs) and non-ELLs in the 2022 grade 8 reading assessment). Unlike static web pages, dynamically rendered content is difficult to scrape programmatically. To address this issue at the state and district levels, we reconstructed API calls to the respective profiles and collected summary texts for each state and district. This data is stored separately and used to answer questions about specific states or districts. Because the NCES website rarely removes pages for older releases, an ongoing challenge is ensuring that the retrieved web page links are both relevant and up to date. In some cases, pages pertain to specific years, grades, and subjects, which we identify through keyword detection in the user's query and apply as a filter. If no such keywords appear in the query, no prefiltering is applied to the knowledge base, and all pages are considered in the similarity search.

Evaluation Approach

Testing Queries

We evaluated Ask NAEP using a bank of expert-generated questions. We selected 55 questions that experts thought most representative of common and important questions that individuals might seek answers to on the NAEP website. These questions were categorized into the topics shown in Table 1 for further analysis:

Table 1

NAEP Questions

Topic	Example Question	Number of Questions
NAEP Content Areas and Assessments	What content is in the 2018 NAEP Technology and Engineering Literacy assessment?	12
NAEP Data Analysis and Statistical Techniques	How are NAEP plausible values used to conduct secondary analysis?	11
NAEP Scores and Achievement Levels	What are the achievement levels for NAEP in general? What was the average 4th-grade math score for NAEP in 2022?	8
NAEP Participation and Accommodations	Can I opt my child out of participating in the NAEP assessment?	7
NAEP Scoring and Assessment Process	When do constructed-response items need to be rescored?	7
NAEP Sample Design and Methodology	What is stratification in NAEP sample design?	5
NAEP Validity and Reliability	How are items treated if the fit is not good in NAEP?	5

Two human raters from the research team evaluated the responses from various versions of the Ask NAEP chatbot using the CCC framework rubric. Interrater reliability was calculated using Cohen's Kappa (Cohen, 1960).

Evaluation Framework and Metrics

Ideally, interacting with a chatbot should feel like a natural conversation, where the chatbot's written responses are as comprehensible as a text message produced by a human author. With this in mind, we created an initial framework based on Grice's Maxims of Conversation (1989), which views conversation as a collaborative product between two parties who share a common aim. In this case, the aim is to gain a better understanding of some aspect of NAEP, whether it involves procedural information or specific test results.

Within this conversational exchange, there are four maxims that ensure a quality response: quantity, quality, relation, and manner (Grice, 1989). These are especially relevant to the presentation of statistical chatbot responses, which should ideally be long enough to include all necessary information without being burdensome (quantity), be truthful (quality), include only relevant information (relation), and be as concise and clear as possible (manner). Since several of these criteria are specific to individual users, for our purpose, we simplified the system to include three criteria—Correctness, Completeness, and Communication—as outlined in Table 2 below.

Table 2

Framework for Generative Component Evaluation

Construct	Annotation	Description
Correctness	$Q_{correct}$	Is the content of the chatbot's answer factually correct?
Completeness	$Q_{complete}$	Does the chatbot's answer include the relevant facts and information needed to answer the question?
Communication	Q_{comm}	Is the chatbot's answer written in a clear and concise fashion?

The following weights are applied to generate a composite score from the three constructs. Since the primary goal of this chatbot is to deliver accurate, complete, and reliable responses to user queries, we prioritize correctness and completeness over communication by assigning greater weight to the first two dimensions. It is worth noting that these weightings are not based on prior studies or established theories.

$$Q_{composite} = \frac{2}{5}Q_{correct} + \frac{2}{5}Q_{complete} + \frac{1}{5}Q_{comm}$$

Table 3 below describes how these dimensions were graded by human reviewers. Some evaluation analyses are based on "pass/fail" grading. The rubric was constructed so that grades of 3, 4, and 5 represent qualitatively acceptable answers for a published chatbot, while grades of 1 and 2 do not. This is why the threshold for a passing answer is 3 or higher for all dimensions.

Table 3
Dimension Scoring Rubric

Grade	Pass/Fail	Correctness	Completeness	Communication
1 (Poor)	Fail	Significant factual errors or misinformation	Incomplete, missing crucial information	Unclear, convoluted, or difficult to follow
2 (Below Average)	Fail	Some inaccuracies or lack of precision	Lacks relevant details or fails to address all aspects	Lacks coherence and may confuse the reader
3 (Average)	Pass	Several minor inaccuracies, but generally correct	Covers the basics but could benefit from more depth	Clear but could be more concise
4 (Good)	Pass	Generally accurate with 1-2 minor inaccuracies	Sufficiently complete, addressing the main points	Generally clear and concise
5 (Excellent)	Pass	Completely correct with no errors	Comprehensive with thorough information	Succinct, well organized, and easy to understand

A major concern in the present chatbot evaluation process is that, since this chatbot represents the interests of a federal statistical agency, it is imperative that it does not hallucinate—that is, it should not produce any answers that are partially or completely incorrect. These three criteria were applied in various forms to all of the human evaluation work conducted.

Research Questions

One could argue that the only component that needs to be evaluated properly is the generative component and how effective the generated responses are. In a RAG bot, however, the retrieval is an important intermediary that can help diagnose why a chatbot responds correctly or incorrectly to queries. If receiving the correct retrieval is unimportant, RAG is not providing a significant improvement over unaltered ChatGPT, so testing the retrieval is one of the evaluation’s research questions.

Our evaluation of Ask NAEP centers around four research questions (RQ):

RQ1. How satisfied are users with Ask NAEP?
RQ2. Which LLM performed better in the RAG chatbot?
RQ3. How important is good retrieval at generating a quality answer?
RQ4. Does the Ask NAEP retrieval process and bot configuration produce quality answers?

Method

To answer RQ1, we conducted the user testing when Ask NAEP was using GPT-3.5 as its generative model. We consider user testing to be an important component to ensure that the chatbot effectively meets real-world user needs and satisfaction. This method allows for iterative improvements that align the chatbot's performance with actual user behavior and preferences.

Participants included NAEP users from various states across the United States who used Ask NAEP and recorded any unsatisfactory interactions; the focus of this round of human evaluation was negative experiences with the chatbot. Among these users, seven interacted with the chatbot and filled out 13 forms, representing a total of 58 problematic interactions with the chatbot out of a much larger pool of interactions. Users also provided feedback on why the output they received was problematic and answered multiple-choice questions regarding why they flagged the output, whether it was easy to understand (correctness), whether it contained relevant information (completeness), and how the output was communicated (communication). The feedback from this user testing was used to improve the performance of Ask NAEP. The current paper presents the results from this round of user testing.

To answer the second RQ, the research team evaluated the Open AI generative models (e.g., GPT-3.5 and GPT-4o) within the RAG framework to identify the best-performing model. Due to our institutes’ security and efficiency concerns, only OpenAI’s GPT models were tested for powering the generative answers that Ask NAEP produces. Development of the chatbot began when GPT-3.5 was the latest

OpenAI LLM available. However, GPT-4 and GPT-4o have since been released. As part of our evaluation, we assessed whether GPT-4o performed better than GPT-3.5 as the underlying chatbot model. We did this by generating answers to all 55 test questions using GPT-3.5 with no retrieval augmentation, then repeating this process for GPT-4o. We then performed a round of human evaluation of each answer across all dimensions.

In RQ 3, we assessed how well the bot answered questions given proper context, as well as its effectiveness in retrieving relevant content to support its answers. Finally, we combined the two components to address the last RQ. Does the chatbot produce quality answers?

Given the relative novelty of these applications, evaluation methods for a RAG chatbot are still emerging, and the research community has not yet reached a consensus on the most effective evaluation approaches (Abeysinghe & Circi, 2024; Celikyilmaz et al., 2021; Howcroft & Rieser, 2021; Iskender et al., 2021; Schoch et al., 2020; Smith et al., 2022; van der Lee et al., 2019). Consequently, it is evident that a variety of evaluation methods exist, some of which are domain-specific while others are more general in nature (Abd-Alrazaq et al., 2020; Chan et al., 2023; Fu et al., 2023; Sellam et al., 2020; Zhang et al., 2020). To evaluate chatbots, researchers use both human evaluation and automated methods. For this research, we adopted the human evaluation approach.

Results

How Satisfied Are Users with Ask NAEP, the GPT-3.5 Version?

We conducted the user testing when GPT-3.5 was the latest OpenAI LLM available. After users interacted with Ask NAEP and provided feedback between May and June 2024, the data were analyzed to better understand the typical issues with that version of the chatbot needing improvement. These responses were further coded using the Correctness-Completeness-Communication (CCC) framework. One of the authors also investigated why these chatbot responses were flagged as problematic.

In general, out of the 58 problematic responses flagged by the experts, 10 (17.2%) occurred because this version of the chatbot declined to answer. For example, when asked, “How many were below basic?” the chatbot provided responses such as: “I'm sorry, but I cannot provide an answer to your question as it cannot be answered using information about NAEP. Is there anything else I can assist you with?” In this case, the chatbot may have declined to answer because the question wasn’t specific to a particular subject area, grade level, or year.

Additionally, 9 responses (15.5%) were flagged because they included some form of hallucination, meaning that part or all of the answers consisted of completely incorrect information. For example, when asked, “Can I find out how my student did on NAEP?” the chatbot responded, “Yes, you can find out how your student did on NAEP. The NAEP Data Explorer allows users to search for state results by student demographic groups and other variables....” The first sentence is considered a hallucination because NAEP never provides scores down to the individual student level.

At first glance, these values seem very high; however, the experts were instructed to report any problematic chatbot responses, so these values are likely higher than they would be for the entire set of chatbot interactions.

In addition, we asked the experts overall why they flagged each question. The two most common reasons were that the chatbot only partially answered the question (38.5%) or did not answer it at all (16.9%). Despite these issues, most respondents found the chatbot easy to understand (51.6%). When asked whether all responses included relevant information, 36.1% of the respondents agreed that they did. Finally, the experts found that the chatbot communicated in a logical manner with a beginning, middle, and end 71.4% of the time. This feedback suggests that while the chatbot’s information may need refining, its communication style is generally accessible.

Finally, the chatbot output was also scored by one of the authors using the CCC rubric. The results are presented in Table 4, which includes both averages and medians. However, it is important to note that human evaluations often treat rubric scores as continuous values, which may not always be appropriate, as they are ordinal categories (Howcroft & Reiser, 2021). Given this distinction, the scores from this analysis are much lower than those from the larger set of sample questions; however, they remain consistent with the types of response values that were flagged.

Table 4

Average and Median Correctness-Completeness-Communication (CCC) Scores for Flagged Chatbot Output

Question	Average Correctness	Median Correctness	Average Completeness	Median Completeness	Average Communication	Median Communicati on
1	2.33	3	2.46	3	2.85	3
2	3.25	3	3.17	3	3.33	3.5
3	2.82	3	2.91	3	3.27	3
4	2.18	3	2.18	3	3.00	4
5	2.09	2	2.36	3	3.09	3

Which LLM Performed Better in the RAG chatbot?

We compared outcomes from the Ask NAEP GPT-4o without retrieval augmentation to those from the GPT-3.5 version, also without retrieval augmentation. Results are presented below in Table 5.

Table 5

Percentage of Passing Answers for Ask NAEP Without Retrieval by LLMs and Dimension

Dimension	N	GPT-4o, No Retrieval[1]	GPT-3.5, No Retrieval[2]	Percentage Point Difference	Permutation test p-value
Correctness	55	87.2%	70.9%	16.4%**	0.00
Completeness	55	89.1%	74.5%	14.5%**	0.00
Communication	55	98.2%	84.5%	13.6%**	0.01
Overall	55	87.2%	71.8%	15.5%**	0.01

Significant at the **5% confidence level

[1]To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .64.

[2]To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .61.

Table 5 shows the percentage of answers produced by GPT-4o and GPT-3.5 that were rated as acceptable by human evaluators for each dimension. The table reveals that GPT-4o outperformed GPT-3.5 in all three dimensions, as well as in overall performance, with differences statistically significant at various confidence levels. The largest difference was observed in the Communication dimension, where GPT-4o achieved 98.2% acceptable answers, compared to 76.4% for GPT-3.5. This difference was significant at the 1% confidence level. The smallest difference was observed in the Correctness dimension, where GPT-4o achieved 85.5% acceptable answers, compared to 70.9% for GPT-3.5. This difference was still significant at the 10% confidence level. These results suggest that GPT-4o is a better model than GPT-3.5 for this chatbot, so Ask NAEP currently uses GPT-4o.

How Important is Good Retrieval at Generating a Quality Answer?

To begin addressing this question, we first examine whether the Ask NAEP retrieval process performs any better than no retrieval at all. We do this by comparing the performance of Ask NAEP with a version of Ask NAEP that performs no content retrieval (which is simply stock GPT-4o with a system context prompt explaining that it is a helpful assistant that answers questions about NAEP). As a reminder, the dimension scores shown in this section are the GPT-assessed scores, and a passing score is a 3, 4, or 5 for the dimension. Table 6 shows that the Ask NAEP retrieval process leads to improvements in passing answer percentages on the Completeness and Correctness dimensions, as well as in overall performance (though the differences are not statistically significant), with no change in the Communication dimension.

Table 6

Percentage of Passing Answers for Ask NAEP (GPT-4o version) With and Without Retrieval by Dimension

Dimension	N	With Retrieval[1]	No Retrieval[2]	Percentage Point Difference	Permutation test p-value
Correctness	55	92.7%	87.3%	5.4%	0.27
Completeness	55	93.6%	89.1%	4.5%	0.38
Communication	55	97.2%	98.2%	-0.9%	0.99
Overall	55	92.7%	87.2%	5.4%	0.27

Significant at the **5% confidence level.

[1] To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .65.

[2] To test human interrater reliability, two human reviewers rated independently. The overall dimension interrater Cohen's Kappa was .64.

Table 6 shows that Ask NAEP provides acceptable answers to more questions than GPT-4o with no retrieval, but we also want to know whether it provides higher quality answers. To do this, we perform a 5-level ordinal logit regression of a binary Ask NAEP response indicator on each of the four dimensions. For each regression, the dimension score is treated as an ordinal dependent variable Y with 5 ordered categories j. The ordered logistic regression model can be expressed as:

$$logit(P(Y \geq j)) = a_j - \beta X$$

where:

- $logit(P(Y \geq j))$ is the log-odds of the dependent variable Y being greater than or equal to category j.
- a_j are the threshold parameters (cut points) for each category j.
- β is the vector of regression coefficients.
- X is the vector of independent variables. In this case, the only independent variable included is a binary indicator $X_{AskNAEP}$, which equals 1 when the answer was generated by Ask NAEP and 0 when it was generated by the no-retrieval model.

What we are interested in is the value of $\beta_{AskNAEP}$, whose value is the log-odds that the response generated by Ask NAEP is in a higher quality category compared to the response generated by the no-retrieval model. If $\beta_{AskNAEP} > 0$, then answers from Ask NAEP are more likely to be in higher or equal quality categories than those from the no-retrieval model. If $\beta_{AskNAEP} < 0$, then answers from Ask NAEP are more likely to be in lower quality categories. If $\beta_{AskNAEP} = 0$, then there is no difference in quality between the answers from Ask NAEP and the no-retrieval model. Table 7 shows the results of the ordinal regression.

Table 7

Ordinal Regression Estimated Probability of Higher or Equal Rating Using Ask NAEP Retrieval Process

Dimension	*N*	Log Odds	Odds Ratio	*p* value |Log Odds| > 0
Correctness	55	1.14	3.13	0.00**
Completeness	55	1.37	3.92	0.00**
Communication	55	0.02	1.02	0.96
Overall	55	0.71	2.04	0.03**

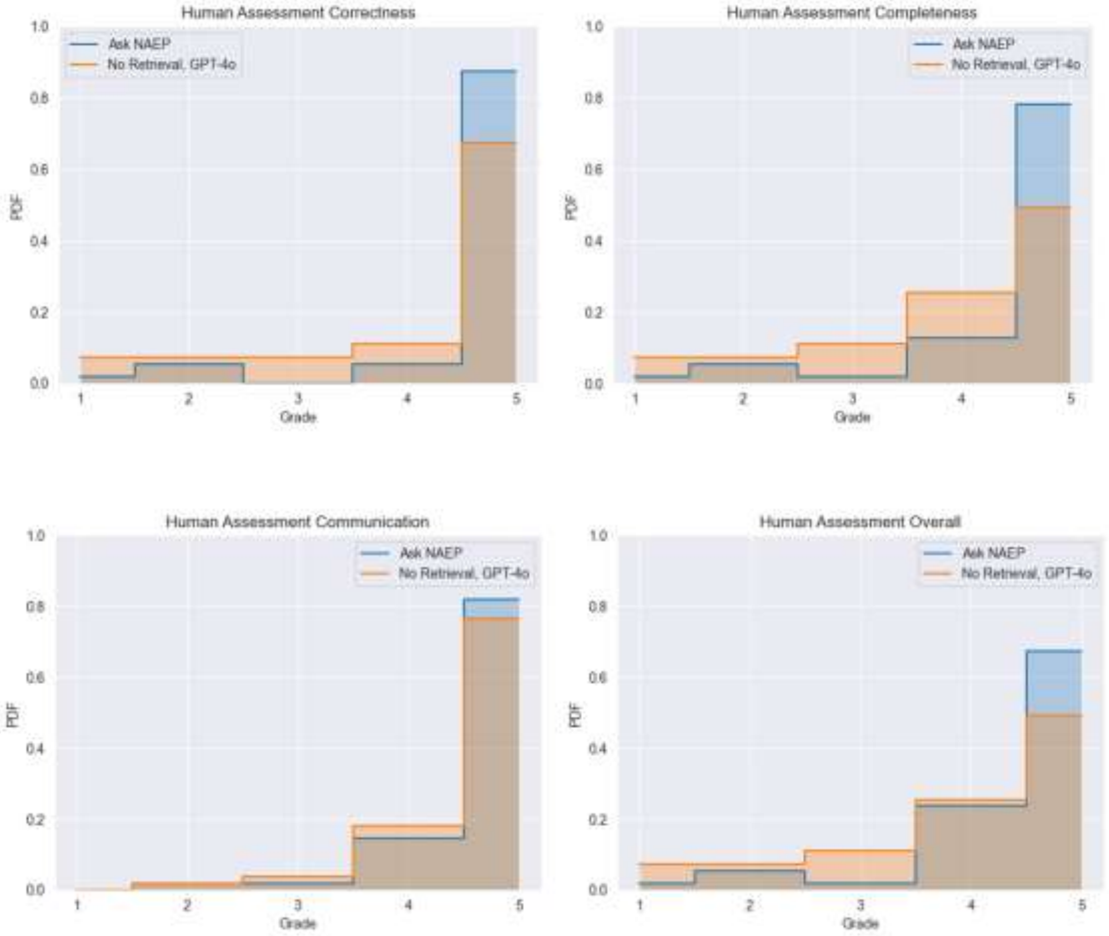
Significant at the **5% confidence level.

The results show that the retrieval process significantly enhances performance in the Completeness and Correctness dimensions, as well as in overall quality. Specifically, the odds of achieving a higher Completeness rating are 3.13 times higher with the Ask NAEP retrieval process, compared to the process with no retrieval. The odds of a higher Correctness rating and a higher Overall rating are 3.92 times and 2.04 times higher, respectively, with the retrieval process. All of these odds are statistically significant. However, the retrieval process does not have a significant impact on the Communication dimension.

These results suggest that the retrieval process is effective in improving the correctness and completeness aspects of response quality, but not necessarily the communication aspect. The cumulative density functions of the assessments for Ask NAEP and GPT-4o with no retrieval are shown in Figure 2.

Figure 2

Probability Density Functions of Dimension Ratings



In Table 8, we examine how the retrieval process impacts the overall score by topic. The NAEP Scores and Achievement Levels topic showed the most improvement. However, the statistical power of this comparison is limited due to the low sample size of questions in each category, so this analysis should be considered exploratory.

Table 8

Overall Percentage of Passing Answers for Ask NAEP (GPT-4o version) With and Without Retrieval by Topic

Topic	*N*	With Retrieval	No Retrieval	Difference
NAEP Sample Design and Methodology	5	100%	100%	0.00
NAEP Data Analysis and Statistical Techniques	11	100%	91%	0.09
NAEP Scoring and Assessment Process	7	100%	86%	0.14
NAEP Scores and Achievement Levels	8	100%	75%	0.25
NAEP Participation and Accommodations	7	93%	100%	-0.07
NAEP Content Areas and Assessments	12	88%	75%	0.13
NAEP Validity and Reliability	5	80%	80%	0.00

Significant at the **5% confidence level.

Does the Ask NAEP Retrieval Process and Bot Configuration Produce Quality Answers?

Ask NAEP attempted to answer all the test questions, and human reviewers gave generally high reviews to these answers across all dimensions. For all dimensions, over 94% of answers received passing grades from human reviewers. Table 9 presents these results, and Figure 3 provides a histogram showing the frequency of each grade for every dimension. Note that *N* is 110 for Table 9 and Figure 3 in this section, as two human reviewers rated bot responses separately for each of the 55 questions, producing 110 reviews in total.

Table 9

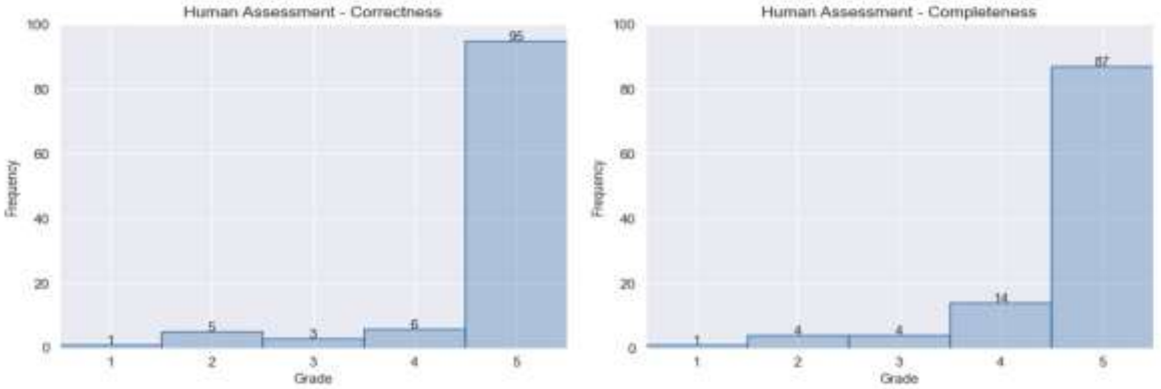
Percentage of Passing Answers for Ask NAEP by Dimension, According to Human Evaluation

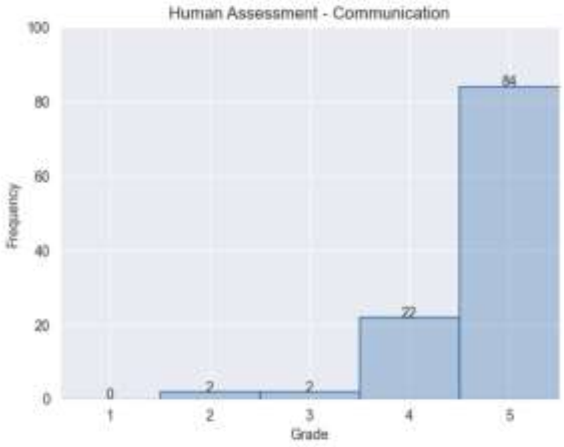
Dimension	*N*[1]	Percentage of Passing Answers
Correctness	110	94.5%
Completeness	110	95.5%
Communication	110	98.2%
Overall	110	94.5%

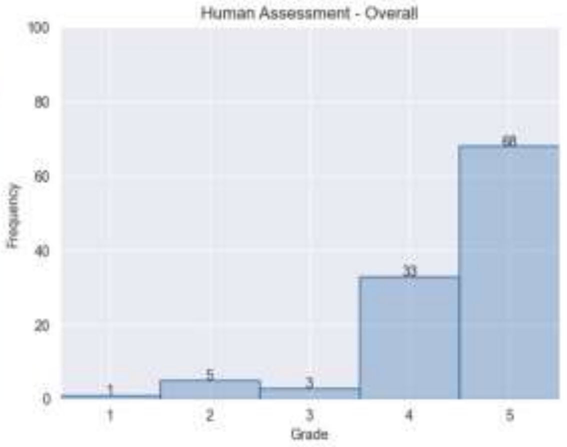
[1] The sample size is 110 because two human reviews are available for each of the 55 questions.

Figure 3

Histograms of Human Assessment of Answered Questions







To better understand which types of queries Ask NAEP answers well, Table 10 shows the percentage of questions with a passing rating for each dimension by topic. Overall, Ask NAEP at this stage is best at answering questions on NAEP Data Analysis and Statistical Techniques, Sample Design and Methodology, Scores and Achievement Levels, and the Scoring and Assessment Process. However, the results indicate a need for improvement in addressing questions related to NAEP Validity and Reliability. This insight has guided the team on which additional documents and data should be integrated in the next phase.

Table 10

Percentage of Passing Answers, Human Assessment of Answered Questions by Topic

Topic	*N*	Correctness	Completeness	Communication	Overall
NAEP Data Analysis and Statistical Techniques	22	100%	100%	100%	100%
NAEP Sample Design and Methodology	10	100%	100%	100%	100%
NAEP Scores and Achievement Levels	16	100%	100%	100%	100%
NAEP Scoring and Assessment Process	14	100%	100%	100%	100%
NAEP Participation and Accommodations	14	93%	93%	100%	93%
NAEP Content Areas and Assessments	24	88%	92%	100%	88%
NAEP Validity and Reliability	10	80%	80%	80%	80%

Discussion and Conclusion

Significance

Ask NAEP demonstrates potential in assisting users to locate the information they need and providing accurate, complete, and comprehensive responses on various NAEP topics. This is particularly true for queries that are summation-based rather than investigative (e.g., questions like “Why did group A perform better than group B?”). In RAG, our corpus is sourced from a federal statistical agency’s website, which undergoes an extensive quality control process. This ensures that the information retrieved is accurate and approved. However, some user questions, particularly ‘why questions,’ may not align with the agency's mission and therefore lack supporting text. Since NAEP information is published by NCES, a statistical agency known for presenting only facts without causal explanations, our chatbot cannot answer investigative questions, especially 'why questions'.

Evaluating language models in generative applications is a challenging task. In this work, we present our evaluation framework, which is an ordinal scale evaluation across three dimensions chosen to assess the quality of Ask NAEP in the context of a federal statistical organization. While other studies have explored dimensional evaluation (e.g., Abeysinghe & Circi, 2024; Fu et al., 2023; Gehrmann et al., 2023; van der Lee et al., 2019, 2021), they are generally more applicable to broader contexts rather than a statistical agency. Therefore, the selection of the proper dimensions for this task is a unique application of our evaluation and is our contribution.

An additional aspect of our contribution is addressing the complexity of existing human evaluation tools, which are often multidimensional and difficult to work with. Previous research has shown that general-purpose rubrics with five or more categories can be challenging for evaluators to use effectively (Wolf et al., 2008). In contrast, the current CCC approach is a much simpler tool for evaluating chatbot output. In developing the CCC approach from our earlier multidimensional framework based on Grice's maxims, we found it easier to apply to chatbot output and more time-efficient compared to the full framework.

Finally, our results indicate that by combining a well-developed RAG mechanism with a more advanced LLM (in this case, GPT-4o), Ask NAEP reduced the occurrence of hallucinations. This improvement is reflected in an increase in our correctness measure from 85.5% to 92.7%. The nonpassing response percent is 100% minus the percent correct and is reduced from 14.5% to 7.3%, this 7.2 percentage point increase in correctness is probably better viewed as a 50% reduction in nonpassing responses.

Principal Findings

In this section, we further analyze and interpret the results that were presented earlier. We also discuss results categorized into the research questions and present the findings accordingly.

The first research question is about user testing. Despite using an older version of Ask NAEP (the bot with GPT-3.5), the results are consistent with the above findings. The GPT 3.5 version performs well on questions based on NAEP's procedures, methodologies, and definitions, including understanding the NAEP assessment process, statistical methods, type of data collected, and assessment purposes. For example, it can accurately answer questions about the NAEP assessment process, how plausible values are drawn, and how biases are addressed in NAEP research studies. It also effectively handles questions about the subjects assessed by NAEP and how NAEP benefits schools and communities.

Conversely, the GPT 3.5 version struggles with questions requiring specific knowledge or data about NAEP assessments, such as the number of items in specific assessments, average scores for specific years, or content from specific years. It also has difficulty with questions about accommodations for disabilities or options for opting out of the assessment. The information and feedback obtained from this round of user testing have been used to enhance Ask NAEP, resulting in improvements to the current version.

Lastly, further investigation of user feedback allowed us to explore additional issues with the chatbot. This analysis revealed that hallucination and refusal to answer remain ongoing issues. Both are generative issues, which may be difficult to resolve without fine-tuning the LLM.

The second research question investigated what LLM should be used in the Ask NCES chatbot context. While acknowledging the existence of other language models, such as Claude and Llama, we limited our experiments to the GPT family for this initial proof of concept. In this work, we present the choice between two large language models, GPT-3.5 and GPT-4o, excluding other elements of the chatbot, such as the embedding process and prompts.

The goal of the second research question was to determine which LLM generates higher quality responses, as judged by human evaluators. For this purpose, a human assessment carried out on 55 questions across different NAEP subjects and administration years revealed that GPT-4o generates much more favorable responses. Our evaluation found that human evaluators rated GPT-4o responses higher than GPT-3.5 responses across all dimensions, and the difference was statistically significant for all dimensions. This finding suggests that despite the increased expenses associated with GPT-4o, its use in critical situations is justified by its superior performance.

In the third research question, we investigated whether adding the retrieval component to GPT-4o would improve the performance on the CCC measures. Our findings show a significant improvement with the addition of retrieval. We are continuing to explore other avenues through which we may enhance retrieval, including alternative embedding models, frameworks for semantic chunking of text, and alternative vector stores that natively support hybrid search.

In addressing the fourth research question, we examined whether combining retrieval with generative processes would result in higher quality bot responses. To test this, we conducted a human evaluation with the CCC measures across two sets of questions: one general and one specific to various NAEP subjects and administrations. Both experiments showed that Ask NAEP attempted to answer the majority of the questions. The human evaluations showed a high passing rate for responses across all dimensions, with a passing score defined as 3 or above on the ordinal scale.

Additionally, we examined whether Ask NAEP generates higher quality answers on specific topics. At this stage, Ask NAEP performs best on questions related to NAEP Data Analysis and Statistical Techniques, Sample Design and Methodology, Scores and Achievement Levels, and the Scoring and Assessment Process. However, the results indicate a need for improvement in addressing questions related to NAEP Validity and Reliability, a finding that aligns with the ongoing efforts to integrate NAEP-published data. This insight has guided the team on which additional documents and data should be integrated in future phases.

Challenges and Limitations

Implementing the Ask NAEP chatbot revealed to us some of the challenges and limitations associated with this type of application. Developing the chatbot involved scraping and storing a large amount of web articles and PDF documents. Dynamic websites, which require human interaction to reveal certain content, proved particularly difficult to scrape. This prompted us to look for other resources for the same information, such as using APIs for state and district profiles to collect summary texts. Another challenge was managing and storing a large amount of unstructured text information, for which vector stores are currently the state-of-the art solution.

Sometimes, a user may ask about a specific NAEP assessment year. Through experimentation, we found that intercepting the user’s query and parsing it to identify the requested year provides better responses. However, we are still working on the ongoing challenge of ensuring that the most up-to-date content is retrieved when the user does not specify a particular year.

Opportunities and Future Directions

Ask NAEP is a proof-of-concept tool that we developed for NCES, with the intention of expanding it to include a larger corpus, such as NCES’s entire website, to meet the broader demands of NCES data users.

Our ongoing efforts involve integrating NAEP-published data (e.g., NAEP summary data tables) and PDFs (such as white papers and methodology reports). However, in this evaluation round, we focused solely on the knowledge base derived from HTML content and state and district data APIs, which we acknowledge as a limitation of this chatbot version.

Future directions include conducting user testing with a more diverse group of stakeholders. For example, although our "Communication" criteria have largely been reviewed by researchers with advanced degrees, most of the responses might still be too technical to be understood by the general public, based on their readability scores. This process would help us better understand the kinds of questions these user groups might ask and give us time to ensure that the responses to the most common questions are consistently accurate.

Another avenue we would like to explore is evaluating other LLMs to see how they perform on the same tasks. As mentioned above, we limited the selection of LLMs to GPT-3.5 and GPT-4o for the convenience of conducting human evaluations. However, there are other models trained on different datasets and using different training procedures. Without testing these models on our knowledge base, it would be difficult to compare their performance with Ask NAEP. Therefore, we plan to conduct similar experiments with other LLMs, such as Claude (Anthropic) and PPLX (Perplexity).

Declarations

Gen-AI Use: The authors of this article declare (Declaration Form #: 2611241800) that Gen-AI tools have NOT been used in any capacity for content creation in this work.

Conflict of Interest: None of the authors have any competing interests that would be interpreted as influencing the research.

Funding: This project has been funded at least in part with Federal funds from the U.S. Department of Education under contract numbers 91990022C0053 and 91990023D0006/91990023F0350. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Acknowledgments: We would like to express our gratitude to Joseph Wilson for his review of the manuscript and constructive feedback. We also thank Jillian Harrison and Martin Hahn for their assistance with editing and formatting.

References

Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical Metrics Used to Evaluate Health Care Chatbots: Scoping Review. Journal of Medical Internet Research, 22(6), e18301. https://doi.org/10.2196/18301

Abeysinghe, B., & Circi, R. (2024, June 13). The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches. The First Workshop on Large Language Models for Evaluation in Information Retrieval, Washington D.C. https://doi.org/10.48550/arXiv.2406.03339

Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., … Vinyals, O. (2024). Gemini: A Family of Highly Capable Multimodal Models (arXiv:2312.11805). arXiv. https://doi.org/10.48550/arXiv.2312.11805

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). Language Models are Few-Shot Learners. arXiv:2005.14165 [Cs]. http://arxiv.org/abs/2005.14165

Celikyilmaz, A., Clark, E., & Gao, J. (2021). Evaluation of Text Generation: A Survey (arXiv:2006.14799). arXiv. http://arxiv.org/abs/2006.14799

Chan, C.-M., Chen, W., Su, Y., Yu, J., Xue, W., Zhang, S., Fu, J., & Liu, Z. (2023). ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate (arXiv:2308.07201). arXiv. http://arxiv.org/abs/2308.07201

Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and psychological measurement, 20(1), 37-46.

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A., & Weston, J. (2023). Chain-of-Verification Reduces Hallucination in Large Language Models (arXiv:2309.11495). arXiv. https://doi.org/10.48550/arXiv.2309.11495

Fu, J., Ng, S.-K., Jiang, Z., & Liu, P. (2023). GPTScore: Evaluate as You Desire (arXiv:2302.04166). arXiv. http://arxiv.org/abs/2302.04166

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey (arXiv:2312.10997). arXiv. https://doi.org/10.48550/arXiv.2312.10997

Gehrmann, S., Clark, E., & Sellam, T. (2023). Repairing the Cracked Foundation: A Survey of Obstacles in Evaluation Practices for Generated Text. Journal of Artificial Intelligence Research, 77, 103–166. https://doi.org/10.1613/jair.1.13715

GitHub. (2024a). Scrapy. GitHub. https://github.com/scrapy/scrapy

GitHub. (2024b). Selenium. GitHub. https://github.com/SeleniumHQ/selenium

Grice, P. (1989). In the way of words. London: Harward University Press.

HuggingFace. (2024). cross-encoder/ms-marco-MiniLM-L-6-v2. HuggingFace. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-6-v2

Howcroft, D. M., & Rieser, V. (2021). What happens if you treat ordinal ratings as interval data? Human evaluations in NLP are even more under-powered than you think. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (pp. 8932–8939). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.703

Iskender, N., Polzehl, T., & Möller, S. (2021). Reliability of Human Evaluation for Text Summarization: Lessons Learned and Challenges Ahead. In A. Belz, S. Agarwal, Y. Graham, E. Reiter, & A. Shimorina (Eds.), Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval) (pp. 86–96). Association for Computational Linguistics. https://aclanthology.org/2021.humeval-1.10

Malkov, Y. A., & Yashunin, D. A. (2018). Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs (arXiv:1603.09320). arXiv. https://doi.org/10.48550/arXiv.1603.09320

National Center for Education Statistics. (2024a). NAEP. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/

National Center for Education Statistics. (2024b). The Nation's Report Card. U.S. Department of Education. https://www.nationsreportcard.gov/

National Center for Education Statistics. (2024c). Technical documentation. U.S. Department of Education. https://nces.ed.gov/nationsreportcard/tdw/

Schoch, S., Yang, D., & Ji, Y. (2020). "This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation. In S. Agarwal, O. Dušek, S. Gehrmann, D. Gkatzia, I. Konstas, E. Van Miltenburg, & S. Santhanam (Eds.), Proceedings of the 1st Workshop on Evaluating NLG Evaluation (pp. 10–16). Association for Computational Linguistics. https://aclanthology.org/2020.evalnlgeval-1.2

Sellam, T., Das, D., & Parikh, A. P. (2020). BLEURT: Learning Robust Metrics for Text Generation (arXiv:2004.04696). arXiv. https://doi.org/10.48550/arXiv.2004.04696

Smith, E. M., Hsu, O., Qian, R., Roller, S., Boureau, Y.-L., & Weston, J. (2022). Human Evaluation of Conversations is an Open Problem: Comparing the sensitivity of various methods for evaluating dialogue agents (arXiv:2201.04723). arXiv. http://arxiv.org/abs/2201.04723

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models (arXiv:2302.13971). arXiv. https://doi.org/10.48550/arXiv.2302.13971 van der Lee, C., Gatt, A., van Miltenburg, E., & Krahmer, E. (2021). Human evaluation of automatically generated text: Current trends and best practice guidelines. Computer Speech & Language, 67, 101151. https://doi.org/10.1016/j.csl.2020.101151

van der Lee, C., Gatt, A., Van Miltenburg, E., Wubben, S., & Krahmer, E. (2019). Best practices for the human evaluation of automatically generated text. Proceedings of the 12th International Conference on Natural Language Generation, 355–368.

Wolf, K., Connelly, M., & Komara, A. (2008). A Tale of Two Rubrics: Improving Teaching and Learning Across the Content Areas through Assessment. 8(1).

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT (arXiv:1904.09675). arXiv. https://doi.org/10.48550/arXiv.1904.09675