Advances in Health Sciences Education (2022) 27:405–425
https://doi.org/10.1007/s10459-022-10092-z

Feasibility assurance: a review of automatic item generation in medical assessment

Filipe Falcão[1,2] · Patrício Costa[1,2] · José M. Pêgo[1,2]

Received: 17 November 2020 / Accepted: 23 January 2022 / Published online: 1 March 2022

Abstract

Background Current demand for multiple-choice questions (MCQs) in medical assessment is greater than the supply. Consequently, an urgency for new item development methods arises. Automatic Item Generation (AIG) promises to overcome this burden, generating calibrated items based on the work of computer algorithms. Despite the promising scenario, there is still no evidence to encourage a general application of AIG in medical assessment. It is therefore important to evaluate AIG regarding its feasibility, validity and item quality.

Objective Provide a narrative review regarding the feasibility, validity and item quality of AIG in medical assessment.

Methods Electronic databases were searched for peer-reviewed, English language articles published between 2000 and 2021 by means of the terms 'Automatic Item Generation', 'Automated Item Generation', 'AIG', 'medical assessment' and 'medical education'. Reviewers screened 119 records and 13 full texts were checked according to the inclusion criteria. A validity framework was implemented in the included studies to draw conclusions regarding the validity of AIG.

Results A total of 10 articles were included in the review. Synthesized data suggests that AIG is a valid and feasible method capable of generating high-quality items.

Conclusions AIG can solve current problems related to item development. It reveals itself as an auspicious next-generation technique for the future of medical assessment, promising several quality items both quickly and economically.

Keywords Assessment · Automatic item generation · Computer-based testing · Medical Assessment · Multiple-choice questions

✉ Filipe Falcão
jmpego@med.uminho.pt

[1] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Largo do Paço, 4700-000 Braga, Portugal

[2] ICVS/3B's, PT Government Associate Laboratory, Braga/Guimarães, Portugal

Demands for new multiple-choice questions (MCQs) items are now higher than the supply in medical assessment. Traditional item development methods are ponderous and expensive, which builds an urgency for new solutions in item development. Automatic Item Generation (AIG) is a next-generation assessment method that mixes human expertise with computer algorithms, promising to overcome the item development burden. Despite the promising scenario, there is not enough evidence to encourage a general application of AIG in medical assessment. This review aims to summarise the state of the art regarding the feasibility, validity and item quality of AIG.

Literature review

Educational measurement is evolving: The medical education case

Thanks to the unified effort of cognitive sciences, statistical theories of test scores, psychology, technology and computer sciences, educational measurement is evolving. This evolution translates into changes in computer-based testing (CBT), test designs and cognitive diagnostic assessment (Gierl & Haladyna, 2012). Alongside this evolution, the way in which tests are administered today has been adapted to the popularity of digital media and the internet with new types of assessments and resources. Consequently, tests once given in paper are now administered by computers via internet, offering advantages to both students and educators compared with more traditional methods of testing (Gierl & Lai, 2012). Among the advantages, we highlight the possibility of exempting educators from time-consuming tasks associated with paper-based tests; the use of diverse item formats; and the possibility for students to take exams when/where they want, while receiving immediate feedback (Gierl et al., 2016; Kosh et al., 2019).

These advances are evident in medical education, which introduced new strategies to measure complex performances and competencies (David et al., 2001; Gierl & Lai, 2013). Due to the high number of contents medical students must learn and large class sizes, the most frequently used assessment method is written tests with multiple-choice questions (MCQs) (Batalden et al., 2002; Royal et al., 2018). Besides being automatically scored, MCQs can be used in items that comprise different skills in an efficient and economical process marked by reduced human intervention (Batalden et al., 2002; McCoubrie, 2004; Royal et al., 2018).

However, MCQ-based assessment is challenging. The reasons for this challenge relate to the difficulty of the process, the time and money required to develop the items, as well as possible security and validity issues (McCoubrie, 2004). These trials are accompanied by increasing demands for new items, a problem that conventional item development methods seem unable to solve (Gierl et al., 2012a, b; Gierl & Lai, 2013a; Royal et al., 2018). Through the use of computer technology, Automatic Item Generation (AIG) emerges as a major breakthrough in psychometric sciences that holds potential to overcome these limitations (Gierl et al., 2012a, b).

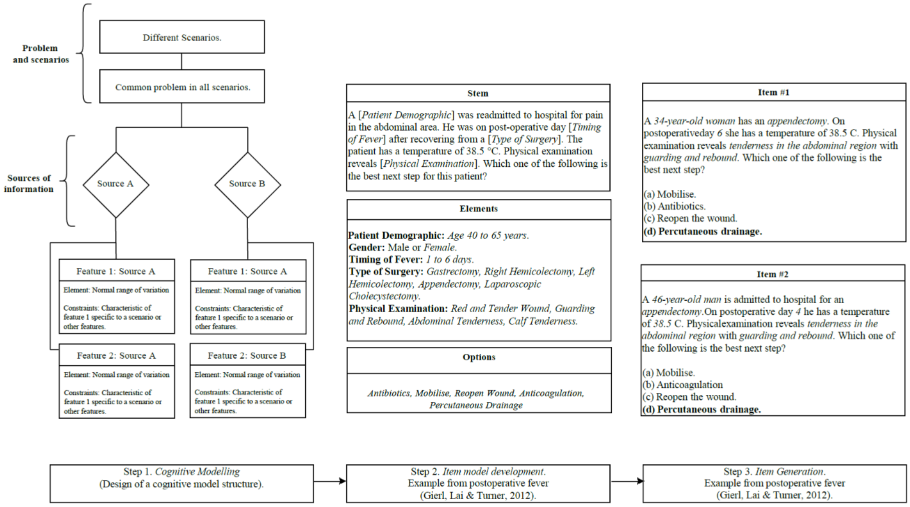
A review of a next-gen assessment theory: Automatic Item Generation

AIG allows content experts the ability to generate a large numbers of test items by integrating domain expertise with computer algorithms (Lai, Gierl, Byrne, et al., 2016). It is a next-generation assessment theory where cognitive models generate statistically calibrated items based on computer modules, producing high-quality and content-specific test items quickly and efficiently (Arendasy & Sommer, 2007; Gierl et al., 2016).

According to Gierl & Lai (2012), the process used to generate medical MCQs based-on AIG runs along three different steps (Cf. Figure 1):

In step one, specialist identify the exam content and outline a framework with the necessary knowledge and skills to formulate a diagnosis, forming a cognitive model structure (Blum & Holling, 2018; Pugh et al., 2016). Within the model, a problem specific to the test item is identified, along with different scenarios and the information needed to solve it. Features within the sources of information are also marked, each one containing elements (variables that can be manipulated to generate new items) and constraints (variables to manipulate in specific scenarios) (Gierl et al., 2012a, b). In step two, the content of the cognitive model is cast into an MCQ, forming an item model. An item model resembles a template highlighting the variables to be manipulated. It is composed by vignettes with the necessary information to answer the item, allowing the definition of the questions and correct options, along with distractors (Gunabushanam et al., 2019; Pugh et al., 2016). Finally, in step three, computer modules work on the item models to generate new items. Within medical assessment, this process can be seen as a method by which the representation of clinical reasoning is used to generate new items based on computer algorithms (Lai et al., 2016a, b).

AIG promises unlimited resources for assessment, exam security and the ease of the item development burden due to the re-use of item models (Cole et al., 2020; Gierl & Lai, 2018; Gierl et al., 2015; von Davier, 2018). Bearing this in mind, and since we continually make


Problem and scenarios
Different Scenarios.
Common problem in all scenarios.
Sources of information
Source A
Source B
Feature 1: Source A
Element: Normal range of variation
Constraints: Characteristic of feature 1 specific to a scenario or other features.
Feature 1: Source A
Element: Normal range of variation
Constraints: Characteristic of feature 1 specific to a scenario or other features.
Feature 2: Source A
Element: Normal range of variation
Constraints: Characteristic of feature 1 specific to a scenario or other features.
Feature 2: Source B
Element: Normal range of variation
Constraints: Characteristic of feature 1 specific to a scenario or other features.
Stem
Elements
Options
Item #1
Item #2
Step 1. Cognitive Modelling
(Design of a cognitive model structure).
Step 2. Item model development.
Example from postoperative fever
(Gierl, Lai & Turner, 2012).
Step 3. Item Generation.
Example from postoperative fever
(Gierl, Lai & Turner, 2012).

Stem

A [*Patient Demographic*] was readmitted to hospital for pain in the abdominal area. He was on post-operative day [*Timing of Fever*] after recovering from a [*Type of Surgery*]. The patient has a temperature of 38.5 °C. Physical examination reveals [*Physical Examination*]. Which one of the following is the best next step for this patient?

Elements

Patient Demographic: *Age 40 to 65 years.*
Gender: Male or *Female.*
Timing of Fever: *1 to 6 days.*
Type of Surgery: *Gastrectomy, Right Hemicolectomy, Left Hemicolectomy, Appendectomy, Laparoscopic Cholecystectomy.*
Physical Examination: *Red and Tender Wound, Guarding and Rebound, Abdominal Tenderness, Calf Tenderness.*

Options

Antibiotics, Mobilise, Reopen Wound, Anticoagulation, Percutaneous Drainage

Item #1

A *34-year-old woman* has an *appendectomy*. On postoperativeday *6* she has a temperature of 38.5 C. Physical examination reveals *tenderness in the abdominal region* with *guarding and rebound*. Which one of the following is the best next step?

(a) Mobilise.
(b) Antibiotics.
(c) Reopen the wound.
(d) Percutaneous drainage.

Item #2

A *46-year-old man* is admitted to hospital for an *appendectomy*.On postoperative day *4* he has a temperature of *38.5* C. Physicalexamination reveals *tenderness in the abdominal region* with *guarding and rebound*. Which one of the following is the best next step?

(a) Mobilise.
(b) Anticoagulation
(c) Reopen the wound.
(d) Percutaneous drainage.

Fig. 1 AIG three-step process for generating medical MCQs

judgements and decisions regarding medical students based on assessment, there is a need to understand AIG's strengths and limitations as an assessment tool (Cook et al., 2015). This need is related to the fact that AIG applications in educational settings have not yet been validated, as well as the psychometric characteristics of the items generated remain unreported (Choi et al., 2018; Gierl et al., 2016).

Considering what was mentioned above, this review aims to evaluate AIG regarding its feasibility, validity and item quality. AIG's feasibility was addressed regarding the time-spent implementing its processes, ease of learning and number of items generated. Item quality was assessed by exploring the various methodologies used in the literature to evaluate the quality of items generated by AIG. Finally, AIG's validity was assessed through the implementation of a validation theory articulated by Kane (2013) to gather evidences of validity.

By synthesizing a range of literature focused on AIG, we hope to give an overview of AIG and obtain evidence that supports its implementation in medical assessment. By demonstrating that AIG is feasible, generates high-quality items and presents validity properties, we expect it to be gradually disseminated by medical schools, promoting more productive and higher quality education.

Method

Narrative reviews are the most common type of article in the medical literature, influencing doctors in clinical practice and research (Baethge et al., 2019). They are adequate for obtaining a broad perspective on a topic since they perform a comprehensive syntheses of the available literature, which is useful to promote discussions with no focused questions and no stated hypothesis (Green et al., 2006). Considering that our interest here is to perform a general evaluation of the feasibility, validity and item quality of AIG in medical assessment, this type of review seems to be the most appropriate to clarify this matter.

Quality assessment of the review

The quality of this review was measured using the Scale for the Assessment of Narrative Review Articles (SANRA) (Baethge et al., 2019). The SANRA is a validated tool that evaluates the quality of non-systematic articles and can be used by authors and journal editors to assess narrative reviews (Baethge et al., 2019). Two reviewers (FF, PC) critically appraised this review. Appraisals were matched and disagreements were discussed. A senior reviewer (JMP) served as a tiebreaker if a consensus was not met.

Sources of information

We conducted a literature search within the PubMed®, Web of Science—Core Collection®, Education Resources Information Center (ERIC®), Taylor & Francis Online® and SCOPUS-ELSEVIER® electronic databases between January and March 2021. The search used terms combined into the following query: 'Automatic Item Generation' OR 'Automated Item Generation' OR 'AIG' AND 'medical assessment' OR 'medical education' from 2000 through March 2021. A total of 152 publications were extracted. 33 duplicated results were eliminated. In total, 119 publications advanced to the screening phase.

Data extraction

Studies found through database search were exported to Rayyan Management Software (Ouzzani et al., 2016). Study selection, data records, search results and eligibility criteria were conducted within the software.

Selection criteria

This review considered studies with quantitative or qualitative designs related to the implementation of AIG in medical assessment. Interventions of interest included the implementation of AIG to generate medical MCQs. Inclusion criteria were: i) studies reporting on the generation of MCQs through AIG; (ii) studies assessing the quality of items generated by AIG; (iii) studies conducted on the medical education area; (iv) studies written in English language.

Exclusion criteria

Studies were excluded if they: i) reported a systematic review or meta-analysis; (ii) represented grey literature; (iii) represented letters, editorials, commentaries, reviews, book chapters or conference papers; or iv) presented incomplete results.

Screening phase

Screening was conducted in two stages: initial screening and full-text screening. In both stages, articles were examined based on the inclusion and exclusion criteria. Initial screening

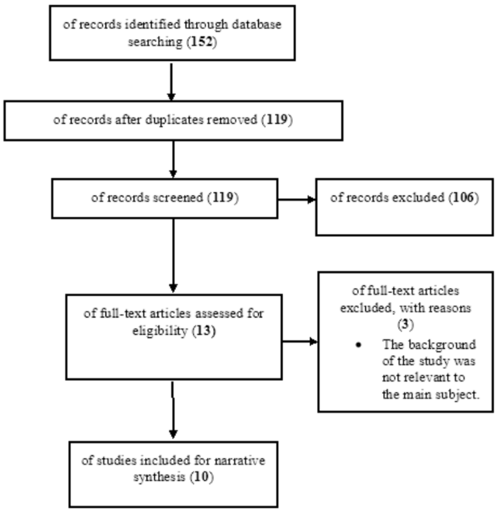

of records identified through database searching (152)
of records after duplicates removed (119)
of records screened (119)
of records excluded (106)
of full-text articles assessed for eligibility (13)
of full-text articles excluded, with reasons (3)
• The background of the study was not relevant to the main subject.
of studies included for narrative synthesis (10)

Fig. 2 Flow chart of the included studies

was conducted by two authors (FF, PC), who examined articles based on title and abstract. A final decision regarding the selected studies was made through discussion between them and the last author (JMP), resulting in a set of 13 publications eligible for the second stage of screening. In the second stage, one reviewer (FF) independently conducted a full-text screening with the papers approved in the first stage. In case of doubt, the reviewer (FF) discussed his decision with the other authors to reach a final decision. A final set of 10 publications was included in this review. The process used for study search and selection is detailed in Fig. 2.

Quality assessment of included studies

To the best of our knowledge, there is no checklist available to evaluate the quality of studies included in a review on this topic. Considering this limitation, we developed a checklist to assess the quality of the papers included in this review inspired by Gierl & Lai's (2012) study. In their work, the authors described seven fundamental topics for AIG: (i) item modelling: definition and related concepts; (ii) developing item models; (iii) item model taxonomy; (iv) using item models to automatically generate items; (v) benefits of item modelling; (vi) item model bank; and (vii) estimation of the statistical characteristics of generated items. In addition to these topics, we have also added the following to get more accurate assessments: (viii) description of the three-step process for conducting AIG; (ix) assessment of AIG's capacity to generate new items (e.g., quantity of items generated, estimated costs and time); (x) quality assessment of generated items, cognitive model and/or item model; (xi) comparison of AIG with traditional methods of item development; and (xii) limitations of AIG.

In this work, two reviewers (FF, PC) judged the quality of the included studies. The more topics the article dealt with, the higher the quality rating assigned. Each paper was rated on a scale from zero ("*topic not covered in the study*.") to two ("*topic was covered in the study*"). Appraisals were matched and disagreements were discussed on a case-by-case basis. Inter-rater reliability was computed and measured through Cohen's kappa coefficient (κ) between the two reviewers. Inter-rater reliability was computed with R (version 4.0.4, 64 bit), R Studio software (version 1.4.1106) and R package 'vcd' (Hornik et al., 2020). κ was found to be substantial (κ=0.80; SE = 0.186; 95% CI = [0.46, 0.85]) (Landis & Koch, 1977). The quality assessment of the included studies is detailed in Table 1. A final decision regarding the inclusion of the studies was made through a discussion among the reviewers. All 10 studies proceeded to data synthesis.

Validity framework

Based on Kane's (2013) framework and a practical guide to Kane's work written by Cook (2015), the validity of AIG was assessed through a statement about the proposed use of AIG and four inferences: (i) scoring; (ii) generalisation; (iii) Extrapolation; and (iv) Implication. The *scoring* inference evaluates the process of item construction. More specifically, it refers to the process of moving from observable performances to observable scores. *Generalisation* refers to the degree to which the assessment tool represents all possible events. The *extrapolation* inference is related to an item's ability to predict real-world performances and refers to evidence of how well candidates will perform in future events. Finally, *implication*

Table 1 Quality assessment of included studies

Quality assessment topics.	i) item modelling: definition and related concepts		ii) developing item models		iii) item model taxonomy		iv) using item models to automatically generate items		v) benefits of item modelling		vi) item model bank		vii) estimation of the statistical characteristics of generated items.		viii) description of the three-step process for conducting AIG		ix) assessment of AIG's capacity to generate new items		x) quality assessment of generated items, cognitive model and/or item model		xi) comparison of AIG with traditional methods of item development		xii) limitations of AIG.	
Study	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC	FF	PC
(Gierl et al., 2012a, b)	2	2	2	2	0	0	2	2	2	2	0	1	0	0	2	2	2	2	0	1	1	0	2	2
(Gierl & Lai, 2013a)	2	2	2	2	0	0	2	2	1	1	2	2	0	0	2	2	2	2	2	2	2	2	2	2
(Gierl & Lai, 2013b)	2	2	2	2	2	1	2	2	2	2	2	2	1	1	2	2	2	2	2	1	0	1	0	0
(Gierl et al., 2016)	2	2	2	2	0	0	2	2	2	2	2	2	2	2	2	2	2	1	2	2	2	2	2	2
(Gierl & Lai, 2016)	2	2	2	2	0	0	2	2	2	2	2	1	2	1	2	2	1	2	2	2	2	1	2	2
(Gierl & Lai, 2018)	2	2	2	2	0	0	2	2	1	1	2	2	0	0	2	2	1	1	2	2	0	0	1	1

Table 1 (continued)

Quality assessment topics.	i) item modelling: definition and related concepts		ii) developing item models		iii) item model taxonomy		iv) using item models to automatically generate items		v) benefits of item modelling		vi) item model bank		vii) estimation of the statistical characteristics of generated items.		viii) description of the three-step process for conducting AIG		ix) assessment of AIG's capacity to generate new items		x) quality assessment of generated items, cognitive model and/or item model		xi) comparison of AIG with traditional methods of item development		xii) limitations of AIG.	
(Lai et al., 2016a, b)	2	2	2	2	0	0	2	2	2	1	0	0	2	2	2	2	2	1	2	2	0	0	2	1
(Pugh et al., 2016)	2	2	2	2	0	0	2	2	2	2	0	0	0	0	2	2	1	1	2	1	2	0	0	0
(Pugh et al., 2020)	2	2	2	2	0	0	2	2	2	2	0	0	0	0	2	2	1	2	2	2	0	0	2	2
(Shappell et al., 2020)	2	2	2	2	0	0	1	1	2	2	2	2	0	0	2	2	2	2	0	0	0	0	1	1

Note-0: 'topic not covered in the study'; 1: 'topic coverage was unclear'; 2: 'topic was covered in the study' is related to decisions made after the test results are known (Cook et al., 2015; Kane, 2013; Tavares et al., 2018). Through this framework, we aim to obtain evidence that AIG presents validity properties that allows it to be employed in medical assessment.

Results

Description of included studies

Included studies were published between 2012 and 2020. Canada (n = 9, 90%) was the country with the highest number of included studies. Remaining data was from the USA. Considered studies focused on implementations of AIG in medical assessment. The purpose of these studies focused on generating medical MCQs through AIG and/or evaluating the quality of the items generated by AIG. Medical domains addressed in the studies were as follows: surgery, therapeutics, neonatal jaundice, upper gastrointestinal bleed, liver disease in adults and emergency medicine. Studies with medical students responding to the generated items collected samples ranging from 45 (Shappell et al., 2020) to 455 (Gierl et al., 2016; Lai, Gierl, Touchie, et al., 2016) students. The remaining studies involved small groups of experts or the process of developing/revising cognitive and item models.

Data synthesis

Data was synthesized in narrative form. Included studies, study country, purpose, method and respective results are detailed in Table 2.

Data synthesis: AIG's feasibility

Synthesized data allowed us to draw conclusions about the feasibility of AIG in medical assessment. We considered the time-spent, the number of items generated, and the ease of learning as indicators of feasibility.

Time-spent

Two studies mentioned the time-spent implementing AIG and the amount of items generated: Pugh et al., (2020) managed to generate 80–100 medical items in 90–120 min, while Gierl et al., (2012a, b) produced 1248 items from one single item model in just 6 h (3 h for step 1; 2 h for step 2; and 1 h for step 3). Both authors used a JAVA-based software called *Item GeneratOR* (IGOR) to generate new MCQs. This means that, approximately, one may generate 208 medical items per hour by using a single cognitive model. Step one seems to be the stage of the three-step process that requires more time (which is understandable since it is necessary to create a cognitive model with all the elements and constraints), while step three requires less time. By using AIG, experts are freed from time-consuming tasks characteristic of the item development process and will be able to devote more time to their educational tasks, which will improve their productivity and quality of teaching.

Table 2 Data synthesis

Reference	Country	Purpose	Method	Results
(Gierl et al., 2012a, b)	Canada	Present a methodology to generate MCQs.	AIG was used to generate MCQs.	In 6 h, 1248 items were generated from one item model: Stage 1 (3 h); Stage 2 (2 h), and Stage 3 (1 h).
(Gierl & Lai, 2013a)	Canada	Determine whether AIG generates high-quality items.	Items generated by AIG and items developed using traditional item development methods were blindly rated for quality by experts. Independent-samples Student t-tests were conducted to assess differences between the items in terms of quality. Subsequently, expert classified each item as generated by AIG or as an item developed using traditional methods.	Specialists developed 25 items using traditional item development methods. The same specialists then created 9496 using AIG. A second group of specialists developed 25 items using traditional item development methods. One t-test produced a statistically significant result ($p \leq 0.05$). Mean quality ratings were significantly higher for items developed using traditional methods on one indicator regarding the plausibility of the distractors (t [173] = 5.49, $p < 0.05$). On average, panellists correctly identified items generated by AIG 82% of the time; and incorrectly identified items generated by AIG 63% of the time. Overall predictive accuracy of the expert medical panellists was 42%.
(Gierl & Lai, 2013b)	Canada	Describe a method for generating test items.	AIG was conducted to generate MCQs using two types of item models: (i)1-layer item model and (ii) n-layer item model.	256 items were generated with the 1-layer item model; 16,384 items were generated with the n-layer item model.
(Gierl et al., 2016)	Canada	Assess the psychometric characteristics of items generated by AIG.	Items generated by AIG were distributed within nine tests. Students responded to the items across different forms, Item analysis was conducted using CTT.	465 items were generated using AIG. For the correct options, items used measured examinees' performance across a broad range of ability levels and provided strong levels of discrimination. For the incorrect options, items consistently differentiated the low from the high performing examinees.
(Gierl & Lai, 2016)	Canada	Assess the quality of items generated by AIG.	Authors describe a method to evaluate the quality of items generated by AIG.	If the instructions for item generation in the models are adequate, the generated items will be appropriate for testing.
(Gierl & Lai, 2018)	Canada	Describe a method for generating items using AIG and rationales required for formative testing.	AIG was used to generate MCQ and the corresponding rationale for each item.	48 items were generated using the content from the cognitive model. Rationale generation added extra time to the AIG process. Rationales satisfied the required characteristics of feedback.

Table 2 (continued)

Reference	Country	Purpose	Method	Results
(Lai et al., 2016a, b)	Canada	Describe and validate a method of generating distractors using AIG (*systematic distractor generation*)	Systematic distractor generation was integrated with AIG's 3-step process. 13 items were selected for field test. Generated items were distributed across examination forms. 455 medical students responded to the items. Item analysis was conducted following indices from CTT.	Results for the correct option: items measured a wide range of difficulty from the same model and presented consistent levels of discrimination. Results for the incorrect options: generated distractors were effective alternatives as they contained information that consistently appealed to lower performing candidates.
(Pugh et al., 2016)	Canada	Provide a framework for the development of quality MCQs.	Authors detail a framework for the development of high-quality MCQs using cognitive models.	The approach allowed the efficient generation of MCQs. Authors found that even a group of novices could apply the method to create a complete cognitive model within about 2 h, resulting in 5–10 new items.
(Pugh et al., 2020)	Canada	Compare the quality of items developed using AIG and the quality of items developed with traditional methods.	Items developed using AIG and traditional methods were blinded reviewed by content experts. A Wilcoxon two-sample test was employed for each quality metric rating scale as well as for the overall cognitive domain judgment scale.	AIG generated between 80–100 items. The entire process required 90–120 min. 51 items created with traditional methods and 51 items generated using AIG were evaluated for quality; AIG items were not perceived as differing from traditionally developed items.
(Shappell et al., 2020)	USA	Investigates an approach to item generation for mastery learning tests.	47 residents of an emergency medicine program took a mastery learning test. 20 item models were created and reviewed by educators. Two versions of the test were created. Consistency was evaluated using the test—retest k statistic and decision-consistency classification indices.	912 MCQ were developed using AIG. Unique iterations per item model ranged from 24 to 128, offering millions of unique 20-question tests. No significant differences in mean learner performance, mean item difficulty and item discriminations across the tests were found.

Number of items generated

Gierl & Lai (2013b) illustrated item generation using two types of item models: (i)1-layer item model and (ii) n-layer item model. In the 1-layer item model, only a small number of elements are manipulated to generate new items, while in the n-layer item model many elements are manipulated. The authors generated 256 items with the 1-layer item model, while 16,384 items were generated with the n-layer item model. Gierl & Lai (2018) used IGOR to generate 48 items and rationales required for formative testing from a single cognitive model, while Shappell et al., (2020) generated 912 MCQs and millions of single 20-question tests for a mastery learning test in emergency medicine. AIG enables the productions of hundreds of new items, which can significantly reduce the shortages affecting item banks and ease the item development burden in medical assessment.

Ease of learning

In Gierl & Lai's (2013a) study, a group of specialists was able to use the IGOR software and generate 9496 new items soon after learning the principles of AIG in a single workshop. Before the workshop, the same specialists only developed 25 items using a more traditional methodology (Gierl & Lai, 2013a; Pugh et al., 2016) claim that even novices can learn how to deploy AIG. In their study, the authors report that a group of novices learned to design a complete cognitive model structure which generated 5–10 new items within about 2 h (Pugh et al., 2016). These data reveal that AIG seems to be easy to learn, which is one more point in favour of its feasibility.

Based on these results, one should consider AIG as a feasible method for item development in medical assessment. By producing a vast number of items quickly and without requiring a great level of expertise, AIG seems to have the upper hand over more traditional item development methods in terms of feasibility.

Data synthesis: quality of MCQs generated by AIG

Scholars are determined to use AIG to develop high-quality items. This is evidenced by the number of studies focused on this topic and the different methodologies used to assert the quality of MCQs generated in medical assessment. We list these methodologies below.

AIG vs. Traditional item development methods

One method used by authors to ensure the quality of items generated by AIG is to compare them with items developed using traditional methods through blind review processes. In general, ratings of quality do not seem to differ between traditionally developed items and items generated by AIG (Gierl & Lai, 2013a; Pugh et al., 2020). The only difference found concerns the plausibility of distractors, as AIG seems to include fewer plausible distractors than items developed using more conventional methods (Gierl & Lai, 2013a). However, no differences were found with respect to other quality indicators, which bodes well for the quality of the items generated by AIG (Gierl & Lai, 2013a).

Psychometric analysis of items generated by AIG

Another method used to assess the quality of items generated by AIG is related to psychometric analysis. Following psychometric indices from Classical Test Theory (CTT), items generated by AIG seem to measure examinee performance across a wide range of ability levels and provide strong levels of discrimination. These indices also revealed that the distractors generated by AIG consistently differentiated between low-and high- performing examinees. Furthermore, item statistics for the correct options were found to be comparable between items developed using traditional methods and items generated by AIG. Index of discrimination and the biserial correlations for the distractors were similar as well. These data ensures that items generated by AIG present quality. This is true for the distractors generated automatically as well, even if they present slightly lower quality ratings than the ones developed using conventional methods (Gierl et al., 2016; Lai, Gierl, Touchie, et al., 2016).

Revisions of the cognitive model

One innovative method used to assess items for quality refers to revisions of the cognitive and item model prior to item generation. Using this method, experts evaluate the content outlined in the cognitive model and the task specified in the item model. Assuming that the generation instructions formulated in the models are adequate, then the generated items should be suitable for testing (Gierl & Lai, 2016). By reviewing the models, experts are exempt from reviewing each item individually (as in conventional item development methods), saving time and effort.

These results ensure that there is a plurality of methods that can be used to assess the quality of items generated by AIG. In general, the available literature does not seem to find major differences in the quality of items generated by AIG and items developed using more traditional methods.

Validity assessment

Validity assessment of AIG is detailed on Table 3. The table lists the evidence found in the included studies regarding each inference from Kane's (2013) framework. Firstly, as expected, the proposed use of AIG in medical assessment is to generate MCQs, especially for medical licensure tests. The main type of assessment appears to be summative assessment. However, in two studies, the implementation of AIG occurred in formative assessment (Gierl & Lai, 2018) and in mastery learning assessment (Shappell et al., 2020). Furthermore, references to computerized adaptive testing are notable throughout the included studies, with several papers mentioning this type of assessment as a field where AIG can be used (e.g. Gierl et al., 2016).

Regarding the scoring inference, we can point out as appraisal of evidence the fact that, in most included studies, cognitive and item models were developed and reviewed by clinical experts. Through this type of analysis, scoring rubrics, rules and procedures become clearly outlined, allowing for the generation of quality items. There is also a concern of the authors of the included studies regarding the quality of the items generated. This is evident when we think about the different methodologies that the authors used to assess items for quality. These best practices seem to support AIG towards the scoring inference.

Table 3 AIG validity assessment

Inferences		(Gierl et al., 2012a, b)	(Gierl & Lai, 2013a)	(Gierl & Lai, 2013b)	(Gierl et al., 2016)	(Gierl & Lai, 2016)	(Gierl & Lai, 2018)	(Lai et al., 2016a, b)	(Pugh et al., 2016)	(Pugh et al., 2020)	(Shappell et al., 2020)
Proposed use of AIG		Generate MCQs for medical licensure testing.	Generate MCQs for medical licensure testing.	Generate MCQs for medical licensure testing.	Generate MCQs for medical licensure testing.	Generate MCQs for medical assessment.	Generate MCQs and rationales for medical formative testing.	Generate MCQs and distractors for medical licensure testing.	Generate MCQs for medical assessment.	Generate MCQs for medical assessment.	Generate MCQs for medical mastery learning assessment.
Scoring	**Existing evidence**	Cognitive and item models were developed and reviewed by specialists.	Items were blindly evaluated for quality by a panel of experts.	Cognitive and item models were developed and reviewed by specialists.	Cognitive and item models were developed and reviewed by specialists.	Experts evaluated the content and the logic specified in the cognitive model and in the item model.	Experts blindly reviewed the rationales generated for formative testing.	Cognitive and item models were developed and reviewed by specialists.	Cognitive and item models were developed and reviewed by specialists.	Quality of items generated was evaluated by experts.	Item models were developed and reviewed by specialists.

Table 3 (continued)

Inferences		(Gierl et al., 2012a, b)	(Gierl & Lai, 2013a)	(Gierl & Lai, 2013b)	(Gierl et al., 2016)	(Gierl & Lai, 2016)	(Gierl & Lai, 2018)	(Lai et al., 2016a, b)	(Pugh et al., 2016)	(Pugh et al., 2020)	(Shappell et al., 2020)
Generalisation	**Existing evidence**	UN	UN	UN	Item response theory was used, but not reported. CTT was used. Generated items measured a broad range of difficulty levels.	UN	UN	CTT was used. Generated items measured a broad range of difficulty levels;	UN	UN	No significant differences in item difficulty between tests were found.
Extrapolation	**Existing evidence**	UN	UN	UN	Consistent levels of item discrimination.	UN	UN	Consistent levels of item discrimination.	UN	UN	No significant differences in mean item discrimination between tests were found.
Implications	**Existing evidence**	UN	UN	UN	UN	UN	UN	UN	UN	UN	UN

*UN - Unclear / Unreported

Only studies that have conducted psychometric analysis on the items generated by AIG (Gierl et al., 2016; Lai, Gierl, Touchie, et al., 2016; Shappell et al., 2020) can point to data that support the generalization inference. These studies applied the CTT framework and concluded that the items generated by AIG were able to adequately cover the spectrum of difficulty, presenting both easy and difficult items. This wide distribution is important as it allows estimating different levels of ability, which represents all the possible events and is important in the context of medical assessment. The extrapolation inference was supported by studies that have conducted these types of analysis: since the items generated by AIG presented consistent levels of item discrimination, one can conclude that they allow differentiating subjects with different levels of ability, which helps predict students' performance on future tests.

Finally, we had difficulties in obtaining data that fulfilled the implication inference. This difficulty may initially be interpreted as a limitation of AIG. However, it should be considered that decisions arising from the results of student assessments seem to escape the purpose of AIG, which is only to generate items and to overcome the burden of item development. Furthermore, it is not expected that all elements displayed should be used in the validation process (Cook et al., 2015), which is why we believe we have found evidence pointing to the validity of this assessment methodology.

Discussion

Developing MCQs in medical assessment is challenging, as it requires specialists to transmute the skills and content needed to diagnose medical problems into test item format (Gierl et al., 2012a, b). Due to the transition to CBT, there is an urgency for new MCQs. However, conventional item development methods seem unable to meet this need (Gierl et al., 2012a, b; Gierl & Lai, 2013a; Royal et al., 2018). AIG emerges as a solution to ease the item development burden by generating new test items based on computer modules (Arendasy & Sommer, 2007; Gierl et al., 2012a, b, 2016). Despite this promising scenario, AIG remains unknown in practice, which is why a narrative review on this topic is pertinent.

In this study, we collected data to promote a general discussion on AIG (Green et al., 2006). Given the importance of narrative in medical science (Baethge et al., 2019), such work could be important for the dissemination of AIG in medical schools. By putting information into perspective through understanding AIG while ensuring its feasibility, validity and item quality, teachers, physicians, and specialists struggling with the scarcity of new items can embark on a paradigm shift that will be beneficial to medical assessment.

However, such review has limitations. First, unlike systematic reviews, there are no specific guidelines for conducting narrative reviews (Ferrari, 2015). We tried to overcome these limitations by implementing procedures typical of systematic reviews in order to reduce bias: we designed an appropriate search strategy; used a checklist to evaluate the quality of our review; defined selection and exclusion criteria; performed two types of screening of the studies that emerged with the search strategy; and we assessed the quality of the included studies. Secondly, the number of studies included was small. Third and finally, there was not much variety regarding the authors who conducted the studies on this topic, which proves the need for more literature on AIG.

On the other hand, this review offers, at least, four outcomes. First, it provides readers with an overview of AIG. We framed the changes occurring in medical assessment with the urgency for new test items; introduced the concept of AIG; described the three-step process proposed by Gierl & Lai (2012) for generating medical MCQs; and elaborated on the need for empirical data regarding the feasibility, validity and item quality of AIG.

Second, this review provides data on the feasibility of AIG. We assessed feasibility by considering a series of criteria: time-spent implementing AIG; number of items generated; and ease of learning. Regarding time-spent, AIG enabled the development of 80 to 1248 items in a relatively short time (1–6 h; 208 items per hour) (Gierl et al., 2012a, b; Pugh et al., 2020). Among the three steps described by Gierl & Lai (2012), the first (development of a cognitive model) was found to be the most time consuming (3 h on average) (Gierl et al., 2012a, b). Considering the number of items generated, almost all authors were able to generate more than 100 new MCQs in their AIG-based experiments—except Gierl & Lai (2018), who employed AIG to generate items along with rationales for formative assessment, which is more complex. Finally, implementing AIG seems to be an easy process to learn, with novices producing a considerable number of items soon after learning the basics processes (Gierl & Lai, 2013a; Pugh et al., 2016). Based on these data, we have reasons to believe that AIG is a feasible method for generating medical MCQs.

Third, this review synthesized data regarding the quality of items generated by AIG. According to the available literature, there seems to be a general concern regarding the quality of the items generated by AIG and their distractors. This was evident from the number of methodologies used to assert the quality of items generated by AIG. One of the strategies is to compare the quality of items generated by AIG with items developed by conventional methods. The results obtained revealed that no statistically significant differences were found (Gierl & Lai, 2013a; Pugh et al., 2020). A second strategy used by authors to assess the quality of items generated by AIG lies in the use of psychometric theory. Through the use of psychometric analysis, studies found that the distractors generated by AIG consistently differentiated between low-performing and high-performing examinees (Gierl et al., 2016; Lai, Gierl, Touchie, et al., 2016). Furthermore, they also revealed that the items generated by AIG were able to measure student performance across a wide range of ability levels, while providing strong levels of discrimination and ensuring item quality (Gierl et al., 2016; Lai, Gierl, Touchie, et al., 2016). A final methodology used to ensure item quality refers to conducting revisions to the cognitive models. Authors using this method claim that items are appropriate for testing if the cognitive model is properly structured (Gierl & Lai, 2016).

Fourth and finally, we found evidence regarding the validity of AIG in medical assessment. This validity assessment provided support for the use of automatically generated items in medical assessment. AIG gathered appraisals of evidence in most of the inferences in Kane's (2013) framework. Revisions made to the cognitive and item models ensure that scoring rubrics, rules and procedures are clearly outlined during the item generation process, which meets the requirements for the scoring inference. Furthermore, psychometric analysis revealed that the items generated by AIG cover a wide spectrum of difficulties and present consistent discrimination scores, which supports the generalization and extrapolation inferences. The only inference that AIG was not able to satisfy is related to the implication inference. However, since this inference is not related to the scope of AIG, we believe its validity remains assured.

With this review, some advantages of AIG have become evident. First, we must highlight AIG's ability to develop item banks. Multiple item models can be used to provide item banks with hundreds of new items, minimizing item exposure. Second, it is a scalable process since the item model is used as the unit of analysis. Consequently, one item model is capable of generating many test items (Gierl & Lai, 2013b). Third, it is a flexible approach for the generation of contents in health sciences. Since the model is the unit of analysis, the contents are easily updated to accommodate the latest changes (Gierl & Lai, 2016). Fourth, it is a timely and cost-effective process. While traditional item development methods require per-item costs associated with editing, pilot testing and calibration procedures, AIG efficiently generates new items because it reuses item models (Gierl & Lai, 2013b; Luecht, 2012). Fifth and finally, since items generated by AIG are stored in banks, they can be easily imported/exported, which allows item transfers between users and institutions in order to support testing (Gierl & Lai, 2016). This feature may be especially useful for local medical schools unable to apply AIG if large-scale test organizations or more developed medical schools are predisposed to transfer assessment content and share item banks.

However, AIG also presents limitations. One limitation lies in the revision of the cognitive models prior to item generation, which makes item reliability and validity dependent on these models (Gierl et al., 2012a, b). This limitation points to the validity issues surrounding AIG and the need to find other sources of evidence to solidify it as a valid method. The three step process also has limitations, as there is no clear methodology for generating distractors (Gierl & Lai, 2013a). While there are strategies for this purpose (e.g., systematic distractor generation) they typically require binary encodings that are not always applicable to patient conditions (Lai et al., 2016). Since AIG is an automated process, it requires greater programming efforts by software engineers and more subsequent assessments than traditional methods to ensure the quality of the processes (Kosh et al., 2019). Furthermore, it requires a minimum number of items to offset the investment made in model development and technology deployment, otherwise cost savings may not be achieved (Kosh et al., 2019). A final limitation relates to the scope of AIG. Thus far, the use of AIG with cognitive models has been limited to narrow content domains, which is why we cannot yet consider AIG as a comprehensive method (Gierl & Lai, 2012).

Directions for future research

Evidence to support AIG can be gathered in many ways. For example, through studies addressing the costs necessary to implement AIG and the psychometric properties of its items (Pugh et al., 2016). Since these studies are scarce (e.g., Kosh et al., 2019), a line of literature ensuring AIG as a timely and economical solution would be useful to deepen the knowledge about this assessment method and disseminate it in medical schools. Further lines of research should focus on expanding AIG to cover different item types. Although the focus of this review was on the generation of MCQs, it would be interesting to see studies implementing AIG on the generation of other items types (e.g., clinical decision-making items) (Gierl & Lai, 2018). Another important line of experiments would be the use of psychometric theory to assess the quality of items generated by AIG. Although the procedures used to assess item quality based on appropriate statistics are more challenging, these are more reliable than substantive judgments from content specialists (Gierl et al., 2016). We believe that the lack of information regarding the psychometric properties and validity of the items generated by AIG causes them to be viewed with suspicion by medical schools, test developers and psychometricians.

Another pertinent point for future research concerns the changes in education caused by the COVID-19 pandemic. Due to the health crisis we are experiencing, institutions all over the globe have been/are closed and examinations were suspended. Consequently, teachers were/are not able to evaluate students face to face, which renders conventional assessment methods purposeless (Choi & McClenen, 2020). Future studies should focus on the impact of distance learning using items generated by AIG. We believe AIG may be a force capable of countering the academic impact caused by distance learning: by generating large quantities of quality items and with the ability to frame them with formative assessment and computerized adaptive testing, the risks of distance learning can be mitigated, which may be advantageous for both teachers and students.

Conclusions

AIG is a next gen assessment method that combines human expertise with computer algorithms. Through this review, we collected data that allows us to conclude that AIG is a feasible and valid item development method capable of generating quality items.

Acknowledgements The authors acknowledge the Portuguese Foundation for Science and Technology (FCT) and Icognitus4All—It Solutions, Lda for providing a doctoral fellowship to FF (grant no. PD/BDE/150621/2020) through the PhDiHES program.

Funding This study was funded by National funds, through the FCT—project UIDB/50026/2020 and UIDP/50026/2020. The authors thank all study participants.

References

Arendasy, M., & Sommer, M. (2007). Using psychometric technology in educational assessment: The case of a schema-based isomorphic approach to the automatic generation of quantitative reasoning items. *Learning and Individual Differences*, 17(4), 366–383. https://doi.org/10.1016/j.lindif.2007.03.005

Baethge, C., Goldbeck-Wood, S., & Mertens, S. (2019). SANRA—a scale for the quality assessment of narrative review articles. *Research Integrity and Peer Review*, 4(1), 2–8. https://doi.org/10.1186/s41073-019-0064-8

Batalden, P., Leach, D., Swing, S., Dreyfus, H., & Dreyfus, S. (2002). General competencies and accreditation in graduate medical education. *Health Affairs*, 21(5), 103–111. https://doi.org/10.1377/hlthaff.21.5.103

Blum, D., & Holling, H. (2018). Automatic generation of figural analogies with the IMak package. *Frontiers in Psychology*, 9(AUG), 1–13. https://doi.org/10.3389/fpsyg.2018.01286

Choi, J., Kim, H., & Pak, S. (2018). Evaluation of Automatic Item Generation Utilities in Formative Assessment Application for Korean High School Students. *Journal of Educational Issues*, 4(1), 68–89. https://doi.org/10.5296/jei.v4i1.12630

Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences (Switzerland)*, 10(22), 1–17. https://doi.org/10.3390/app10228196

Cole, B. S., Lima-Walton, E., Brunnert, K., Vesey, W. B., & Raha, K. (2020). Taming the Firehose: Unsupervised Machine Learning for Syntactic Partitioning of Large Volumes of Automatically Generated Items to Assist Automated Test Assembly. *Journal of Applied Testing Technology*, 21(1), 1–11

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, 49(6), 560–575. https://doi.org/10.1111/medu.12678

David, M. F., Ben, Davis, M. H., Harden, R. M., Howie, P. W., Ker, J., & Pippard, M. J. (2001). AMEE medical education guide no. 24: Portfolios as a method of student assessment. *Medical Teacher*, 23(6), 535–551. https://doi.org/10.1080/01421590120090952

Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, 24(4), 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Gierl, M. J., & Haladyna, T. M. (2012). Automatic item generation: Theory and practice. *Automatic Item Generation: Theory and Practice*, 1–246. https://doi.org/10.4324/9780203803912

Gierl, M. J., & Lai, H. (2012). The Role of Item Models in Automatic Item Generation. *International Journal of Testing*, 12(3), 273–298. https://doi.org/10.1080/15305058.2011.635830

Gierl, M. J., & Lai, H. (2013a). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical Education*, 47(7), 726–733. https://doi.org/10.1111/medu.12202

Gierl, M. J., & Lai, H. (2013b). Instructional Topics in Educational Measurement (ITEMS) Module: Using Automated Processes to Generate Test Items. *Educational Measurement: Issues and Practice*, 32(3), 36–50. https://doi.org/10.1111/emip.12018

Gierl, M. J., & Lai, H. (2016). A Process for Reviewing and Evaluating Generated Test Items. *Educational Measurement: Issues and Practice*, 35(4), 6–20. https://doi.org/10.1111/emip.12129

Gierl, M. J., & Lai, H. (2018). Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing. *Applied Psychological Measurement*, 42(1), 42–57. https://doi.org/10.1177/0146621617726788

Gierl, M. J., Lai, H., Pugh, D., Touchie, C., Boulais, A. P., & De Champlain, A. (2016). Evaluating the Psychometric Characteristics of Generated Multiple-Choice Test Items. *Applied Measurement in Education*, 29(3), 196–210. https://doi.org/10.1080/08957347.2016.1171768

Gierl, M. J., Lai, H., & Turner, S. R. (2012a). Using automatic item generation to create multiple-choice test items. Medical Education, 46(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012a.04289.x

Gierl, M., Lai, H., & Turner, S. (2012b). Using automatic item generation to create multiple-choice test items. Medical Education, 46(8), 757–765. https://doi.org/10.1111/j.1365-2923.2012b.04289.x

Gierl, M., Lai, H., Hogan, J., & Matovinovic, D. (2015). A Method for Generating Educational Test Items That Are Aligned to the Common Core State Standards. *Journal of Applied Testing Technology*, 16(1), 1–18

Green, B., Johnson, C., & Adams, A. (2006). Writing narrative literature reviews for peer-reviewed journals: secrets of the trade. *Journal of Chiropractic Medicine*, 5(3), 101–117. https://doi.org/10.1162/ling_a_00246

Gunabushanam, G., Taylor, C. R., Mathur, M., Bokhari, J., & Scoutt, L. M. (2019). Automated Test-Item Generation System for Retrieval Practice in Radiology Education. *Academic Radiology*, 26(6), 851–859. https://doi.org/10.1016/j.acra.2018.09.017

Hornik, K., Gerber, F., & Friendly, M. (2020). & Davidmeyerr-projectorg, M. D. M. *Package ' vcd.'*

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. https://doi.org/10.1111/jedm.12000

Kosh, A. E., Simpson, M. A., Bickel, L., Kellogg, M., & Sanford-Moore, E. (2019). A Cost–Benefit Analysis of Automatic Item Generation. *Educational Measurement: Issues and Practice*, 38(1), 48–53. https://doi.org/10.1111/emip.12237

Lai, H., Gierl, M. J., Byrne, B. E., Spielman, A. I., & Waldschmidt, D. M. (2016a). Three Modeling Applications to Promote Automatic Item Generation for Examinations in Dentistry. *Journal of Dental Education*, 80(3), 339–347. https://doi.org/10.1002/j.0022-0337.2016a.80.3.tb06090.x

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A. P., & De Champlain, A. (2016b). Using Automatic Item Generation to Improve the Quality of MCQ Distractors. *Teaching and Learning in Medicine*, 28(2), 166–173. https://doi.org/10.1080/10401334.2016b.1146608

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174

Luecht, R. (2012). Automatic Item Generation for Computerized Adaptive Testing. In M. J. Gierl & T. M. Haladyna (Ed.), *Automatic Item Generation: Theory and PracticeTheory and Practice* (pp. 196–216)

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, 26(8), 709–712. https://doi.org/10.1080/01421590400013495

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan-a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 1–10. https://doi.org/10.1186/s13643-016-0384-4

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Medical Teacher*, 38(8), 838–843. https://doi.org/10.3109/0142159X.2016.1150989

Pugh, D., De Champlain, A., Gierl, M., Lai, H., & Touchie, C. (2020). Can automated item generation be used to develop high quality MCQs that assess application of knowledge? *Research and Practice in Technology Enhanced Learning*, 15(1), https://doi.org/10.1186/s41039-020-00134-8

Royal, K. D., Hedgpeth, M. W., Jeon, T., & Colford, C. M. (2018). Automated item generation: The future of medical education assessment? *EMJ Innov*, 2(1), 88–93

Shappell, E., Podolej, G., Ahn, J., Tekian, A., & Park, Y. S. (2020). Notes From the Field: Automatic Item Generation, Standard Setting, and Learner Performance in Mastery Multiple-Choice Tests. *Evaluation and the Health Professions*, 1–4. https://doi.org/10.1177/0163278720908914

Tavares, W., Brydges, R., Myre, P., Prpic, J., Turner, L., Yelle, R., & Huiskamp, M. (2018). Applying Kane's validity framework to a simulation based assessment of clinical competence. *Advances in Health Sciences Education*, 23(2), 323–338. https://doi.org/10.1007/s10459-017-9800-3

von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, 83(4), 847–857. https://doi.org/10.1007/s11336-018-9608-y

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.