Artificial Intelligence in Computerized Adaptive Testing

Dena F. Mujtaba and Nihar R. Mahapatra
Department of Electrical and Computer Engineering, Michigan State University
East Lansing, Michigan, USA
{*mujtabad, nrm*}*@egr.msu.edu*

***Abstract*—Artificial intelligence (AI) is increasingly used to provide customized and efficient e-learning, job search, and career development assistance to students and workers. Both students and jobseekers encounter assessments several times throughout their career and during job searches. Organizations now employ *computerized adaptive testing* (*CAT*), a computer-administered assessment that serves questions based upon the ability of a test taker. CAT aims to provide personalized assessments to test takers to accurately estimate their proficiency with respect to a *latent trait* (e.g., general intelligence and personality characteristic) that is not directly observable. There are several challenges in CAT, such as estimating the latent traits of an individual, generating questions, and question selection. Furthermore, these challenges become more complex as the number of latent trait dimensions being measured increases or if item responses are categorical rather than binary (e.g., using a 1 to 5 scale versus true or false). Traditional approaches employ psychometric and statistical models to make estimations. However, many approaches using machine learning, deep learning, and other AI techniques have emerged to provide better performance. In this paper, we provide a technique-oriented review of AI applications in CAT, and highlight the advantages, limitations, and future challenges in this problem area. We also reconcile different terms and notations used across psychometrics and AI to assist future research and development.**

***Keywords*-artificial intelligence, computational psychometrics, computerized adaptive testing, item response theory, machine learning.**

I. Introduction

A. Motivation

Artificial intelligence (AI) is increasingly used to provide efficient and customized support to students and workers in e-learning, job search, and career development. For instance, recommendation approaches have been developed to help student learners navigate a large number of available learning materials and internships [1], [2]. Also, evolutionary algorithms and machine learning methods have been investigated for generating learning and career pathways [3], [4], [5]. The need for efficient and personalized approaches to learning and job search is growing because of the accelerating impact of technology on jobs, which necessitates more frequent reskilling and changes in jobs and even occupational categories [6], [7].

Assessments provide an important means for facilitating personalization because they are frequently used to characterize individual knowledge, skills, abilities, interests, values, etc. of students and jobseekers. Organizations now employ *computerized adaptive testing* (CAT), a computer-administered assessment that serves questions based upon the ability of a test taker/participant [8]. The responses to these questions are used to estimate proficiency with respect to a *latent trait* dimension, denoted by θ, that is not directly observable (such as general intelligence, knowledge, skills, abilities, and personality characteristics) [8]. Typically, at the start of a CAT, the test taker is provided a question/item of medium difficulty. If it is answered correctly, a more difficult question is provided because the estimated θ of the test taker is seemingly higher. However, if the question is answered incorrectly, the estimated θ of the test taker is lower and an easier prompt is given. Given the customization of each assessment to the test taker, CAT can greatly cut down the time needed to complete a test, provide questions suited to the individual, and enhance precision [8]. Therefore, adaptive testing is widely used in exams, from the Graduate Record Examination to the U.S. Department of Defense's Armed Services Vocational Aptitude Battery test [8]. However, CAT presents new challenges for assessment development, such as over-exposure of items (potentially inflating scores after test takers remember question prompts), estimating the latent traits of an individual, generating and creating the initial item pool, and item selection for the individual test taker. Furthermore, these become more complex as the number of latent trait dimensions being measured increases, or if item responses are categorical rather than binary (e.g., using a 1 to 5 scale versus true or false) [9].

Traditional approaches to CAT have employed psychometric and statistical models to make estimations using combinations of maximum likelihood estimation, Bayesian estimation, and Kullback–Leibler divergence [10], [8]. However, in the past decade, approaches using machine learning, deep learning, and other AI techniques to improve performance of each task in CAT have been studied, and a review of these emerging methods is currently lacking.

B. Related Work and Our Contributions

The main contributions of this paper relative to previous work are twofold: (1) a novel review of emerging AI applications in computerized adaptive testing, and (2) identification of opportunities and challenges in current AI-based CAT approaches. Although past surveys have covered

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00116

CAT methods, studies, and challenges [8], [11], [12], [13], [14], [15], [9], there has not been a comprehensive survey of AI for the tasks in adaptive testing. This includes use of machine learning, deep learning, natural language processing, reinforcement learning, and other AI methods for generating questions, scoring, item selection, and other CAT applications. Therefore, our goal in this paper is to present an overview of traditional and AI-based methods in CAT and reconcile many of the disparate terms and notations used across literature.

The remainder of this paper is organized as follows. In Section II, we provide background in CAT and discuss traditional methods used for CAT as covered in the literature in education and industrial psychology. Next, Section III details the applications of AI (e.g., using machine learning, deep learning, and reinforcement learning) in each stage of CAT. Finally, Section IV provides concluding remarks and summarizes limitations of existing work and future challenges in CAT.

II. Background

CAT research started in the 1970s and has continued to expand over the last few decades. In this section, we cover the steps in CAT and a core component of CAT known as item response theory. These are further described next.

A. Steps in computerized adaptive testing

As delineated by Weiss et al., the main steps in CAT are: (1) development of an item pool (or generation of items), (2) selection of items to provide to the test taker, (3) scoring, or modeling, based upon the participant's response to the question, and (4) a termination criterion (e.g., when all question items in the item pool have been exhausted or a particular score has been achieved) [12]. These steps are shown in Figure 1.

First, an initial item pool with questions to be served to the test taker is established. Though question items are commonly written by domain experts, *automatic item generation* (AIG) seeks to create question prompts similar to those found in other assessments, and has recently relied on natural language processing (NLP) to provide state-of-the-art results [12], [16].

Once the items have been calibrated, or mapped to a latent trait dimension, a selection algorithm is used to determine the next item to provide to the test taker. This relies upon the estimated ability of the test taker after the most recent item was administered and the probability of the next item being answered correctly [8]. The estimated ability of the test taker is inferred using *item response theory* (IRT), which models the probability of a correct response to an item as a function of person parameters (such as math ability) and item parameters (such as difficulty) [10]. Research and design of IRT models are a core component of CAT. These models take the form [17]:

$$\mathbb{P}(U = u|\theta) = f(\theta, \eta, u), \tag{1}$$

where $\mathbb{P}$ is the conditional probability that the score U on the test item will take the value u for a person characterized by a vector of latent trait parameters θ, and f is the function relating θ, a vector η of parameters characterizing the item, and u to a probability in $[0, 1]$ [17]. Selection algorithms use this probable response to choose items. The two most commonly used approaches are maximum information and Bayesian item selection [8]. For instance, the maximum information approach selects the item with the highest mean information. In this approach, the information contributed by the score on item g_k, where $k = \{1, 2, ..., N_j\}$ and N_j is the final test length for test taker j, is given by:

$$I_{g_k}(\theta) = \frac{\mathbb{P}'_{g_k}(\theta)}{\mathbb{P}_{g_k}(\theta)[1 - \mathbb{P}_{g_k}(\theta)]}, \tag{2}$$

where $\mathbb{P}_{g_k}(\theta)$ is the conditional probability for g_k given by the IRT model and $\mathbb{P}'_{g_k}(\theta) = \frac{\partial \mathbb{P}_{g_k}(\theta)}{\partial \theta}$ [8]. The total information function for selection is additive and is given by [8]:

$$I(\theta) = \sum_{k=1}^{N_j} I_{g_k}(\theta). \tag{3}$$

IRT is also used in the next stage of CAT to score the latent trait θ for the test taker using their response to the question item. There are several types of models and scoring methods depending on the item response type. A detailed overview of IRT is provided next.

B. Item response theory

IRT, also known as latent trait theory, is built upon *classical test theory (CTT)*, a psychometric method for scoring [11]. Selecting which IRT model to use depends upon two factors: (1) the type of response to the test items, and (2) the number of latent traits θ being measured. These are further described next.

1) Dichotomous items: IRT models based on *dichotomous items*, such as multiple choice, are given a binary score (e.g., true/false, correct/incorrect) [18]. First described in 1960, the Rasch model, or the 1PL (one-parameter logistic) model, defines the simplest model to summarize a person's ability [18], [10]. The probability that test taker i with trait level θ_i will provide a correct response (i.e., $U_{i,j} = 1$) to the j^{th} item with relative difficulty b_j is given by [10], [17]:

$$\mathbb{P}(U_{i,j} = 1|\theta_i, b_j) = \frac{1}{1 + e^{-(\theta_i - b_j)}}. \tag{4}$$

Moreover, this model has been extended to more commonly used versions in recent research. The 2PL model includes a *discrimination parameter* to adjust the logistic curve of the 1PL model for each item, and more quickly separate lower and higher ability test takers [10]. Furthermore, the

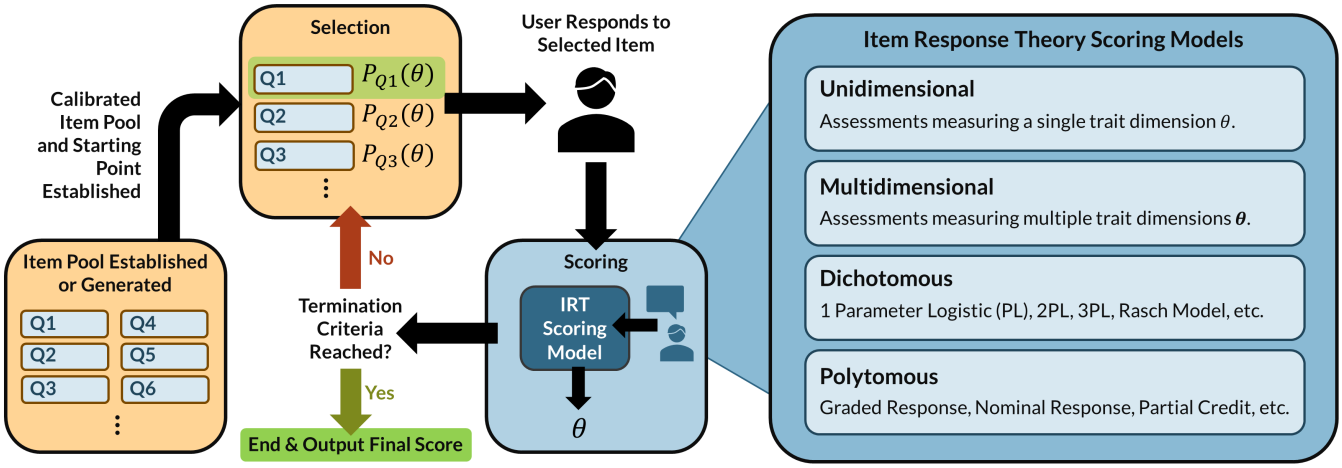

Figure 1. Overview of computerized adaptive testing. First, an item pool is created and calibrated to the desired trait dimension θ. Then, a selection algorithm is applied to determine the assessment question to provide to the user using the probability of a correct/contributing response $P_j(\theta)$, where j is an item. Next, a scoring model is used to calculate the participant's ability in the trait dimension θ. Last, a termination criteria is used to determine if the assessment should prompt the participant for another question or stop (e.g., if the estimated error is low enough for all measured traits or if the entire item pool has been used).

3PL model adds a *pseudo-guessing parameter* that defines the asymptotic minimum of the logistic curve. The 2PL and 3PL models are given by [10]:

$$\mathbb{P}(U_{i,j} = 1|\theta_i, \boldsymbol{d}_j) = c_j + \frac{1 - c_j}{1 + e^{-a_j\theta_i - b_j}}, \tag{5}$$

where a_j is the item discrimination parameter, c_j is the pseudo-guessing parameter, $\boldsymbol{d}_j = \{a_j, b_j, c_j\}$, and $c_j = 0$ for the 2PL model. Depending on the assessment and the dataset size, a model will need to be chosen accordingly (e.g., for a smaller dataset, the 3PL model would not be used because it requires more data to accurately infer ability) [10].

2) Polytomous items: In contrast, *polytomous item* models focus on items with categorical responses, such as a Likert item with a scale of 1 to 5 [19]. The main reason for using a categorical response over binary is the more precise trait estimate that can be obtained from the test taker [19]. Many models were developed to evaluate responses such as the graded response model (GRM), nominal response model, and the partial credit model [19], [20]. These typically build upon dichotomous item models. For instance, GRM expands upon the 2PL model by considering the probability $\mathbb{P}^*_{j_x}(\theta_i)$ that the test taker's response to item j falls at an ordered category x or higher for a given trait level θ_i. This probability is given by [19], [20]:

$$\mathbb{P}^*_{j_x}(\theta_i) = \frac{e^{a_j(\theta_i - t_{j_x})}}{1 + e^{a_j(\theta_i - t_{j_x})}}, \tag{6}$$

where a_j is the item discrimination parameter and t_{j_x} is a threshold parameter that is equal to the latent trait level for which the probability of the test taker responding at or above category x is 50%.

3) Unidimensional versus multidimensional IRT: In *unidimensional IRT*, only one latent trait is estimated. All models discussed so far have been from a unidimensional perspective. However, multiple latent traits are often estimated in exams (e.g., measuring the Big Five personality characteristics, further described by Neito et al. [21]).

In contrast, mutlidimensional IRT is used to measure multiple latent traits in an assessment [9]. This is also referred to as *multidimensional item response theory* (MIRT). MIRT presents new challenges, such as possible correlation between latent traits and the computational difficulty with the increased number of factors in a high dimensional space [22]. One method for MIRT was to build upon unidimensional models and incorporate multiple latent traits. For example, a multidimensional 2PL model is given by,

$$\mathbb{P}(U_{i,j} = 1|\boldsymbol{\theta}_i, \boldsymbol{a}_j, d_j) = \frac{e^{\boldsymbol{a}_j^T\boldsymbol{\theta}_i + d_j}}{1 + e^{\boldsymbol{a}_j^T\boldsymbol{\theta}_i + d_j}} \tag{7}$$

where $\boldsymbol{\theta}_i = [\theta_{i1}, \theta_{i2}, ..., \theta_{iD}]^T$ describes D latent traits, $d_j = -\sum_{k=1}^{D} a_{ik}b_{ik}$, and $\boldsymbol{a}_i$ is the vector of D discrimination parameters [17], [23]. To capture dependencies between items, additional models such as the bifactor model were created for both dichotomous and polytomous data [22].

III. Applications of Artificial Intelligence

Applications of AI in CAT have provided more efficient and accurate results in each stage. We survey the new work using AI for one or more stages in CAT next.

A. Item Generation

Automatic item generation (AIG) can be used to provide highly customized and well-suited questions for test takers in CAT. However, this is still an emerging area and there are

many challenges faced by researchers such as generating text with correct syntax and semantics, capturing domain-specific questions for each test, generating according to the reading comprehension and ability of the test taker, and IRT modeling for AI-generated questions [24].

Question answering has been a prominent challenge in natural language processing, with many benchmark datasets to compare model performance [25]. However, work focusing on AIG is limited, and requires human evaluation and annotation. A survey presented by Pan et al. further covers the current question generation benchmarks and evaluation methods, though from a general perspective [25].

AIG methods have shown promising results using deep neural networks and natural language processing. The approach by Settles et al. uses machine learning and natural language processing to create linguistic exams to measure English language proficiency [26]. Though not generating questions directly, a vocabulary bank was used to create a large dataset of items and to later score item difficulty [26]. In addition, a template-based approach named Job2Questions was also proposed by Shi et al. that generates questions to prompt jobseekers based on a job ad using a deep neural network and pre-defined question structures [27]. Furthermore, the approach proposed by Davier et al. uses a recurrent neural network for AIG [16].

However, another challenge after question generation is modeling the test taker's response and question difficulty for scoring of the exam. The approach presented by Benedetto et al. uses a term frequency-inverse document frequency (TF-IDF) approach to assist IRT modeling of generated questions in this task [28], [29]. The paper by Gier et al. [30] further discusses challenges in AIG and modeling latent abilities with generated questions.

B. Item Selection

Item selection in CAT uses the current estimated latent ability of the test taker and selects the next item to serve that the test taker will likely be able to respond to [31]. Though many items use IRT to predict the test taker's response, we first focus on the selection method itself.

Initial approaches to item selection have used optimization and evolutionary algorithms such as genetic algorithms. In the approach by Phankokkruan et al., a genetic algorithm was used with IRT parameters as the fitness function to generate a decision tree to model item selection [32].

Furthermore, recommendation-based methods, such as collaborative-filtering [33], contextual-bandits [34], and a Markov decision framework [35], have been developed in the past for recommending questions to participants.

Methods relying on reinforcement learning and deep learning have also been developed in the last five years. The approach by Li et al. uses deep Q-learning to estimate the latent abilities of the test taker and provide item recommendations [17]. One challenge in providing recommendations is the possibly long pathway of item transitions to gain an estimate of the test taker's latent ability. Therefore, they use a transition model estimator where, given a state and action, it will output the next probable state to capture the learning behavior of the test taker [17]. Reinforcement learning was also used by Reddy et al. [36] and Tang et al. [37] to model student learning for content-selection and recommendation strategies.

C. Scoring and Item Response Theory

Many methods using AI, such as neural networks and reinforcement learning, have been proposed in the area of item response theory.

The field of *knowledge tracing*, or modeling the ability of a student over time, is a common application and is essential for improving student learning. Recently the area of *deep knowledge tracing* has employed neural networks to model student learning [38]. A study by Khajah et al. compares deep knowledge tracing to past Bayesian models, finding that the deep learning approach provides several advantages [38]. Another model, Deep-IRT, also addresses knowledge tracing using a deep key-value memory network architecture which uses an attention mechanism to model the student's latent ability [39]. The Deep-IRT model has also been compared to standard IRT models, with the authors finding that traditional methods do not consider student learning during the examination [39]. The approach proposed by Benedetto et al. named R2DE (Regressor for Difficulty and Discrimination Estimation) addresses knowledge tracing in the context of newly generated questions [28]. R2DE uses natural language processing with the generated question text to assess the difficulty of the question that is then used in a 2PL dichotomous model for modeling student latent ability [28]. Another work by Benedetto et al. named text2props is a framework to calibrate multiple choice question items for IRT, using natural language processing and TF-IDF for information retrieval from generated questions [29]. Generated multiple-choice questions have additional challenges when estimating their difficulty, due to the number and order of prompts changing difficulty for test takers. Another approach by Xue et al. uses an ELMo (Embeddings from Language Models) and BiLSTM (Bidirectional Long Short-Term Memory) model to predict the p-value, or difficulty of the item as defined by the proportion of test takers answering the question correctly [40]. Overall, the advantage of deep knowledge tracing stems from its ability to model latent student abilities and patterns for learning [38]. However, a limitation in this approach is the lack of interpretability of neural network models and the need for a large training dataset [38].

Other approaches have also used AI to replace the general IRT model itself, rather than building upon it. The approach developed by Li et al. uses a deep reinforcement learning framework (deep Q-learning) to find the optimal learning

policy to select items for the test taker [17]. Another method by Cheng et al. named DIRT (deep item response theory) develops a model for estimating student latent trait ability [41]. The model uses a proficiency vector of knowledge areas for the test taker and a deep neural network is used to predict the latent trait, discrimination factor, and difficulty for IRT [41]. Difficulty was also predicted using a CNN-based neural network in the approach by Sekiya et al. that focuses on mathematical puzzle difficulty [42]. Furthermore, an LSTM-based model by Uto et al. addresses IRT with short-answer questions [43]. These methods can potentially outperform past IRT models and greatly assist future student learning.

IV. Conclusion

In this paper, we survey recent advances in computerized adaptive testing using AI. Many approaches to IRT, AIG, and item selection have included deep learning, natural language processing, and reinforcement learning to provide more custom and accurate CAT. However, there are still a number of future challenges for the use of AI in CAT. First, studying possible algorithmic bias in questions that could influence the responses and inference of the AI system is necessary. Past studies have found AI bias in natural language text and features provided to the model [44]. Second, real-world study of the proposed AI-based methods is needed to assess its performance in a CAT setting. Last, we notice that many of the IRT models have not considered item dependence and MIRT. Therefore, more studies on how AI can be used to measure/model multiple latent traits would assist many MIRT applications. Furthermore, this will influence item selection since recommendation of questions also considers the order and engagement with each item [45]. Overall, work in this area can assist many employers, jobseekers, and students by providing more accurate and efficient assessments.

Acknowledgment

This material is based upon work partly supported by the U.S. National Science Foundation under Grant No. 1936857.

References

[1] M. d. O. C. Machado, N. F. S. Bravo, A. F. Martins, H. S. Bernardino, E. Barrere, and J. F. de Souza, "Metaheuristic-based adaptive curriculum sequencing approaches: a systematic review and mapping of the literature," *Artificial Intelligence Review*, pp. 1–44, 2020.

[2] E. Kurilovas, I. Zilinskiene, and V. Dagiene, "Recommending suitable learning scenarios according to learners' preferences: An improved swarm based approach," *Computers in Human Behavior*, vol. 30, pp. 550–557, 2014.

[3] Q. Meng, H. Zhu, K. Xiao, L. Zhang, and H. Xiong, "A hierarchical career-path-aware neural network for job mobility prediction," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 14–24.

[4] L. Li, H. Jing, H. Tong, J. Yang, Q. He, and B.-C. Chen, "NEMO: Next career move prediction with contextual embedding," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 505–513.

[5] V. Vanitha, P. Krishnan, and R. Elakkiya, "Collaborative optimization algorithm for learning path construction in e-learning," *Computers & Electrical Engineering*, vol. 77, pp. 325–338, 2019.

[6] "Employee tenure in 2020," https://www.bls.gov/news.release/tenure.nr0.htm.

[7] J. Manyika, S. Lund, M. Chui, J. Bughin, J. Woetzel, P. Batra, R. Ko, and S. Sanghvi, "Jobs lost, jobs gained: Workforce transitions in a time of automation," *McKinsey Global Institute*, 2017.

[8] R. R. Meijer and M. L. Nering, "Computerized adaptive testing: Overview and introduction," 1999.

[9] R. P. McDonald, "A basis for multidimensional item response theory," *Applied Psychological Measurement*, vol. 24, no. 2, pp. 99–114, 2000.

[10] M. Wu, R. L. Davis, B. W. Domingue, C. Piech, and N. Goodman, "Variational item response theory: Fast, accurate, and expressive," *arXiv preprint arXiv:2002.00276*, 2020.

[11] L. Cai, K. Choi, M. Hansen, and L. Harrell, "Item response theory," *Annual Review of Statistics and Its Application*, vol. 3, pp. 297–321, 2016.

[12] D. J. Weiss and G. G. Kingsbury, "Application of computerized adaptive testing to educational problems," *Journal of Educational Measurement*, vol. 21, no. 4, pp. 361–375, 1984.

[13] J. M. Linacre *et al.*, "Computer-adaptive testing: A methodology whose time has come," MESA memorandum, Tech. Rep., 2000.

[14] G. Thompson, "Computer adaptive testing, big data and algorithmic approaches to education," *British Journal of Sociology of Education*, vol. 38, no. 6, pp. 827–840, 2017.

[15] Y. Sheng and C. K. Wikle, "Comparing multiunidimensional and unidimensional item response theory models," *Educational and Psychological Measurement*, vol. 67, no. 6, pp. 899–919, 2007.

[16] M. von Davier, "Automated item generation with recurrent neural networks," *psychometrika*, vol. 83, no. 4, pp. 847–857, 2018.

[17] X. Li, H. Xu, J. Zhang, and H.-h. Chang, "Deep reinforcement learning for adaptive learning systems," *arXiv preprint arXiv:2004.08410*, 2020.

[18] G. Rasch, "Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests." 1960.

[19] R. Ostini and M. L. Nering, *Polytomous item response theory models*. Sage, 2006, no. 144.

[20] C. Zanon, C. S. Hutz, H. H. Yoo, and R. K. Hambleton, "An application of item response theory to psychological test development," *Psicologia: Reflexão e Crítica*, vol. 29, no. 1, pp. 1–10, 2016.

[21] M. D. Nieto, F. J. Abad, A. Hernández-Camacho, L. E. Garrido, J. R. Barrada, D. Aguado, and J. Olea, "Calibrating a new item pool to adaptively assess the big five," *Psicothema*, vol. 29, no. 3, pp. 390–395, 2017.

[22] R. P. Chalmers *et al.*, "mirt: A multidimensional item response theory package for the R environment," *Journal of Statistical Software*, vol. 48, no. 6, pp. 1–29, 2012.

[23] R. L. McKinley and M. D. Reckase, "An extension of the two-parameter logistic model to the multidimensional latent space." 1983.

[24] J. Choi, "Automatic item generation with machine learning techniques," *Application of Artificial Intelligence to Assessment*, p. 189, 2020.

[25] L. Pan, W. Lei, T.-S. Chua, and M.-Y. Kan, "Recent advances in neural question generation," *arXiv preprint arXiv:1905.08949*, 2019.

[26] B. Settles, G. T. LaFlair, and M. Hagiwara, "Machine learning–driven language assessment," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 247–263, 2020.

[27] B. Shi, S. Li, J. Yang, M. E. Kazdagli, and Q. He, "Learning to ask screening questions for job postings," *arXiv preprint arXiv:2004.14969*, 2020.

[28] L. Benedetto, A. Cappelli, R. Turrin, and P. Cremonesi, "R2DE: a NLP approach to estimating IRT parameters of newly generated questions," in *Proceedings of the Tenth International Conference on Learning Analytics & Knowledge*, 2020, pp. 412–421.

[29] ——, "Introducing a framework to assess newly created questions with natural language processing," *arXiv preprint arXiv:2004.13530*, 2020.

[30] M. J. Gierl and H. Lai, "The role of item models in automatic item generation," *International journal of testing*, vol. 12, no. 3, pp. 273–298, 2012.

[31] K. C. T. Han, "Components of the item selection algorithm in computerized adaptive testing," *Journal of Educational Evaluation for Health Professions*, vol. 15, 2018.

[32] M. Phankokkruad and K. Woraratpanya, "An automated decision system for computer adaptive testing using genetic algorithms," in *2008 Ninth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing*. IEEE, 2008, pp. 655–660.

[33] J. Xiao, "Collaborative filtering item selection methods for on-the-fly assembled multistage adaptive testing," 2019.

[34] A. S. Lan and R. G. Baraniuk, "A contextual bandits framework for personalized learning action selection." in *EDM*, 2016, pp. 424–429.

[35] Y. Chen, X. Li, J. Liu, and Z. Ying, "Recommendation system for adaptive learning," *Applied psychological measurement*, vol. 42, no. 1, pp. 24–41, 2018.

[36] S. Reddy, S. Levine, and A. Dragan, "Accelerating human learning with deep reinforcement learning," in *NIPS workshop: teaching machines, robots, and humans*, 2017.

[37] X. Tang, Y. Chen, X. Li, J. Liu, and Z. Ying, "A reinforcement learning approach to personalized learning recommendation systems," *British Journal of Mathematical and Statistical Psychology*, vol. 72, no. 1, pp. 108–135, 2019.

[38] M. Khajah, R. V. Lindsey, and M. C. Mozer, "How deep is knowledge tracing?" *arXiv preprint arXiv:1604.02416*, 2016.

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

[40] K. Xue, V. Yaneva, C. Runyon, and P. Baldwin, "Predicting the difficulty and response time of multiple choice questions using transfer learning," in *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, pp. 193–197.

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

[42] R. Sekiya, S. Oyama, and M. Kurihara, "User-adaptive preparation of mathematical puzzles using item response theory and deep learning," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2019, pp. 530–537.

[43] M. Uto and Y. Uchida, "Automated short-answer grading using deep neural networks and item response theory," in *International Conference on Artificial Intelligence in Education*. Springer, 2020, pp. 334–339.

[44] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[45] K. Kenthapadi, B. Le, and G. Venkataraman, "Personalized job recommendation system at LinkedIn: Practical challenges and lessons learned," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 346–347.