TYPE Mini Review
PUBLISHED 10 May 2023
DOI 10.3389/feduc.2023.858273

Check for updates

OPEN ACCESS

EDITED BY
April Lynne Zenisky,
University of Massachusetts Amherst,
United States

REVIEWED BY
Mark Gierl,
University of Alberta, Canada

*CORRESPONDENCE
Ruhan Circi
✉ rcirci@air.org

RECEIVED 19 January 2022
ACCEPTED 06 April 2023
PUBLISHED 10 May 2023

CITATION
Circi R, Hicks J and Sikali E (2023) Automatic item generation: foundations and machine learning-based approaches for assessments. *Front. Educ.* 8:858273.
doi: 10.3389/feduc.2023.858273

COPYRIGHT
© 2023 Circi, Hicks and Sikali. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Automatic item generation: foundations and machine learning-based approaches for assessments

Ruhan Circi[1]*, Juanita Hicks[1] and Emmanuel Sikali[2]

[1]American Institutes for Research, Arlington, VA, United States, [2]National Center for Education Statistics, Washington, DC, United States

This mini review summarizes the current state of knowledge about automatic item generation in the context of educational assessment and discusses key points in the item generation pipeline. Assessment is critical in all learning systems and digitalized assessments have shown significant growth over the last decade. This leads to an urgent need to generate more items in a fast and efficient manner. Continuous improvements in computational power and advancements in methodological approaches, specifically in the field of natural language processing, provide new opportunities as well as new challenges in automatic generation of items for educational assessment. This mini review asserts the need for more work across a wide variety of areas for the scaled implementation of AIG.

KEYWORDS

digital assessments, automatic item generation, item models, machine learning approaches, NLP

Introduction

Due to the increase in large-scale summative (e.g., national assessments, re/certification assessments) and formative assessments (e.g., practice tests/assignments, feedback preparation), items need to be created at a higher pace than ever before to keep up with continuous testing (Attali, 2018; Kurdi et al., 2019). This new era of continuous testing presents a challenge to the traditional methods of item creation and item stability, as it is labor intensive and costly to create items *individually*, and it is difficult to keep a "healthy" item bank where items are not overexposed, especially for computer adaptive testing. In addition to traditional items and methods of item development, more attention has been given to innovative item types, which are needed to measure newer skills that have emerged in the 21st century (e.g., collaborative skills). These innovative and interactive items are even *more* labor intensive and costly to create.

A potential solution to these issues is to generate items automatically. Automatic/automated item generation (AIG) and automated question generation (AQG) are used synonymously to broadly refer to the process of generating items/questions from various inputs, including models, templates, or schemas. In research papers, the terms AIG and AQG are used interchangeably. However, automatic item generation is mostly used in the education domain; therefore, AIG will be used in the continuation of this review.

Historically, AIG was first described by John Bormuth in the 1960's (Bormuth, 1969) but was not developed until much later. Through the years, item generation techniques evolved from using traditional instructional objectives to semi-automated means (Roid and Haladyna, 1978). In 2006, Drasgow et al. (2006) established the base for the theoretical framework and methods of AIG that are widely used in education today. By 2012, with the increase in the number of assessments and the increase in software and computer resources, AIG had become a unique research area with rapid and continuing growth. At this time, there was enough research to provide analysis of both the theoretical concepts and practical applications of AIG (Gierl and Haladyna, 2012). For the last two decades, research on AIG has addressed the current challenges of test/assessment development by generating items on a large scale efficiently (e.g., Gierl et al., 2021).

The main promises of AIG for test/assessment development include: (1) reduced item generation time, (2) reduced cost to create items, (3) support for continuous and rapid item development for large item pools, and (4) support for learning by tailoring items for customized measurement and learning needs. In the educational context, the goal of AIG is defined as creating more items in an efficient and fast manner, such that the items target the same construct but appear unique to test takers (e.g., Pugh et al., 2016). Despite these promises, there is still not enough application of AIG in educational assessment. Therefore, it is critical to understand AIG regarding its feasibility, applicability, and item quality.

The following review covers over 40 papers, two multimedia sources, and one systematic literature review published in the field of automated/automatic item generation for educational purposes. Three data bases (i.e., ACM, IEEE, ERIC), google scholar, AERA and NCME programs, and google searches with selected key words (e.g., "automated item generation," "automated question generation," "machine learning and item generation") were used to extract foundational studies and the most recent work related to AIG to shed light on the most recent developments in this field. For the purpose of this mini review, we scanned and extracted papers to make a finalized list to perform an in-depth review of each selected paper. Our review focuses on the following key points:

- Purpose of AIG in the reviewed material.
- Type of items generated.
- Input type and approaches to generate items.
- Methods used to evaluate generated items.

The following section summarizes the results of the review with a specific focus on the previous key points to show the diversity of thought regarding the topic of automated/automatic item generation and to highlight areas of improvement.

Purpose of AIG in the reviewed material

With regard to the purpose of AIG in the current review, most studies focus on using AIG for assessment purposes, either for large scale assessments (e.g., Gierl et al., 2008; Pugh et al., 2016; Attali et al., 2022), opinion questions (e.g., Baghaee, 2017), classroom/formative assessment purposes such as exam questions (e.g., Fridenfalk, 2013), and practice questions (e.g., Attali, 2018). There are other studies that also use AIG to focus on generating personality items (von Davier, 2018; Hommel et al., 2022). AIG can also be used to expand past the generation of items to more complex tasks for assessments, such as stories and passages (e.g., Harrison et al., 2021; Attali et al., 2022) which is a critical next step in assessment development (Burke, 2020).

Types of generated items

In this review, it was found that multiple choice items are the most frequently generated type of questions in the large-scale educational assessment context as they are the main item type in most large-scale assessments. In addition to the item stem, distractors can also be generated. While using AIG to generate distractors was a challenge (e.g., Embretson and Kingston, 2018), there have been improvement in the methods used to create efficient distractors for multiple choice items over time (Lai et al., 2016; Gierl et al., 2021). For other types of assessments, open ended factual questions (i.e., who, where, when, etc.) are the most common item type to be generated; in comparison, there are fewer studies that are attempting to create open ended questions (e.g., Fridenfalk, 2013; Zhou and Huang, 2019).

Input type and approaches to generate items

In this review, commonly used input types for item generation can be divided into two groups: (a) structured inputs, such as item model templates (e.g., Gierl et al., 2012, 2016; Latifi et al., 2013; Colvin et al., 2016; Attali, 2018; Blum and Holling, 2018), and (b) unstructured inputs such as available written material (e.g., Khodeir et al., 2018; Wang et al., 2018; von Davier, 2019; Zhou and Huang, 2019; Attali et al., 2022). A third input type can be described as a combination of both structured and unstructured inputs (e.g., Atapattu et al., 2012; Wang et al., 2018).

In the educational context, item models are commonly used (Gierl et al., 2008; Gierl and Lai, 2013; Blum and Holling, 2018; Embretson and Kingston, 2018; Pugh et al., 2020). An item model is defined as "… a template that specifies the features in an item that can be manipulated" (LaDuca et al., 1986; Bejar et al., 2003). There are multiple approaches to generate an item model (Drasgow et al., 2006) and they include: (a) weak theory [e.g., generate item sets that are derived from a parent item but look different from one another (Geerlings et al., 2011)], (b) cognitive theory/strong theory [e.g., systematic variations of the parts in an item supported by an underlying theory (e.g., Gierl and Lai, 2013)], and (c) automatic min-max [e.g., introduction of the construct to be measured into the item development process (Arendasy and Sommer, 2012)]. The most applied approach, in an educational context, for generating item models is the cognitive theory/strong theory approach. The main steps in this approach are (a) highlighting the skills and knowledge required for the problem to be solved, (b) subject matter experts (SME) developing cognitive models, (c) creating item models based on the cognitive models that specify the features that can be manipulated, and (d) finally manipulating item models using computer-based algorithms (e.g., software called IGOR).

The approaches used for question/item generation from available written text have been gradually developed and are more diverse than item models. In the field of educational assessment (i.e., large scale), there is a limited amount of work using available written text for automatic item generation (e.g., Attali et al., 2022). Hence, most of the examples in this review are from different assessment domains such as practice quiz generation, factual question generation (e.g., Fattoh et al., 2015; Baghaee, 2017; Wang et al., 2018; Kumar et al., 2019; Blšták and Rozinajová, 2022), personality item generation (e.g., von Davier, 2018; Hommel et al., 2022).

The approaches include use of machine learning/deep learning architectures (e.g., RNN and variants of RNNs as in Kim et al., 2019), natural language processing (NLP) based models (as in Wang et al., 2018; Blšták and Rozinajová, 2022), and large pre-trained language models (e.g., GTP2, GPT3, BERT as in Attali et al., 2022). The focus of neural/deep networks, some have even integrated Natural Language Processing (NLP) based approaches into their models (e.g., Zhou et al., 2017; Wang et al., 2018), is to train models on large data sets and they have the potential to learn implicit rules from the data itself, e.g., GPT-2 fine-tuned using the International Personality Item Pool in Hommel et al. (2022); free medical articles on GPT-2 in von Davier (2019); Stanford Question Answering Dataset (SQuAD) in Kumar et al. (2019); Dolphin18K in Zhou and Huang (2019); Amazon Question/Answer data set in Baghaee (2017); Wikipedia in Harrison et al. (2021). For example, SQuAD (Rajpurkar et al., 2016) consists of more than 100,000 questions from more than 500 articles. It is critical to note that for the existing data sets which include question-answer pairs, none are tailored for educational assessment subjects. Research in automated item generation for education assessments can benefit from the availability of targeted resources (e.g., science subject specific input data) to take advantage of emerging approaches.

Among others utilizing deep learning technology, sequence-to-sequence models have come a long way from its inception (e.g., Kurdi et al., 2019; Pan et al., 2019), which aims to produce plausible questions with minimal human intervention. From the baseline barebones sequence-to-sequence model Du et al. (2017) come up with a model which uses an encoder to take sentence level and paragraph level information and convert it to hidden vectors. Then the decoder takes the vectors from the encoder and creates hidden vectors to predict the next word. This approach is used to generate questions from text passages to measure reading comprehension. In this work, the authors realized that generated questions also included parts of the answer. To address this issue, Zhao et al. (2018) and Kim et al. (2019) proposed various sequence-to-sequence models that are answer aware (i.e., which takes an answer as additional information) and position aware (i.e., using distance between the context words and the answer). Similarly, Sun et al. (2018) addressed issues of unmatching words and unrelated copied context words using complex model.

Methods utilizing RNNs as sequence-to-sequence models to generate questions from sentences or passages (Du et al., 2017; Kim et al., 2019) are most common. However, RNN models suffer from long context/sequences, that is, performance of the models decreases when they are applied to paragraph level context. Chan and Fan (2019) showed that pre-trained language models can also be efficiently used to generate questions. By altering the architecture of one language model (i.e., BERT) to allow sequential generation of words, the authors demonstrated the ability of those models to produce appropriate questions from a paragraph context.

Methods used to evaluate generated items

Development of test specifications and producing items are the first steps of item development for operational purposes. Traditionally, items go through a very rigorous review process including multiple rounds of review and editing, they are also pilot tested, and their psychometric characteristics are evaluated, which include item difficulty, discrimination, and differential item functioning, for operational use (e.g., Haladyna and Rodriguez, 2013). The main promise of AIG is to reduce one-by-one item production and produce items in large quantities. Yet, another time-consuming part of item generation is reviewing and approving items for use in operational settings. Specifically, evaluation of the psychometric characteristics of automatically generated items is critical for the operational use in large scale educational settings. Therefore, it is important to have a systematic and automated evaluation of items generated by AIG.

The item model approach provides a more comprehensive method for the evaluation of AIG generated items. Those approaches include both qualitative and empirical methods. One qualitative approach is to mix the AIG items, traditional items, and then ask content experts to review the items to examine if the items are differentiable (e.g., Gierl and Lai, 2018; Pugh et al., 2020). Empirical methods include examination of psychometric properties of the items using classical test theory, item response theory measures (e.g., Gierl et al., 2016; Attali, 2018; Blum and Holling, 2018; Embretson and Kingston, 2018), and similarity metrics (e.g., Gierl and Lai, 2013; Latifi et al., 2013). Among these methods, the evaluation of psychometric properties for multiple choice items is more established compared to other item types [e.g., most of the items in Attali (2018); Latifi et al. (2013)].

Neural net/deep learning approaches have more variety in relation to the evaluation of generated items; however, these evaluation procedures are still in their early stages and are less standardized. In addition to human evaluation of generated items, some studies use machine transformation evaluation metrics, such as Bilingual Evaluation Understudy (BLEU) or Metric for Evaluation of Translation with Explicit ORdering (METEOR) and text summary measures such as Recall-Oriented Understudy for Gisting Evaluation (ROUGE) to compare their approach with other AIG methods (e.g., Baghaee, 2017; Wang et al., 2018; Zhou and Huang, 2019). One study by von Davier (2018) used factor analysis, as an evaluation method, to show that dimensionality is the same for generated items. Another study, Fattoh et al. (2015) used confusion matrix measures (i.e., precision, recall, f-measure) to evaluate the prediction of item types (i.e., who, what, when etc. types), not the actual item created using AIG.

Conclusion

In this mini review, we have shown that there are various approaches for item generation for assessment purposes. Similar to Kurdi et al. (2019) our review of the literature suggests that almost all the work conducted using AIG is experimental, not operational. However, it is hard to conclude that AIG is not commonly used in operational settings, as there is limited access to the methods used by testing organizations due to the privacy and confidentiality policies. There are a few well-known testing organizations that mention the use of AIG to produce operational items (e.g., Bo et al., 2020). We observed that most of the work specific to AIG in large scale assessments uses template or rule-based approaches as the primary method for creating item models from which to generate items (e.g., Gierl and Lai, 2018). They provide an advantage of creating items aligned well with the intended constructs. Yet, they are mostly used in subjects (e.g., math, medical assessments) where question types are more conventional (e.g., multiple choice, fill-in-blanks). There are a few exceptions where data driven methods such as deep learning and natural language processing are employed (von Davier, 2019; Burke, 2020; Attali et al., 2022). These models aim to generate questions/items with minimal human involvement, in the realm of neural networks and large pre-trained models. The use of neural-based models dominates state of the art question generation in various domains. In the last few years, pretrained large models significantly increased performance gains on many tasks. Researchers now leverage existing models to generate semantically coherent and fluent questions, which is critical in the context of educational assessments as digitalization in education continues to grow. However, it is important to reiterate that non-template data driven models still have a long way to meet the standards of quality expected in operational testing situations.

Automated/automatic item generation is a process, therefore the steps used in AIG are very important. While terminology is not common across fields or papers, it is helpful to understand two main stages of AIG (different approaches differ in the number stages): (1) the input stage (or encoder), and (2) the transformation stage (generation or decoder). Algorithmic generation approaches lead to scalable item development and produce large numbers of items. However, with the abundance of items, the next challenge becomes differentiation of high quality items. Evaluation of generated items was not provided in all the studies; however, for studies that did provide item evaluation there was high variation in evaluation approaches. Some of the evaluation approaches include: (1) blind review of both AIG items and traditionally developed items by expert panels (e.g., Khodeir et al., 2018; Pugh et al., 2020), (2) factor analysis to examine the internal structure (e.g., von Davier, 2018), (3) comparison of psychometric properties of AIG items with operational items (e.g., Attali, 2018; Embretson and Kingston, 2018), (4) examination of the similarity of generated items (using cosine similarity index, e.g., Gierl and Lai, 2013; Latifi et al., 2013; Kaliski et al., 2020), and (5) comparison of different models with the original text or human judgement using machine translation indices together with or without human evaluation (e.g., Wang et al., 2018; Chan and Fan, 2019; Zhou and Huang, 2019). The variety of AIG evaluation approaches included in the literature suggests that there is a clear need for more research in this area.

Over the years, increasing item demands have led to multiple approaches to automatically generate items. Various researchers and practitioners have helped AIG to be more consistent and established, but at the same time have increased its complexity. All approaches used for AIG still need to be thoroughly tested to become well understood. Thus, this summary review suggests that the topic of automated/automatic item generation is wide and varied with its unique strengths and limitations as an assessment tool.

Author contributions

RC and JH: wrote the original draft, reviewed, and edited. All authors contributed to the article and approved the submitted version.

Funding

Research in this working paper was developed with funding from NCES under Contract No. ED-IES-12-D-0002/0004. The views, thoughts, and opinions expressed in the paper belong solely to the author(s) and do not reflect NCES position or endorsement.

Acknowledgments

We would like to thank our colleagues, Fusun Sahin, Tiago Calico, and Xiaying Zheng, for their contributions to the original literature review referenced in this work. We also would like to thank Emmanuel Sikali, Senior Research Scientist/Mathematical Statistician at NCES, for his support to expand the original work.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arendasy, M., and Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes assessment. *Learn. Individ. Differ.* 22, 112–117. doi: 10.1016/j.lindif.2011.11.005

Atapattu, T., Falkner, K., and Falkner, N. (2012). "Automated extraction of semantic concepts from semi structured data: supporting computer-based education through the analysis of lecture notes," in *Database and Expert Systems Applications. DEXA 2012*. eds. S. W. Liddle, K. D. Schewe, A. M. Tjoa and X. Zhou, Lecture Notes in Computer Science (Berlin, Heidelberg: Springer).

Attali, Y. (2018). "Automatic item generation unleashed: an evaluation of a large-scale deployment of item models," in *Artificial intelligence in education: 19th International Conference*, eds C. P. Rosé, R. Martínez-Maldonado, H. U. Hoppe, R. Luckin, M. Mavrikis, K. Porayska-Pomsta, B. McLaren, and B. du Boulay (Cham: Springer), 17–29.

Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., et al. (2022). The interactive reading task: transformer-based automatic item generation. *Front. Artif. Intell.* 5:903077. doi: 10.3389/frai.2022.903077

Baghaee, T. (2017). Automatic neural question generation using community-based question answering systems. Unpublished master's thesis. University of Lethbridge.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., and Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *J. Technol. Learn. Assess.* 2, 1–32.

Blšták, M., and Rozinajová, V. (2022). Automatic question generation based on sentence structure analysis using machine learning approach. *Nat. Lang. Eng.* 28, 487–517. doi: 10.1017/S1351324921000139

Blum, D., and Holling, H. (2018). Automatic generation of figural analogies with the IMak package. *Front. Psychol.* 9, 1–13. doi: 10.3389/fpsyg.2018.01286

Bo, E., He, W., Javurel, A., Miller, S., Scheuring, M. S., and Simpson, M. A. (2020). Items and item models: AIG traditional and modern. In P. Kaliski (Chair), *Challenges with automatic item generation implementation: Research, strategies, and lessons learned [virtual symposium]. Annual meeting of the National Council on measurement in education.*

Bormuth, J. (1969). *On a Theory of Achievement Test Items*. Chicago, IL: University of Chicago Press.

Burke, A. (Host). (2020). Creating test items with automated item generation: AIG, AIGL, & POE (no.7) [audio podcast episode]. In ACT next navigator. Available at: https://www.youtube.com/watch?v=XUSGAIdfVsg

Chan, Y.-H., and Fan, Y.-C. (2019). A recurrent BERT-based model for question generation. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering. Association for Computational Linguistics*.

Colvin, K. F., Keller, L. A., and Robin, F. (2016). Effect of imprecise parameter estimation on ability estimation in a multistage test in an automatic item generation context. *J. Comput. Adapt. Test.* 4, 1–18. doi: 10.7333/1608-040101

Drasgow, F., Luecht, R. M., and Bennett, R. E. (2006). "Technology and testing," in *Educational Measurement*. ed. R. L. Brennan (Westport, CT: Praeger Publishers), 471–516.

Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. ArXiv:1705.00106 [Cs] [Epub ahead of preprint].

Embretson, S., and Kingston, N. M. (2018). Automatic item generation: a more efficient process for developing mathematics achievement items? *J. Educ. Meas.* 55, 112–131. doi: 10.1111/jedm.12166

Fattoh, I. E., Aboutabl, A. E., and Haggag, M. H. (2015). Semantic question generation using artificial immunity. *Int. J. Mod. Educ. Comput. Sci.* 7, 1–8. doi: 10.5815/ijmecs.2015.01.01

Fridenfalk, M. (2013). "System for automatic generation of examination papers in discrete mathematics," in *Proceedings of IADIS International Conference on e-Learning 2013 IADIS Multi Conference on Computer Science and Information Systems*, 365–368.

Geerlings, H., Glas, C. A. W., and van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika* 76, 337–359. doi: 10.1007/S 11336-011-9204-X

Gierl, M. J., and Haladyna, T. M. (Eds.). (2012). *Automatic Item Generation: Theory and Practice* (*1st*). England: Routledge.

Gierl, M. J., and Lai, H. (2013). Instructional topics in educational measurement (items) module: using automated processes to generate test items. *Educ. Meas. Issues Pract.* 32, 36–50. doi: 10.1111/emip.12018

Gierl, M. J., and Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Appl. Psychol. Meas.* 42, 42–57. doi: 10.1177/0146621617726788

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A.-P., and De Champlain, A. (2016). Evaluating the psychometric characteristics of generated multiple-choice test items. *Appl. Meas. Educ.* 29, 196–210. doi: 10.1080/08957347.2016.1171768

Gierl, M. J., Lai, H., and Tanygin, V. (Eds.). (2021). *Advanced Methods in Automatic Item Generation* (*1st*). England: Routledge.

Gierl, M. J., Lai, H., and Turner, S. R. (2012). Using automatic item generation to create multiple-choice test items. *Med. Educ. J.* 46, 757–765. doi: 10.1111/j.1365-2923.2012.04289.x

Gierl, M. J., Zhou, J., and Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *J. Technol. Learn. Assess.* 7, 1–51.

Haladyna, T., and Rodriguez, M. (2013). "Developing the test item," in *Developing and Validating Test Items*. eds. T. Haladyna and M. Rodriguez (New York, NY: Routledge), 17–27.

Harrison, B., Purdy, C., and Riedl, M. (2021). Toward automated story generation with markov chain Monte Carlo methods and deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, pp. 191–197.

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, 749–772. doi: 10.1007/s11336-021-09823-9

Kaliski, P., Clauser, J., and Burke, M. (2020). Exploring the utility of semantic similarity indices for automated item generation. In P. Kaliski (Chair), *Challenges with automatic item generation implementation: Research, strategies, and lessons learned [virtual symposium]. Annual meeting of the National Council on measurement in education*.

Khodeir, N. A., Elazhary, H., and Wanas, N. (2018). Generating story problems via controlled parameters in a web-based intelligent tutoring system. *Int. J. Inf. Learn. Technol.* 35, 199–216. doi: 10.1108/IJILT-09-2017-0085

Kim, Y., Lee, H., Shin, J., and Jung, K. (2019). Improving neural question generation using answer separation. *Proc. AAAI Conf. Artif. Intell.* 33, 6602–6609. doi: 10.1609/aaai.v33i01.33016602

Kumar, V., Muneeswaran, S., Ramakrisnan, G., and Li, Y.-G.. (2019). ParaQG: a system for generating questions and answers from paragraphs. arXiv [Epub ahead of preprint]: 1–6.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2019). A systematic review of automatic question generation for educational purposes. *Int. J. Artif. Intell. Educ.* 30, 121–204. doi: 10.1007/s40593-019-00186-y

LaDuca, A., Staples, W. I., Templeton, B., and Holzman, G. B. (1986). Item modeling procedures for constructing content-equivalent multiple-choice questions. *Med. Educ.* 20, 53–56. doi: 10.1111/j.1365-2923.1986.tb01042.x

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A., and De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teach. Learn. Med.* 28, 166–173. doi: 10.1080/10401334.2016.1146608

Latifi, S., Gierl, M. J., Lai, H., and Fung, K. (2013) Establishing item uniqueness for automatic item generation [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Pan, L., Lei, W., Chua, T. S., and Kan, M. Y. (2019). Recent advances in neural question generation. arXiv [Epub ahead of preprint].

Pugh, D., De Champlain, A., Gierl, M. J., Lai, H., and Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Med. Teach.* 38, 838–843. doi: 10.3109/0142159X.2016.1150989

Pugh, D., De Champlain, A., Gierl, M. J., Lai, H., and Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Res. Pract. Technol. Enhanc. Learn.* 15, 1–13. doi: 10.1186/s41039-020-00134-8

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. arXiv [Epub ahead of preprint].

Roid, G. H., and Haladyna, T. M. (1978). A comparison of objective-based and modified-Bormuth item writing techniques. *Educ. Psychol. Meas.* 38, 19–28. doi: 10.1177/001316447803800104

Sun, X., Liu, J., Lyu, Y., He, W., Ma, Y., and Wang, S. (2018). Answer-focused and position-aware neural question generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3930–3939.

von Davier, M. (2018). Automated item generation with recurrent neural networks. *Psychometrika* 83, 847–857. doi: 10.1007/s11336-018-9608-y

von Davier, M. (2019). Training optimus prime, M.D.: generating medical certification items by fine tuning OpenAI's gpt2 transformer model. arXiv [Epub ahead of preprint], pp. 1–19.

Wang, Z., Lan, A. S., Nie, W., Waters, A. E., Grimaldi, P. J., and Baraniuk, R. G. (2018). QG net: a data driven question generation model for educational content. In *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pp. 1–10.

Wang, H., Zhang, X., Ma, S., Sun, X., Wang, H., and Wang, M. (2018). A neural question answering model based on semi-structured tables. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1941–1951.

Zhao, Y., Ni, X., Ding, Y., and Ke, Q. (2018). Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 3901–3910.

Zhou, Q., and Huang, D. (2019). Towards generating math word problems from equations and topics. In *Proceedings of the 12th international conference on natural language generation*, pp. 494–503.

Zhou, Q., Yang, N., Wei, F., Tan, C., Bao, H., and Zhou, M. (2017). Neural question generation from text: a preliminary study. arXiv [Epub ahead of preprint].