Trustworthy AI Meets Educational Assessment: Challenges and Opportunities

Sheng Li

University of Virginia
Charlottesville, VA 20903 USA
shengli@virginia.edu

Abstract

Artificial intelligence (AI) has made substantial impacts in numerous fields, including education. Within education, learning and assessment are two key areas. Although many AI techniques have been applied to improve teaching and learning, their potential in educational assessment remains underexplored. This paper explores the intersection of AI and educational assessment and presents a rich landscape of challenges and opportunities, especially in the context of trustworthy AI, including fairness, transparency, accountability, explainability, and robustness. We will begin by outlining the foundations of trustworthy AI and educational assessment. Next, we will delve into the application of trustworthy AI for various assessment tasks, such as test item generation, test design, and automated scoring. In addition, the talk will also discuss how insights from educational measurement theory, such as item response theory (IRT) and validity frameworks, can inform the development and evaluation of trustworthy AI models. These frameworks help ensure that AI systems in education are not only accurate, but also equitable and aligned with educational goals. Finally, we will highlight future research directions, focusing on the integration of ethical AI principles into educational technology and the need for interdisciplinary collaboration to tackle the emerging challenges in this field. The aim is to foster a new generation of AI-powered educational tools that are both innovative and trustworthy, ultimately contributing to a more equitable and more effective educational landscape.

Introduction

The rapid advancement of artificial intelligence (AI) has revolutionized numerous domains, including education. For example, AI technologies have been used to improve traditional teaching methods by enabling personalized learning experiences and providing real-time feedback to students (Luckin and Holmes 2016; Berendt, Littlejohn, and Blakemore 2020; Fitria 2021; Holmes and Tuomi 2022; Jiao, He, and Yao 2023; Harry and Sayudin 2023; Hao et al. 2024). From adaptive learning platforms that tailor instructions to individual student needs (Kabudi, Pappas, and Olsen 2021; Minn 2022), to intelligent tutoring systems (Lin, Huang, and Lu 2023; Mousavinasab et al. 2021) and automated scoring tools (Yan, Rupp, and Foltz 2020; Ercikan and McCaffrey 2022), AI is reshaping education in very diverse ways. These innovations not only improve efficiency and accessibility but also open up new possibilities to create more engaging and effective learning environments. By far, AI techniques have been mainly applied to improve the teaching and learning process in the education realm, and their potential in educational assessment remains underexplored.

Educational assessment and measurement focuses on assessing and quantifying a student's knowledge, skills, learning, and performance (Newton 2007; Bennett 2015; Brennan 2023; Kubiszyn and Borich 2024). It encompasses various methods, from traditional exams to more dynamic approaches such as project-based and formative assessments. However, educational assessment faces many challenges (Jiao and Lissitz 2020; Bulut et al. 2024), including ensuring fairness, maintaining validity and reliability across diverse student populations, and adapting to different learning styles. Additionally, the need for scalable, efficient, and personalized assessments has become more pressing in today's educational landscape, making it crucial to address these issues while maintaining the integrity of the assessment process.

These challenges motivate us to consider the possibility of integrating trustworthy AI to educational assessment. Several key elements of trustworthiness, including fairness, transparency, accountability, explainability, and robustness, are essential in this context (Li 2023; Li et al. 2023; Zhu et al. 2024). We believe that incorporating trustworthy AI will further expedite the adoption of AI technologies and ensure the safe use of AI in educational assessment and measurement. Additionally, insights from the educational measurement theory, such as item response theory (IRT) (Embretson and Reise 2013) and validity frameworks (Moss 1992), would help guide the development and evaluation of trustworthy AI and generative AI models, such as large foundation models. The recognition of these mutual benefits will encourage researchers from both fields to collaborate more extensively and create innovative solutions.

The major contributions of this vision paper are summarized as follows:

- We review the core concepts in trustworthy AI and educational assessment, and discuss the rationality and potential impacts of bridging these two fields.

- We review existing work on using AI to assist educational assessment, such as test item generation, test design, automated scoring, and test security.
- We discuss how to leverage the theory and statistical tools from educational assessment to assist the development of trustworthy AI models, such as creating theory-guided benchmarks.
- We summarize research challenges when bridging these two areas and also point out several potential opportunities for future research.

Background

Educational Assessment and Measurement

Educational assessment and educational measurement are foundational fields in education (Newton 2007; Bennett 2015; Brennan 2023; Kubiszyn and Borich 2024; Ackerman et al. 2024)., focusing on evaluating and understanding students' learning, skills, and abilities. Educational assessment involves the design, implementation, and analysis of various tools and methods, such as tests and quizzes, to gather meaningful insights into student learning outcomes. Educational measurement, on the other hand, emphasizes the development and application of quantitative approaches, especially statistical and psychometric techniques, to ensure the reliability and validity of assessments. Together, educational assessment and measurement play a crucial role in informing teaching practices, guiding policy decisions, and supporting student growth in diverse learning environments.

Trustworthy AI (TAI)

Trustworthy AI (TAI) refers to the development and deployment of artificial intelligence systems that are ethical, reliable, and aligned with human values (Li et al. 2023; Zhu et al. 2024). TAI involves multiple aspects, such as robustness, transparency, fairness, explainability, accountability, and privacy, with the goal of ensuring the safe and responsible use of AI. Trustworthy AI has become a very active research topic in the AI and machine learning communities. Prior research efforts mainly focus on the design and evaluation of generic trustworthy AI algorithms. The development of trustworthy AI methods for educational assessment has not been well studied yet.

Trustworthy AI for Educational Assessment

Trustworthy AI can be potentially used to improve many tasks in educational assessment and measurement. In the following, we briefly discuss some recent efforts and new ideas on TAI for test item generation, computerized adaptive testing, automated scoring, and test security.

TAI for Test Item Generation

Test development has traditionally been a time-consuming and costly process, requiring extensive resources and expertise. However, the integration of AI, particularly generative AI models, offers the potential to significantly streamline and enhance the entire process. For instance, researchers are already employing large language models to generate test items for a variety of assessments (Hommel et al. 2022; Säuberli 2023; Laverghetta Jr and Licato 2023; Mihalache, Popovic, and Muni 2023), such as K-12 education, licensure and certification, and language proficiency. Beyond standard item creation, AI can also be harnessed to produce more innovative and engaging content, such as multimodal items that incorporate multiple forms of media (e.g., images, videos, and animations), and interactive items that require dynamic responses. This transformation promises not only to increase the efficiency but also to improve the quality and variety of assessments, making them more aligned with modern educational needs. In addition, considering trustworthy objectives such as fairness, AI models have the potential to generate items in accessible formats that help students with disabilities.

TAI for Computerized Adaptive Testing

Computerized adaptive testing (CAT) has become a widely adopted method in many testing programs due to its ability to tailor assessments to the individual needs of examinees (Wainer et al. 2000; Van der Linden, Glas et al. 2000). In CAT, items are selected dynamically from a large item pool, with each item being chosen based on the examinee's performance on previous questions. This personalized approach not only improves test efficiency but also enhances the accuracy of ability estimates.

AI techniques present exciting opportunities to further advance CAT by creating more sophisticated adaptive algorithms and optimizing the selection of test items (Pan et al. 2022; Zhuang et al. 2022; Yu et al. 2023; Liu et al. 2024; Zhuang et al. 2024). For instance, AI can be used to enhance item selection strategies by predicting which items will be most informative for each test-taker, thereby improving both the precision of assessments and the test-taking experience. AI could also be leveraged to analyze the coverage and quality of the existing item pool, according to the test standards or blueprints, and then assist in expanding the item pool. By integrating AI into CAT, testing programs could achieve even greater levels of adaptability, fairness, and test efficiency, ultimately pushing the boundaries of what is possible in personalized assessment.

TAI for Automated Scoring

Natural language processing (NLP) tools have been successfully implemented for automated essay scoring in language assessments, demonstrating the potential of AI to evaluate complex, open-ended responses (Dikli 2006; Madnani and Cahill 2018; Ramesh and Sanampudi 2022; Mizumoto and Eguchi 2023). However, the automated scoring of other item types, such as short-answer questions, problem-solving tasks, and performance-based assessments, remains underexplored and represents a promising area for further research and development.

Beyond simply assigning scores, there is an even greater opportunity to use AI for a more in-depth analysis during the scoring process. Rather than just evaluating responses, AI can be leveraged to automatically identify gaps in a student's knowledge or misunderstandings. This diagnostic approach could generate personalized feedback reports for students,

educators, and parents, providing targeted insights into areas that require improvement. By offering detailed reports on a student's strengths and weaknesses, AI-powered scoring systems could support more tailored learning pathways, enabling a more informed and data-driven approach to education. This shift from traditional scoring to diagnostic evaluation would not only enhance the assessment process but also contribute to more effective learning outcomes by bridging the gap between assessment and instruction.

TAI for Psychometrics

Trustworthy AI has significant potential in the analysis of item properties, such as predicting item difficulty, discrimination parameters, and other psychometric attributes (Yeung 2019; Settles, T. LaFlair, and Hagiwara 2020). By leveraging AI models, testing programs can automate and enhance the precision of item parameter estimations, which traditionally require extensive empirical data and human expertise. These models can quickly analyze large datasets, enabling more efficient calibration of test items while improving the reliability and validity of assessments.

In addition to item-level analysis, AI can be applied to better understand and characterize student proficiency from multiple dimensions. This capability aligns with and advances the field of psychometrics, particularly in the context of multi-dimensional item response theory (MIRT) (Reckase 1997). AI-driven approaches can analyze student performance across various cognitive and skill-based factors, offering a more nuanced view of their abilities. For instance, AI could provide insights into how students perform in different subject areas or cognitive domains, capturing their strengths and weaknesses in a multi-faceted manner.

Furthermore, AI's ability to handle complex, multi-dimensional data, including both the student response data and process data, paves the way for the development of more sophisticated models in educational assessment (Jing and Li 2018; Cheng et al. 2019), allowing for deeper personalization in assessments.

TAI for Test Security

AI plays a pivotal role in enhancing test security in educational assessment, by proactively identifying and mitigating threats that compromise the integrity of exams (Zhou and Jiao 2023; Hao and Fauss 2024). Advanced algorithms can monitor the behavior of examinees during remote and in-person exams to detect anomalies such as cheating attempts. AI-driven systems can also analyze large-scale process data to identify patterns of item preknowledge or content exposure, safeguarding test items from being compromised. Additionally, natural language processing (NLP) tools such as large language models could be used to detect plagiarism or unauthorized sharing of test materials online. By employing AI for real-time monitoring, forensic analysis, and predictive threat modeling, testing programs can ensure the validity of assessments and maintain fairness for all examinees.

Educational Assessment for Trustworthy AI

In this section, we briefly discuss how we can potentially leverage the theory and statistical tools from educational assessment to assist the development of trustworthy AI models, such as establishing theory-guided model assessment and creating theory-guided benchmarks.

Assessment-in-the-Loop for TAI Model Training

In the traditional AI development pipeline, training and assessment are typically viewed as two separate, consecutive stages: first, models are trained, and then they are evaluated for performance. However, by drawing inspiration from educational assessment theory, particularly the way humans gradually learn and master new concepts, we propose a more integrated approach to AI development called the "assessment-in-the-loop" framework.

In this framework, assessment is not just an afterthought or a final step, but an ongoing, integral part of the training process. Just as in human learning, where continuous feedback and evaluation help identify gaps in understanding and guide further learning, this approach involves performing *formative assessments* throughout the AI model's development. It is worthy noting that, unlike the commonly used evaluations during training stage (e.g., loss and training accuracy), the formative assessments we proposed would involve a more comprehensive evaluation on model's capability. By continuously assessing the model's performance on some carefully designed internal tasks, we can detect knowledge gaps or areas where the model is underperforming and use this information to refine and adjust the training process in real time.

This iterative cycle of training and assessment fosters the creation of more trustworthy AI models by ensuring that weaknesses are addressed throughout development, rather than being identified only at the end. It allows for a more dynamic, adaptive process where the AI models can be improved incrementally, learning from its mistakes in a way that mirrors human cognitive development. This method has the potential to produce more robust, fair, and reliable AI systems, as it integrates the principles of formative assessment directly into the learning loop of the model, fostering continuous improvement and better alignment with real-world performance needs.

Theory-Guided Model Assessment

AI models aim to simulate human behavior in various contexts, yet the methods currently available for evaluating the performance and behavior of these models remain relatively limited in scope. While AI models have made significant strides in replicating human-like decision-making, language use, and problem-solving, there is still a need for more robust and comprehensive frameworks to assess how well these models perform across different tasks and scenarios.

In contrast, psychometrics has developed a rich body of theoretical work focused on evaluating human behavior, cognition, and performance. These established theories, such as those used to measure cognitive abilities, personality traits, and learning outcomes, offer a suite of insights that could be potentially adapted to assess the behavior of AI models. We will leverage psychometric principles to evaluate AI systems, which complement the existing evaluation protocols and metrics.

For instance, psychometric models that assess multidimensional human traits could be adapted to evaluate how AI systems balance various performance dimensions, such as precision and generalization. This interdisciplinary approach could help establish new standards for AI evaluation, providing a more rigorous understanding of how AI models perform in real-world applications. Ultimately, integrating psychometric theory into AI assessment could improve the reliability and accountability of AI systems, guiding the development of AI models that better mimic and complement human decision-making.

New Benchmarks for TAI Evaluation

Building upon the principles of educational assessment theory, we aim to develop innovative pipelines and strategies for creating comprehensive benchmarks that can effectively evaluate the trustworthiness of AI models, particularly large foundation models. Traditional evaluation methods are often restricted in capturing the complexity and multifaceted nature of advanced AI systems. By integrating concepts from assessment theory, we can establish new benchmarks that better reflect the diverse capabilities and challenges associated with trustworthy AI.

Furthermore, these benchmarks will encompass a wide range of scenarios and tasks designed to test the AI model's ability to generalize across different contexts, manage uncertainty, and handle edge cases, much like how comprehensive assessments are structured to measure human learning across multiple domains. By employing these educational assessment strategies, we can create benchmarks that not only assess the technical capabilities of AI models but also ensure they align with the ethical, social, and practical requirements of real-world deployment. This approach will push the boundaries of AI evaluation, fostering the development of more reliable and human-centered AI systems.

Discussions and Future Opportunities

In this section, we briefly discuss the future research opportunities when bridging the fields of trustworthy AI and educational assessment.

AI Model Customization

Developing customized AI tools specifically designed for educational assessment and measurement is a crucial step forward, as most existing AI tools function in a broad, general capacity without being finely tuned to the unique needs of education. While these general-purpose models offer impressive capabilities, they often lack the specific focus required to address the complexities in practical assessments. To unlock the full potential of AI in this domain, it is essential to create tailored solutions that cater to the specific goals of educational assessment, learning outcomes, and student progress tracking. Also, customizing large language models (LLMs) and other AI systems for educational assessment and measurement requires a deep and collaborative effort between AI researchers and experts in education.

Responsible Use of AI in Educational Assessment

Ensuring the responsible use of AI in education remains a significant and complex challenge, as AI tools become increasingly integrated into various educational processes such as student learning, classroom instruction, and performance measurement. While these technologies offer the potential to enhance learning experiences and improve educational outcomes, they also bring critical concerns that must be addressed to ensure they are used ethically and responsibly. Theories, algorithms, and tools should be developed to monitor the responsible use of AI technologies in educational assessment.

AI Policy and Governance

The use of AI tools in educational assessment and measurement is rapidly gaining popularity, as these technologies offer innovative ways to evaluate student performance, personalize learning, and streamline administrative tasks. However, with this growing reliance on AI, it is essential to establish clear guidelines and frameworks to ensure the responsible and effective governance of AI tools in the education field. Without such guidelines, there is a risk that AI could be misapplied, leading to unintended consequences such as bias, data privacy concerns, or diminished educational quality.

References

Ackerman, T. A.; Bandalos, D. L.; Briggs, D. C.; Everson, H. T.; Ho, A. D.; Lottridge, S. M.; Madison, M. J.; Sinharay, S.; Rodriguez, M. C.; Russell, M.; et al. 2024. Foundational competencies in educational measurement. *Educational Measurement: Issues and Practice*, 43(3): 7–17.

Bennett, R. E. 2015. The changing nature of educational assessment. *Review of research in education*, 39(1): 370–407.

Berendt, B.; Littlejohn, A.; and Blakemore, M. 2020. AI in education: Learner choice and fundamental rights. *Learning, Media and Technology*, 45(3): 312–324.

Brennan, R. L. 2023. *Educational Measurement*. Rowman & Littlefield.

Bulut, O.; Beiting-Parrish, M.; Casabianca, J. M.; Slater, S. C.; Jiao, H.; Song, D.; Ormerod, C.; Fabiyi, D. G.; Ivan, R.; Walsh, C.; et al. 2024. The Rise of Artificial Intelligence in Educational Measurement: Opportunities and Ethical Challenges. *arXiv preprint arXiv:2406.18900*.

Cheng, S.; Liu, Q.; Chen, E.; Huang, Z.; Huang, Z.; Chen, Y.; Ma, H.; and Hu, G. 2019. Dirt: Deep learning enhanced item response theory for cognitive diagnosis. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 2397–2400.

Dikli, S. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, 5(1).

Embretson, S. E.; and Reise, S. P. 2013. *Item Response Theory*. Psychology Press.

Ercikan, K.; and McCaffrey, D. F. 2022. Optimizing implementation of artificial-intelligence-based automated scoring:

An evidence centered design approach for designing assessments for AI-based scoring. *Journal of Educational Measurement*, 59(3): 272–287.

Fitria, T. N. 2021. Artificial intelligence (AI) in education: Using AI tools for teaching and learning process. In *Prosiding Seminar Nasional & Call for Paper STIE AAS*, volume 4, 134–147.

Hao, J.; and Fauss, M. 2024. Test security in remote testing age: perspectives from process data analytics and AI. *arXiv preprint arXiv:2411.13699*.

Hao, J.; von Davier, A. A.; Yaneva, V.; Lottridge, S.; von Davier, M.; and Harris, D. J. 2024. Transforming assessment: The impacts and implications of large language models and generative ai. *Educational Measurement: Issues and Practice*, 43(2): 16–29.

Harry, A.; and Sayudin, S. 2023. Role of AI in Education. *Interdiciplinary Journal and Hummanity (INJURITY)*, 2(3): 260–268.

Holmes, W.; and Tuomi, I. 2022. State of the art and practice in AI in education. *European Journal of Education*, 57(4): 542–570.

Hommel, B. E.; Wollang, F.-J. M.; Kotova, V.; Zacher, H.; and Schmukle, S. C. 2022. Transformer-based deep neural language modeling for construct-specific automatic item generation. *Psychometrika*, 87(2): 749–772.

Jiao, H.; He, Q.; and Yao, L. 2023. Machine learning and deep learning in assessment. *Psychol Test Assessm*, 65: 179–90.

Jiao, H.; and Lissitz, R. W. 2020. What hath the coronavirus brought to assessment? Unprecedented challenges in educational assessment in 2020 and years to come. *Educational Measurement: Issues and Practice*, 39(3): 45–48.

Jing, S.; and Li, S. 2018. Contextual collaborative filtering for student response prediction in mixed-format tests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Kabudi, T.; Pappas, I.; and Olsen, D. H. 2021. AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, 2: 100017.

Kubiszyn, T.; and Borich, G. D. 2024. *Educational Testing and Measurement*. John Wiley & Sons.

Laverghetta Jr, A.; and Licato, J. 2023. Generating better items for cognitive assessments using large language models. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, 414–428.

Li, B.; Qi, P.; Liu, B.; Di, S.; Liu, J.; Pei, J.; Yi, J.; and Zhou, B. 2023. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9): 1–46.

Li, S. 2023. Towards Trustworthy Representation Learning. In *Proceedings of the SIAM International Conference on Data Mining (SDM)*, 957–960. SIAM.

Lin, C.-C.; Huang, A. Y.; and Lu, O. H. 2023. Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. *Smart Learning Environments*, 10(1): 41.

Liu, Q.; Zhuang, Y.; Bi, H.; Huang, Z.; Huang, W.; Li, J.; Yu, J.; Liu, Z.; Hu, Z.; Hong, Y.; et al. 2024. Survey of Computerized Adaptive Testing: A Machine Learning Perspective. *arXiv preprint arXiv:2404.00712*.

Luckin, R.; and Holmes, W. 2016. Intelligence unleashed: An argument for AI in education.

Madnani, N.; and Cahill, A. 2018. Automated scoring: Beyond natural language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, 1099–1109.

Mihalache, A.; Popovic, M. M.; and Muni, R. H. 2023. Performance of an artificial intelligence chatbot in ophthalmic knowledge assessment. *JAMA ophthalmology*, 141(6): 589–597.

Minn, S. 2022. AI-assisted knowledge assessment techniques for adaptive learning environments. *Computers and Education: Artificial Intelligence*, 3: 100050.

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

Moss, P. A. 1992. Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of educational research*, 62(3): 229–258.

Mousavinasab, E.; Zarifsanaiey, N.; R. Niakan Kalhori, S.; Rakhshan, M.; Keikha, L.; and Ghazi Saeedi, M. 2021. Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. *Interactive Learning Environments*, 29(1): 142–163.

Newton, P. E. 2007. Clarifying the purposes of educational assessment. *Assessment in education*, 14(2): 149–170.

Pan, Y.; Sinharay, S.; Livne, O.; and Wollack, J. A. 2022. A machine learning approach for detecting item compromise and preknowledge in computerized adaptive testing. *Psychological Test and Assessment Modeling*, 64(4): 385–424.

Ramesh, D.; and Sanampudi, S. K. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, 55(3): 2495–2527.

Reckase, M. D. 1997. The past and future of multidimensional item response theory. *Applied Psychological Measurement*, 21(1): 25–36.

Säuberli, A. 2023. *Automatic Generation and Evaluation of Multiple-Choice Reading Comprehension Items with Large Language Models*. Ph.D. thesis, Ph. D. thesis, University of Zurich.

Settles, B.; T. LaFlair, G.; and Hagiwara, M. 2020. Machine learning–driven language assessment. *Transactions of the Association for computational Linguistics*, 8: 247–263.

Van der Linden, W. J.; Glas, C. A.; et al. 2000. *Computerized adaptive testing: Theory and practice*, volume 13. Springer.

Wainer, H.; Dorans, N. J.; Flaugher, R.; Green, B. F.; and Mislevy, R. J. 2000. *Computerized adaptive testing: A primer*. Routledge.

Yan, D.; Rupp, A. A.; and Foltz, P. W. 2020. *Handbook of Automated Scoring: Theory into Practice*. CRC Press.

Yeung, C.-K. 2019. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *arXiv preprint arXiv:1904.11738*.

Yu, J.; Zhenyu, M.; Lei, J.; Yin, L.; Xia, W.; Yu, Y.; and Long, T. 2023. SACAT: Student-Adaptive Computerized Adaptive Testing. In *Proceedings of the Fifth International Conference on Distributed Artificial Intelligence*, 1–7.

Zhou, T.; and Jiao, H. 2023. Exploration of the stacking ensemble machine learning algorithm for cheating detection in large-scale assessment. *Educational and Psychological Measurement*, 83(4): 831–854.

Zhu, R.; Guo, D.; Qi, D.; Chu, Z.; Yu, X.; and Li, S. 2024. A Survey of Trustworthy Representation Learning Across Domains. *ACM Transactions on Knowledge Discovery from Data*.

Zhuang, Y.; Liu, Q.; Huang, Z.; Li, Z.; Shen, S.; and Ma, H. 2022. Fully adaptive framework: Neural computerized adaptive testing for online education. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 4734–4742.

Zhuang, Y.; Liu, Q.; Zhao, G.; Huang, Z.; Huang, W.; Pardos, Z.; Chen, E.; Wu, J.; and Li, X. 2024. A bounded ability estimation for computerized adaptive testing. *Advances in Neural Information Processing Systems*, 36.