



Progress is impossible without change: implementing automatic item generation in medical knowledge progress testing

Filipe Manuel Vidal Falcão^{1,2,3}  · Daniela S.M. Pereira^{1,2,3} · José Miguel Pêgo^{1,2,3} · Patrício Costa^{1,2}

Received: 16 May 2023 / Accepted: 26 June 2023 / Published online: 8 July 2023
© The Author(s) 2023

Abstract

Progress tests (PT) are a popular type of longitudinal assessment used for evaluating clinical knowledge retention and long-life learning in health professions education. Most PTs consist of multiple-choice questions (MCQs) whose development is costly and time-consuming. Automatic Item Generation (AIG) generates test items through algorithms, promising to ease this burden. However, it remains unclear how AIG-items behave in formative assessment (FA) modalities such as PTs compared to manually written items. The purpose of this study was to compare the quality and validity of AIG-items versus manually written items. Responses to 126 (23 automatically generated) dichotomously scored single best-answer five-option MCQs retrieved from the 2021 University of Minho PT of medicine were analyzed. Procedures based on item response theory (IRT), dimensionality testing, item fit, reliability, differential item functioning (DIF) and distractor analysis were used. Qualitative assessment was conducted through expert review. Validity evidence of AIG-items was assessed by using hierarchical linear modeling (HLM). The PT proved to be a viable tool for assessing medical students cognitive competencies. AIG-items were parallel to manually written-items, presenting similar indices of difficulty and information. The proportion of functional distractors for both AIG and manually written items was similar. Evidence of validity for AIG-items was found while showing higher levels of item quality. AIG-items functioned as intended and were appropriate for evaluating medical students at various levels of the knowledge spectrum.

Keywords Item development · Automatic item generation · Progress testing · Medical education · Item response theory · Hierarchical Linear Models

1 Introduction

Health professions educators have long recognized the importance of fostering a lifelong learning culture in their students (Demuth et al., 2018). However, learning should not be a matter of whether students learn but how they learn (Görlich & Friederichs, 2021). The purpose of progress tests (PTs), which involve testing students with a comprehensive exam that covers every aspect of the curriculum in succession, is to break the examination's steering effect and encourage deep and long-term learning (Albanese & Case, 2016). More specifically, it is a methodical and ongoing evaluation that seeks to gauge students' knowledge at the level they should have attained by the end of the programme (Cecilio-Fernandes et al., 2018).

The majority of medical examinations include multiple-choice questions (MCQs) (Royal et al., 2018). However, creating MCQs for formative assessment (FA) modalities such as PTs is expensive and time-consuming since specialists must manually create and fine-tune each item individually. This onerous task is impractical when hundreds of items are required for multiple test versions or to populate item banks (Lai, Gierl, Byrne et al., 2016). Consequently, there is an urgent need for content-specific test items (and respective feedback) to be available for FA because these questions are regularly given to students, especially in PTs (Gierl & Lai, 2015).

In the next-generation assessment theory known as Automatic Item Generation (AIG), computer algorithms swiftly develop large sets of testing items with precise content (Falcão et al., 2022, 2023). Despite the encouraging structure and applicability in high-stakes examinations (e.g., Gierl and Lai, 2013; Pugh, de Champlain et al., 2016), few is known about how AIG-items behave in FA modalities such as PTs compared to manually written items used in higher education institutions (HEI). This study compared the psychometric properties of PT-generated AIG items versus manually written items. Based on a robust statistical pipeline, we examined data from a single medical PT conducted at the University of Minho, delivered in December 2021. Our research adds to the body of knowledge by determining whether AIG-items can live up to the high criteria required by testing procedures used in FA such as PTs.

1.1 Background

1.1.1 Assessment of learning or assessment for learning?

Educational institutions today require strong tools to accurately evaluate students' understanding of the concepts covered in their study materials (Nwafor & Onyenwe, 2021). Traditionally, the focus has been on summative assessment (SA), which is considered reliable (Knight, 2002). SA is conducted under standardized conditions using end-of-unit exams that provide a snapshot of educational achievement to guide critical decisions (Gierl & Lai, 2018). However, this approach merely calls for assessments of students' learning, not for providing them with feedback throughout the learning process (Boston, 2002).

Evaluation should not only determine if pupils have acquired knowledge (*assessment of learning*) but also encourage learning (*assessment for learning*) (Prashanti

& Ramnarayan, 2019). Teachers should therefore identify students' misconceptions and deliver feedback on their progress (Preston et al., 2020). FA refers to any practice that improves student learning based on feedback (Irons, 2007). It is a methodical procedure that entails acquiring data on student learning and using this knowledge to clarify which instructions should be modified to better meet students' needs (Black & Wiliam, 2009). Providing pupils feedback results in a positive evaluation of their performance, which in turn helps them develop their knowledge and shape their performances (Xinxin, 2019). Consequently, the developmental underpinnings of assessment become more prominent, and the roles of both teachers and students are acknowledged, making FA a cyclical programme of high and low-stake activities in which the student actively participates (Leenknecht et al., 2021).

1.1.2 Making progress: formative assessment within medical education

Student feedback moved centre stage and FA is currently used in medical education to monitor students learning, develop clinical competences and promote clinical reasoning (Schüttpelz-Brauns et al., 2020). However, educators typically refrain from delivering feedback since many students often report that the feedback they receive leaves them frustrated (Chowdhury & Kalu, 2004). Considering the shift from judgments based on test scores to student-based instruction, there is a need to engage students with learning in a safe environment for them to learn and for educators to deliver feedback (Watling & Ginsburg, 2019).

PTs are long-term, feedback-focused educational assessment strategies for determining how well student's knowledge develops and endures over time (Couto et al., 2019). Typically, all students receive them on a regular basis (2–4 times annually) during the academic program. The test samples all of the knowledge domains expected of graduates to master upon degree completion, regardless of the student's year level (Öztürk, 2013). Its main purpose is to promote deep learning while reducing the steering effect of exams (Albanese & Case, 2016). Considering their properties, PTs are effective and trustworthy instruments for assessing how much is being learned during undergraduate health professions education (Görlich & Friederichs, 2021).

FA such as PT were traditionally conducted using paper-based tests. However, this was logistically challenging and presented drawbacks, such as requiring students to gather in a single location at a specific time to administer the test and to obtain feedback (Olson & McDonald, 2004). To make matters worse, PTs are very resource-binding as they demand lots of testing items and time required for planning, implementing, and evaluating results (Koşan et al., 2019). Thanks to the advent of modern technology, we have witnessed significant growth in online FA systems, which have maximised the value of instant feedback (Joyce, 2018). Of these, online computerized formative testing (CFT) are technology-enhanced assessment systems that deliver students feedback on their learning progress (Gierl & Lai, 2018). These systems have a solid theoretical basis and are prevalent in HEI, being widely recognized in health professions schools as a resource for self-directed learning (Bijol et al., 2015). Through CFT, students are able to assess their knowledge and determine where they need more training. Additionally, since students are being virtu-

ally tested, they provide equity and inclusiveness, allowing students to be evaluated anonymously, anytime, and anywhere, providing a safe environment where trial and error is permitted (Mitra & Barua, 2015).

1.1.3 The underlying need for a next-gen assessment theory: automatic item generation

Most tests within medical education, such as PT, are composed of multiple-choice questions (MCQ) (Dijksterhuis et al., 2009). MCQs are economical, efficient, instantly scored and are appropriate for item types that include a variety of skills (McCoubrie, 2004). These types of questions rapidly adapted to the hypermedia environment of CFTs given the ability of these systems to produce automated tests, which facilitates the delivery of feedback (Farrell & Leung, 2004). However, developing MCQ is costly (with prices ranging from US\$1500 to US\$2000 per item) and time-consuming for specialists since they must manually write and refine each item individually (Kosh et al., 2019). This daunting task is not feasible when hundreds of items are needed for multiple test versions or when thousands of items are needed to fill item banks (Lai, Gierl, Byrne et al., 2016). Since these items are continuously administered to students and their production is arduous, there is a pressing need for content-specific items and feedback to be available for CFT (Gierl & Lai, 2018; Xinxin, 2019).

Digitally based assessment methods produce sophisticated data that more closely reflects how students interacted with the items than in traditional settings (von Davier et al., 2021). Automatic Item Generation (AIG) is a next-generation assessment theory, validated empirically and theoretically, that promises to ease this burden (Falcão et al., 2022, 2023; Jendryczko et al., 2020). It is a contemporary method that combines the knowledge of content experts with computer modules to produce large numbers of high-quality and content-specific test items, both quickly and effectively, following specific guidelines (Lai, Gierl, Byrne et al., 2016). These guidelines cover the (a) creation of a model containing the variables to be manipulated (item model), which includes the *stem* (part of the item model with the data required for problem-solving), the *response options* (both correct and incorrect ones), a *lead-in-question* (the complete sentence with the question) as well as supporting information; and the (b) systematic fusion of the components listed above by computer algorithms to produce a large number of new items (Bejar, 2012).

According to Gierl and Lai (2012), a three-step approach is necessary to generate MCQs in health professions education using AIG: (i) in the first step, content specialists outline a framework for item generation with the knowledge and skills expected to be used by students to formulate a diagnosis (*cognitive model*) (Pugh, de Champlain et al., 2016). This framework identifies the problem specific to a test item, presenting different scenarios related to it, the variables to be manipulated for item generation, and the data required to establish a diagnosis (Gierl et al., 2012); (ii) In the second step, the contents of the framework are added to MCQ's to form item models, which are similar to templates highlighting the variables to be manipulated and contain the relevant information to answer each question and respective options (Gunabushanam et al., 2019); (iii) In the third and final step, specialized computer

modules work on the item models, manipulating data in components such as the stem and the options, systematically generating massive high-quality digital items (Prasetyo et al., 2020). Figure 1 outlines this process.

1.1.4 Implementing AIG in online PT

AIG promises to produce tests based on unlimited item banks rapidly. It hinders item exposure, predicts the psychometric properties of generated items and presents construct validity since it relies on cognitive mechanisms underlying task performance (Harrison et al., 2017). Although the many advantages of AIG have been proven in high-stakes exams, it is not known how AIG-items compare to handwritten items in FA modalities such as PTs. In response to these concerns, strategies to enhance the quality/validity of AIG-items have undergone a lot of development. However, the quality/validity of these items has only been the subject of a small number of research (e.g., Falcão et al., 2022; Pugh et al., 2020), particularly in high-stakes examinations administered by licensing bodies whose resources to develop and control the quality of developed items and inherent costs are significantly different from HEI. The University of Minho’s School of Medicine (EMUM) has added AIG to the MCQ content development process during the past few years, increasing the capacity of the item banks available to evaluate its students. To gauge the potential of these items, AIG items have progressively been included on tests. In the present study, we first provide a psychometrical analysis of a PT conducted at the EMUM, where 23 AIG-items were included. Along with the psychometric approach considered, the AIG and manually written items included in the PT were both subjected to a qualitative evaluation and a validity assessment procedure. Our study contributes to the body of knowledge by evaluating whether AIG-items can meet the high standards expected of testing methodologies used in FA, such as PT in HEI.

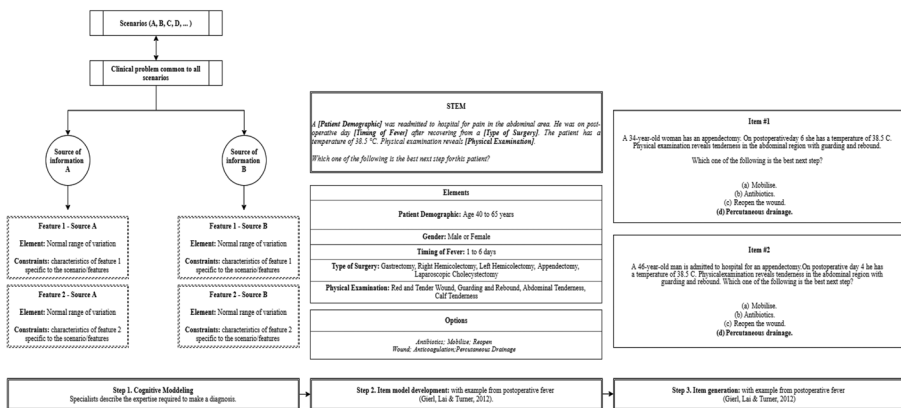


Fig. 1 Three-step process for generating medical MCQs based on AIG

2 Statistical and psychometric methods

2.1 Study design

This study involved a mixed-methods analysis of real data using software capable of gauging the psychometric qualities of the items included in the PT.

2.2 Data collection and sample

We analysed the responses to 126 dichotomously scored single best-answer five-option MCQs from the EMUM PT of Medicine, administered in December 2021. The questions, which were presented in the form of clinical scenarios, were designed to gauge how well medical knowledge—including that from the foundational medical sciences—is applied. Of these, 23 (18%) were automatically generated from previously designed cognitive models, generating hundreds of different items. The selection of the AIG-items was random and solely based on the PT topics. Different topics/disciplines, each with unique target competencies, conditions and unique item numbers, were covered in the PT (Cf. Appendix A). Candidates should understand subject-specific elements as part of the content objectives of the PT. 279 medical students of the EMUM (clinical years only) were used as our sample. Most students were females (72.9%) with ages ranging from 21 to 40 ($M=24$; $SD=3.20$). 243 of these students have been enrolled in the school's new curriculum plan (MinhoMD)¹ since 2020. The remaining students were enrolled in an alternative curricular plan.

2.3 Procedure

Students completed the 3.5-hour PT through an electronic testing platform (*QuizOne*®) under online supervision from their teachers. QuizOne is an e-assessment management system integrated with AIG functionalities that are designed to deliver and administer knowledge tests, among other features. Results obtained by the students in the PT did not contribute to SAs. After finishing the PT, students submitted their responses, and the platform closed the respective session.

2.4 Psychometric approach

The psychometric properties of the PT questions were analyzed using an Item Response Theory (IRT) approach. The Rasch model (RM) (Rasch, 1960) was employed in the PT data due to its determinants of item response (respondent's ability and item difficulty) and its relevance for achievement tests, providing a proper scalling method to establish measures based on students' response patterns (Hohen-

¹ The Minho MD program provides a major emphasis on the needs of the students, allowing them to design their own course of study, select their own path through optional units, and even sign up for classes in a variety of areas including biomedical research, management, or economics. The alternative curricular plan, on the other hand, provides an opportunity for students already trained in other areas to take the medicine course. These students join the MinhoMD students after a year of study in which they learn the fundamental sciences underlying medicine.

sinn & Kubinger, 2011; Tor & Steketee, 2011). The RM explains the conditional probability of a binary outcome, considering the person's latent trait level (θ) and the item's difficulty level (Rasch, 1960). Mathematical representation of this relationship is as follows:

$$P(Y_{ij} = 1 | \theta_j, b_i) = \frac{\exp(D(\theta_j - b_i))}{1 + \exp(D(\theta_j - b_i))} \quad (1)$$

Where $P(Y_{ij} = 1)$ is the probability of correctly answering an item, θ_j is the level of the latent trait of respondent j ($j=1, \dots, J$), b_i is the item difficulty parameter for item I ($i=1, \dots, I$), and D is a scaling constant that maps the model's parameters to the scale of a typical ogive model (Desjardins & Bulut, 2018). We recommend De Champlain's (2010) work for a brief overview of IRT.

Our statistical pipeline was as follows: first, we evaluated if fundamental pre-sumptions were true prior to running the RM: (a) *unidimensionality* (i.e., the PT was optimally measuring a single underlying construct); (b) *local independence* (i.e., absence of systematic conditional covariance among items); and (c) *monotonicity between θ and true scores* (i.e., the requirement that the probability of endorsing an item increases as θ increases). Appendix B contains the strategies for evaluating RM assumptions and respective results. Second, after ensuring model-data fit (Cf. Appendix C), we conducted a calibration process to estimate item properties (item difficulty – b_i) and obtain estimates of θ . Third, item reliability was evaluated using item information function (IIF) plots, whereas exam reliability was examined through the Kuder-Richardson 20 statistic (KR-20) (Kuder & Richardson, 1937), the Person Separation Index (PSI) and test information function (TIF) plots. Fourth, differential item functioning (DIF) was tested for the variable “curricular plan” through the Mantel-Haenszel (MH) method (Mantel & Haenszel, 1959). DIF is assessed in educational data to detect item-level bias and occurs when respondents from different subgroups display the same θ but answer differently across some items (Shea et al., 2012). The MH procedure is one nonparametric method for detection of DIF. It is based on comparing matched groups, so that item functioning can be evaluated conditional on θ (Socha et al., 2015). For *curricular plan* DIF, the RM again branched into two groups: (0) students enrolled in the new curricular plan; and (1) students enrolled in the alternative plan. Students from the new curricular plan were used as reference (focal group). Fifth and finally, we conducted a traditional distractor analysis to measure how well the incorrect options contributed to the quality of these MCQs. Since student's performance is influenced by how the distractors are designed, it is necessary to include plausible distractors that are more likely to attract examinees with partial knowledge (Desjardins & Bulut, 2018; Wind et al., 2019). Distractor analysis was conducted by examining the percentage of students who chose a particular distractor (Desjardins & Bulut, 2018; Gierl et al., 2017). Items with distractors selected by 5% or less of respondents were candidates for potential revisions as they were endorsed at such a low level as to suggest that most examinees did not consider them viable. A two-sample equal variance Student t-test compared the number of distractors needing revisions of both AIG-items and 23 manually written items that were selected at random through a random integer generator. The “N-1” Chi-squared

test(Campbell, 2007) compared the proportion of functional distractors between these items. All statistical procedures were based on the full sample. A significance level of $\alpha=0.05$ was set for all analyses.

2.5 Qualitative assessment (expert review)

When evaluating the quality of an item, it is important to consider not only its statistical data but also its qualitative information (Rzasa, 2002). A test development specialist with experience in evaluating MCQs conducted a blind qualitative assessment of each item administered in the PT. The expert used a proper item quality rating scheme (Jozefowicz et al., 2002) to evaluate the quality of each MCQ. The authors of the rating system created it to be consistent with standard item-writing guidelines. Each question was given a score between 1 (“*the item tested recall only and was technically flawed*”) and 5 (“*the item used a clinical or laboratory vignette, required reasoning to answer, and was free of technical flaws*”). The test development specialist scored the items independently, without being aware of the study’s objective. Data obtained was compiled using descriptive statistics. To determine whether there were statistically significant variations in the quality ratings between the two item types, a Mann-Whitney-U-Test (Mann & Whitney, 1947) was conducted.

2.6 Validity assessment (hierarchical linear modeling)

Validity refers to the extent to which available evidence supports the intended use of test results (American Educational Research Association, 2018). Response processes are an understudied yet promising source of validity evidence (Hubley, 2021). They describe the mental activities that a respondent employs when responding to test items (Russell & Hubley, 2017). It reaches the core of validity assessments, revealing how theory and evidence support the interpretation of test results. Additionally, they assist researchers in creating better test items, reducing construct irrelevant variance (Hubley, 2021). In the present paper, AIG-items response processes were evaluated using response time (RT; in seconds) as a measure for validity evidence. RT is one useful approach for examining response processes (Padilla & Benítez, 2014). However, it only provides indirect information regarding the difficulty and the degree of processing involved. Therefore, the validity of AIG-item response processes should be evaluated by examining the effect of the item type on RT in conjunction with student related variables through proper statistical models (Hubley, 2021).

In light of this restriction, and since we’re dealing with nested data, we employed hierarchical linear modelling (HLM) analysis to our PT data to disentangle possible within- and between-student variance and to predict RT by reference to random and fixed item and student-level predictors. Using HLM, one obtains accurate estimates of standard errors of beta coefficients and information on the variance distribution between various levels of analysis (Klusmann et al., 2008). We examined the impact of the item type on student’s RT while considering answer switching (i.e., the number of changes from the initial response to other option(s) that the students considered more appropriate during the PT). The reason for choosing this variable is that items where response changes occur may be prompting students to explore their

doubts, leading them to change their answers. Item (i.e., item type=Manual vs. AIG) and student variables (i.e., answer switching) were specified at the first and second level, respectively. This may be an interesting way to complement the analysis of the response processes used to answer AIG-items, as we can evaluate whether these can explore students' doubts as handwritten items do. In total, four different models were developed: (i) a random-intercept model (i.e., a model with no predictors; null model); (ii) an intermediate constrained model (ignoring between-cluster variation of the level-1 variable); (iii) an intermediate augmented model (considering between-cluster variation of the level 1-effect variable); and (iv) a final model including slope residuals and cross-level interactions - for a detailed insight at HLM, see Sommet and Morselli (2017). Continuous predictor variables were centered at the sample's grand mean to enhance the regression coefficients' interpretability. Dichotomous variables kept their original metrics.

2.7 Software

The procedure was implemented in Winsteps (Version 5.3.0) and within the R open-source statistical programming environment (<http://www.r-project.org>) with the R-packages: “*Psych*” (Revelle & Revelle, 2015); “*mirt*” (Chalmers, 2012); “*eRm*” (Patrick et al., 2018); and “*Lme4*” (Kuznetsova et al., 2015).

3 Results

A detailed explanation of the results of each procedure follows.

3.1 Item calibration: student and question comparison

Appendix D provides parameter estimations and standard errors from the RM. 3 questions (X7; X28; X75) were previously dropped for lack of variance (100% correct answers). None of these items were automatically generated. The mean θ of the students was centered on 0.0 logits with a standard error of 0.19. Positive values of θ indicated better medical knowledge and negative values represent less knowledge. A θ value of 0 demonstrates that the PT was well calibrated for the students and that most respondents had a 50% chance of correctly answering most of the provided items. Overall, it can be concluded that the quality of the PT was not affected by the presence of AIG-items. Figure 2 displays the b_i hierarchy of the questions as answered by the students (i.e., the distribution of subject's θ and the distribution of item b_i levels within the same scale). The questions of the PT were parallel compared to the subject's θ and spread evenly. b_i estimates of the items represented the minimum θ required to correctly answer an item and ranged from -5.7 (AIG106) to 3.89 (AIG83) logits (the mean b_i value for the RM is 0). These AIG-items were at the extreme ends of the b_i spectrum, which may suggest that they were mismatched and require adjustment. Mean item b_i was 0.26 logit below the mean person θ . Items with b_i levels above and below the b_i average value had a balanced amount of items. This means that the items included in the PT covered a wide portion of the θ . The RM

Fig. 2 Map of the difficulty level of the questions from the RM

Note. Area on the left represents the distribution of the subject’s θ ; Area on the right represents the distribution of items; Items with the highest difficulty level are at the top, while the easiest items are at the bottom; Each # represents 4 students; each ‘.’ represents 4 students; The values on the left of each scale are in logits. T=2 standard deviations from the mean; S=1 standard deviation from the mean; M=mean

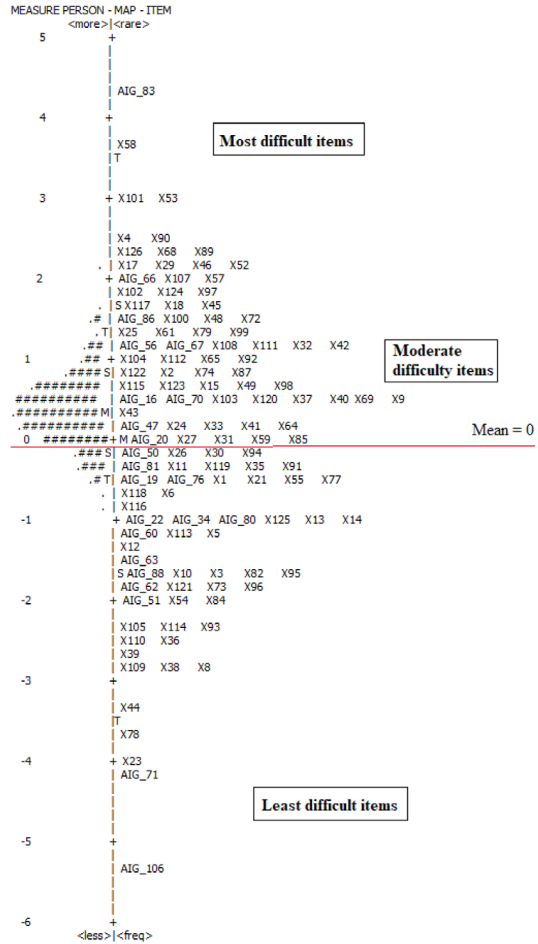
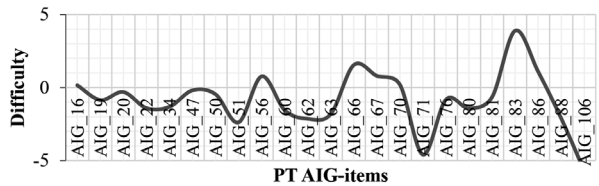


Fig. 3 Bi values for the PT AIG-items



estimates of the b_i values suggested that the PT was reasonably “moderate” and was targeted toward samples with medium levels of θ .

The range of b_i for AIG-items was between -5.7 (AIG106) and 3.89 (AIG83) logits, while the b_i for manually written items ranged between -4.31 (Q23) and 3.27 (Q58) logits. The RM calibration b_i levels for AIG-items were generally comparable to those of manually written items, supporting validity evidence for AIG. Figure 3 displays the b_i values for the AIG-items included in the PT. There were some AIG-items with b_i measures below the least able student (easy items) and few items with

b_i beyond the most able one (difficult items). However, most of the AIG-items were located at the same level that majority of the students (medium difficulty), demonstrating that they functioned as intended in the PT and were, therefore, appropriate for evaluating medical students at various levels of the θ continuum in HEI.

Figure 4 displays the item characteristic curves (ICCs) for the AIG-items included in the PT. Most items followed an increasing monotonic function. As we can see from this figure, most of the curves shift to the centre, which means there was a 50% probability that students with an ability of 0 logits answered most of the AIG-items correctly. Through this graphic visualization, it becomes clear that items AIG_83 and AIG_106 need to be revised. The first one was simple and could be easily solved even by those students with less θ . On the other hand, the second was extremely difficult and required great knowledge to be solved correctly.

3.2 Reliability

The KR20 was high (KR20=0.76), suggesting good internal consistency of the PT. PSI for the PT data was 0.76, which means that 76.4% of the variance in the observed scores was due to the estimated true variance in students' levels of clinical competence. The PSI determines how well students can be differentiated, with a value >0.70, meaning that the exam was adequate for group evaluation (Tennant & Conaghan, 2007). Once again, we found that the AIG-items successfully adapted to

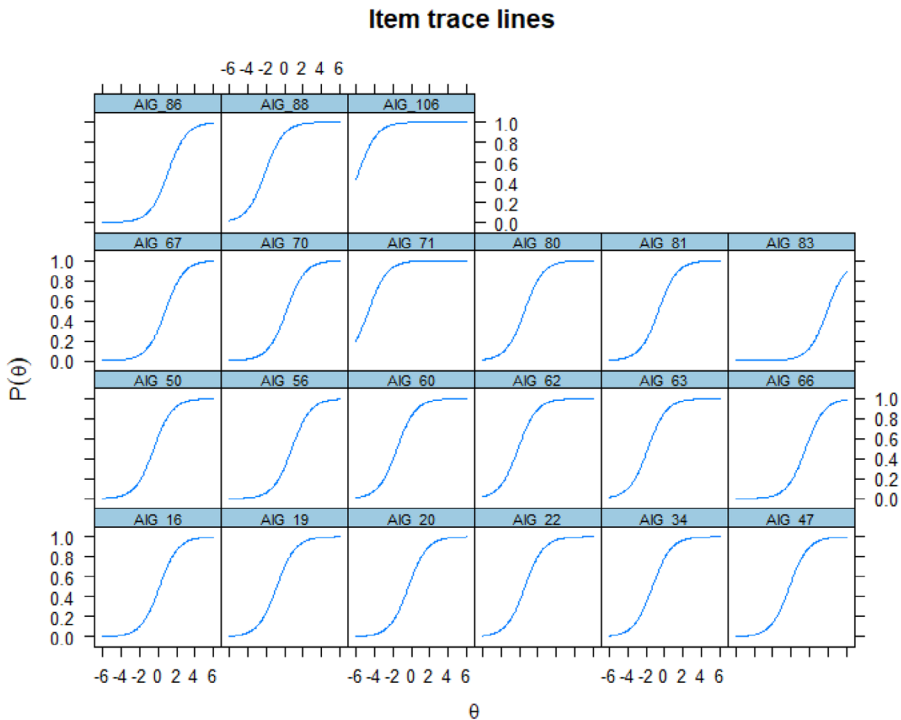


Fig. 4 ICCs for the PT AIG-items

the PT at the global level. Both item types provided an equivalent quantity of information, suggesting that the AIG-items were of good quality and offered measurement the same way as manually written items. The item information curves in Fig. 5 demonstrate how the AIG-items of the PT discriminated between different levels of θ . The high information values of these functions suggest that θ at relevant points may be precisely measured and used to distinguish students from adjacent θ levels. Once again, items AIG106 and AIG83 appeared to be problematic. Item AIG_71 also presented rather flat information curves for the left of the θ continuum, suggesting that these items could also require revision.

The information curve of all items of the PT was then summed up into the overall TIF presented in Fig. 6. The TIF was computed with the conditional standard error of measurement (CSEMs) to evaluate the accuracy with which the PT measured different values of θ throughout the continuum. For the PT, the TIF peaked at a trait level of around 0 (lowest amount of CSEM), consistent with the moderate difficulty values of the exam. The TIF illustrates the region of the underlying θ that is measured most precisely, revealing the reliability of the PT at different levels of student's θ . The PT provided an appropriate test information profile for the intended usage. The test provided reliable information between the middle and high end of the trait; thus, it was able to more accurately classify students in that area. Figure 6 also reveals that the PT was precise in differentiating between students with low and above average θ because high amounts of test information were gathered from these students.

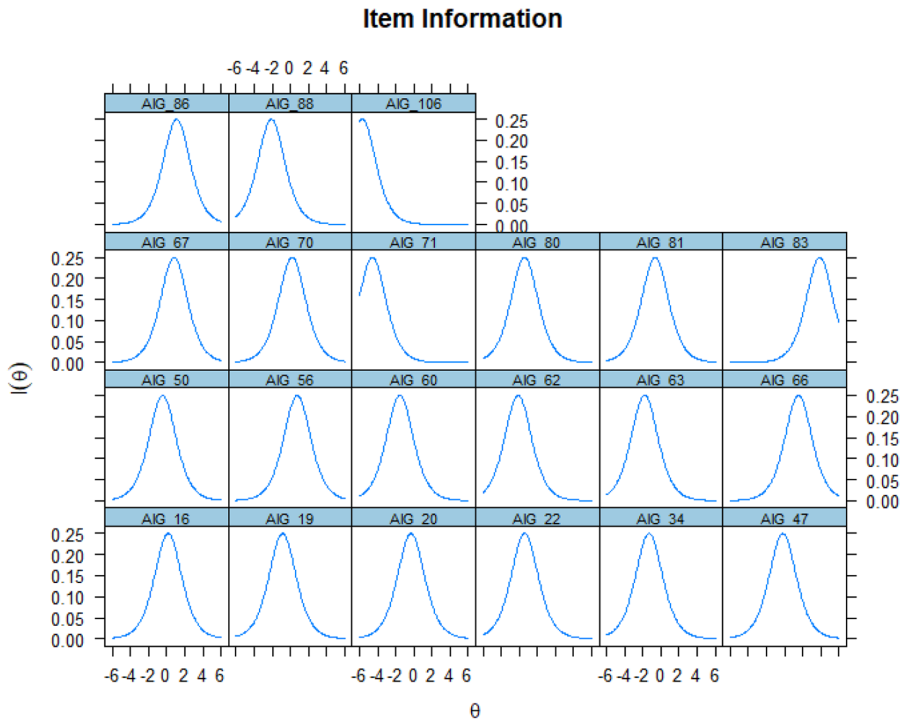


Fig. 5 IIFs for the PT-items

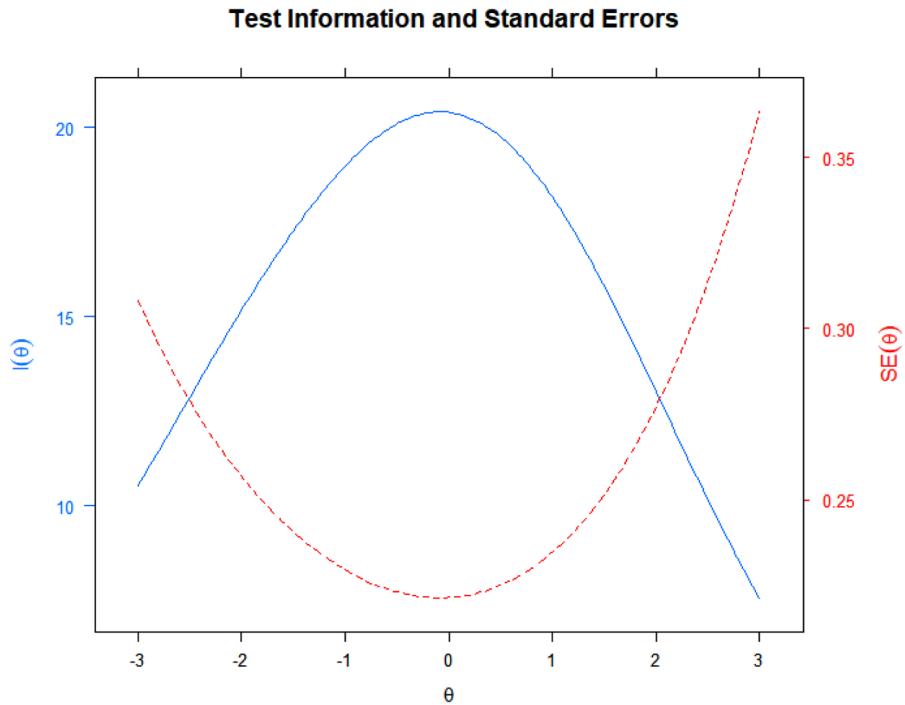


Fig. 6 TIF for the PT

Table 1 MH procedure for measuring and detecting DIF

Item	MH DIF statistic	α MH	Δ MH	Effect size
X3	4.99*	4.99	-1.84	Large
X33	4.76*	0.31	2.79	Large
X43	5.71*	3.20	-2.74	Large
AIG_66	5.20*	3.37	-2.85	Large
X72	6.87**	4.57	-3.57	Large
X84	14.41***	0.15	4.40	Large
X108	4.46*	3.05	-2.62	Large
X115	7.04**	3.29	-2.80	Large
X118	4.37*	0.31	2.73	Large
X119	3.93*	0.38	2.28	Large

Note: $p=0$ ****; $p<.001$ ***; $p<0.01$ **; $p<.05$ *

3.3 Quantifying DIF

Results from the MH chi-square (MH χ^2) test revealed that 10 items in the PT were flagged for the alpha level of $\alpha=0.05$ for exhibiting uniform DIF between students enrolled in the new curricular plan and students enrolled in the alternative plan (Cf. Table 1). 9 of them were manually written (X3; X33; X43; X72; X84; X108; X115; X118; X119). Only 1 AIG-item (AIG_66) presented DIF. Effect size measures (Δ MH) were used to supplement the chi-square test of statistical significance. DIF in the PT was balanced. Most Δ MH values were negative, indicating that most items

were not advantageous to the focal group. Only 4 items (X33, X84, X118, X119) presented positive ΔMH values, indicating that they benefited the focal group. Using the classification scheme for measuring the effect size of DIF developed by Dorans and Holland (1992), we found that the above mentioned items presented large DIF. Items that fall under this classification present values of $MH \chi^2$ that differs from one at the 5% level and a ΔMH that is more than 1.5 in absolute value – see Socha et al. (2015).

3.4 Distractor analysis

A total of 230 distractors were analysed. Distractor performance for the AIG-items was slightly positive. 39% of the distractors had a choice frequency $\leq 5\%$. Nearly 61% of the distractors were functional (choice frequency $> 5\%$). Distractor performance for the manually written items was slightly superior, with 32% of the distractors presenting a choice frequency $\leq 5\%$. This means that 68% of the manually written distractors were functional. Table 2 contains the frequency of choice distribution for the AIG-items' distractors. 2 AIG-items revealed problems with all distractors (AIG_71; AIG_106). 3 manually written items (X23; X44; X84) presented problems with all distractors. These problems may explain the ease of these items and respective poorer quality. The two-sample equal variance Student t-test revealed no significant differences between the number of distractors needing revisions of both item types ($t(45)=0.42, p=.15$). The N-1 test revealed no significant difference between the proportion of functional distractors of both item types ($\chi^2(1)=0.001, p=.977, 95\% \text{ CI} [-14.7, 14.9]$).

3.5 Qualitative assessment (expert review)

The majority of AIG-items received higher quality ratings ($M=4.40; SD=1.10$). Manually written items scored slightly lower ($M=4.12; SD=1.39$). These average scores for both item types means that not all items in the PT were able to satisfy, at least, one of three conditions: (i) a vignette; (ii) a one-best answer format (i.e., not true-or-false); and (iii) no technical flaw (Jozefowicz et al., 2002). The Mann-Whitney-U-Test revealed the existence of statistically significant differences in the quality ratings of both item types ($U=831, p<.05, r=.11$). Higher quality ratings were associated with AIG-items.

3.6 Validity assessment (HLM)

Table 3 contains the results of the HLM analysis. A separate null model for RT was first specified. The model revealed whether the means of RT differed across students (level-2 unit). The Intraclass Correlation Coefficient (ICC) (Hox et al., 2017) of the null hypothesis was 0.11 ($p<.001$), justifying the use of HLM. This means that 11% of the variance in RT could be attributed to between student differences. Conversely, 89% of the variance in RT could be attributed to within-student differences. The intercept (B) was 61, meaning that students answered each MCQ of the PT in an average time of 61 ± 1.02 (standard errors - SE), regardless of other variables. Next, we addressed whether/how item type and answer switching predicted RT (model 1). The

Table 2 Distractor analysis

	Item	Key	<u>% of students choosing option</u>					Number of distractors needing revision
			A	B	C	D	E	
AIG	AIG_16	D	0%	35%	15%	-	4%	2
	AIG_19	E	3%	3%	24%	1%	-	3
	AIG_20	C	1%	10%	-	1%	30%	2
	AIG_22	C	5%	7%	-	3%	5%	1
	AIG_34	B	18%	-	1%	3%	0%	3
	AIG_47	E	5%	18%	2%	21%	-	1
	AIG_50	D	6%	22%	8%	-	3%	1
	AIG_51	B	1%	-	5%	3%	0%	3
	AIG_56	C	30%	9%	-	0%	29%	1
	AIG_60	B	2%	-	5%	4%	6%	2
	AIG_62	E	2%	5%	4%	0%	-	2
	AIG_63	A	-	1%	1%	13%	0%	3
	AIG_66	D	37%	41%	4%	-	0%	2
	AIG_67	D	18%	49%	1%	-	0%	2
	AIG_70	E	26%	6%	6%	16%	-	0
	AIG_71	D	1%	0%	0%	-	0%	4
	AIG_76	E	3%	4%	24%	0%	-	3
	AIG_80	B	1%	-	0%	1%	18%	3
	AIG_81	B	7%	-	8%	3%	18%	1
	AIG_83	C	80%	11%	-	7%	0%	1
AIG_86	A	-	1%	1%	8%	65%	2	
AIG_88	E	10%	1%	1%	0%	-	3	
AIG_106	C	0%	0%	99%	0%	0%	4	

Table 2 (continued)

Manually written	X1	A	-	26%	1%	2%	0%	3
	X2	A	-	9%	28%	4%	10%	1
	X10	B	7%	-	1%	4%	0%	3
	X15	D	13%	10%	14	-	20%	0
	X21	A	-	1%	14%	6%	10%	1
	X23	C	0%	0%	-	0%	0%	4
	X24	A	-	16%	23%	0%	4%	2
	X32	A	-	6%	18%	20%	23%	0
	X35	D	3%	5%	10	-	16%	1
	X44	D	0%	0%	1%	-	0%	4
	X45	C	15%	15	-	42%	5%	0
	X49	C	26%	3	-	9%	19%	1
	X52	B	17%	-	1%	1%	66%	2
	X53	A	-	3%	12%	62%	15%	1
	X58	E	13%	35%	7%	4%	-	1
	X77	D	4%	2%	13%	-	11%	2
	X84	D	4%	4%	1%	-	1%	4
	X89	E	42%	28%	8%	8%	-	0
	X93	C	94%	3%	-	1%	1%	3
	X94	D	2%	25%	5	-	6%	1
X97	C	7%	28%	-	22%	24%	0	
X99	A	-	22%	1%	35%	12%	1	
X102	A	-	65%	3%	4%	7%	2	

Note. The results presented correspond to all students. Because there were no consequences for giving inaccurate answers, there were no missing data

Table 3 Hierarchical linear model analysis

	Response time								
	Null Model		Model 1		Model 2		Model 3		
	B	SE	B	SE	B	SE	B	SE	
<i>Intercept</i>	61	1.02	59.9	1.10	59.9	1.05	59.9	1.05	
<i>Student Level</i>									
Number of responses given			15.6**	0.47	15.6**	0.47	15.3*	0.62	
<i>Item Level</i>									
Item type			3.29**	0.77	3.29*	0.78	3.29*	0.78	
<i>Interaction</i>									
Item type * Number of responses given							0.84	0.92	
<i>Residual variance</i>									
Student level		243						223	
Item level		2062						1890	
<i>Explained variance (R²)</i>									
Fixed effects								0.076	
Fixed and random effects								0.19	

Note: SE=Standard error; Item type: 0=Manual, 1=AIG; Null model=intercept only model; Model 1=constrained model; Model 2=Augmented model; Model 3=full model; * p<.05; ** p<.001

model included the predictors mentioned above simultaneously. Results revealed that the fixed effect of both item type and answer switching was positively significant ($B_{\text{switching}}=3.29$; $B_{\text{type}}=15.6$, $p<.001$). The average effect of answer switching for the typical student on the PT was 15.6 s ($SE=0.47$), and the average RT for manually written items was 3.29 s ($SE=0.77$), lower than the average RT for AIG-items. Following this, we built an augmented model (Model 2) considering cluster-specific effects of item type and the overall effect of answer switching. Both predictors were again positively significant of RT, and the coefficients were equal to those of model 1. However, this time we obtained a measure of the differences in the effect (slope) of item type (level 1-effect) on RT between the students (level-2 units). The average deviation of the item type effect of a student from the average effect was about 1.89 s. We then compared the deviance of models 1 and 2 to test whether including the between-student variation of the item type-effect improved the estimation using the likelihood-ratio test (LR χ^2) and the Akaike Information Criterion (AIC) statistic. The p-value of the LR χ^2 was below 0.20 ($p=.069$) and Model 2 had a lower AIC than Model 1 ($AIC_{M1}=132,377$; $AIC_{M2}=132,376$). These values are proof that estimating the variance and covariance terms of item type improved the fit (Sommet & Morselli, 2017), the reason why we decided to include it in the final model (Model 3). Fixed and random effects included in model 3 explained 19% of the variance in the final model. The coefficient estimates of both item type and answer switching didn't vary compared to the other models. In model 3 we also included cross-level interactions between answer switching and item type. The coefficient estimate of this cross-level interaction was 0.836 ($p=.367$). This means that the pooled within-student effect of switching answers was not statistically significant between both item types. After considering this data, we discover that the response processes for both AIG and manually written items appear to be pretty similar, which is a source of validity evidence for automatically generated items.

4 Discussion

4.1 Overview

There is a high need for numerous and valid MCQs due to changes in student assessment brought about by new computer-based exam formats that traditional item construction methods cannot keep pace (Arendasy & Sommer, 2012). It is, therefore, necessary to extend classic assessment methods and psychometric procedures so that it covers modern procedures/techniques (von Davier et al., 2021). AIG is a cutting-edge approach to item development and management that combines cognitive and psychometric theories for futuristic assessment services in digital contexts (Choi & Zhang, 2019; Falcão et al., 2023). The study of AIG is a promising pursuit since it enables computer technology to produce many new items (Hommel et al., 2022).

In the present paper, we provide a fresh approach for developing material for MCQ FA's modalities, such as PTs, using AIG principles and practices. 23 AIG-items were included in a medicine PT along with traditionally developed MCQs. AIG-items were generated using cognitive models outlined in advance of the PT. Data

obtained was analysed from an RM, distractor performance and HLM perspective. First, the psychometric analysis of the PT demonstrated good internal consistency and test reliability, revealing that the exam was a suitable tool for evaluating students' θ . Additionally, the PT presented unidimensionality and the questions that composed the exam were locally independent and presented a monotonic relationship between the scores and values of θ . This means that both AIG and manually written items assessed the same construct. Neither question was redundant, and using true scores in place of θ scores was justified.

In addition to fitting the RM, the AIG-items have successfully fit the PT and proven to be on par with the manually written items in terms of quality. At the item level, psychometric analysis of the PT suggested several strengths of the AIG-items as well as some opportunities for future improvement. AIG-items revealed a broad spread of b_i (which demonstrates more adaptability and possibly more discriminating power), with only 3 items (AIG_71; AIG_83; AIG_106) located at both ends of the b_i continuum. These extreme positions indicate that the b_i of these 3 items was either too simple/too challenging for the students to answer, revealing the need for revision. The remaining AIG-items provided a comprehensive assessment across a wide range of the underlying θ , providing reliable information for sorting students with relatively moderate levels of knowledge. Among the 23 AIG items included in the exam, only the 3 items mentioned above captured very high or very low values of θ , offering little information about the students' knowledge.

Item bias was evaluated via DIF. The RM found evidence for curricular-plan uniform DIF for only 1 AIG-item. The possible existence of DIF in AIG-items may sound problematic for some. The reason for this concern relates essentially to the solution traditionally adopted by subject matter experts to deal with items that present this phenomenon in achievement testing, which consists of their omission or replacement with alternative items from a larger item bank (Silvia et al., 2021). Concurrently, scholars have been questioning the value of analysing DIF in educational measures in recent years, claiming that tests of cognitive ability and educational achievement are not test biased and produce results comparable to test performance. These items are professionally developed to evaluate educational achievement and are subject to extensive reviews before being released, which is why they should not be considered biased towards the test. However, DIF will always exist, and its sources will always be uncertain because there are too many interrelated variables, which is why we believe that these outcomes do not call into question the quality of the AIG-items (Teresi & Fleishman, 2007).

Distractor performance was slightly better, though non significant, for the manually written items, with 68% of them being functional. Approximately 61% of the distractors of the AIG-items used in the PT properly functioned. These outcomes are favorable for the use of AIG. Nevertheless, a significant number of distractors (39%) call for possible revisions. Our results are in line with the literature focusing on the plausibility of AIG item distractors (e.g., Gierl and Lai, 2013; Lai et al., 2016), which appear less plausible than handwritten items and refer to the need of a clear methodology for the generation of distractors. Despite these issues, we must consider that this same research line claims that AIG distractors can already distinguish between

low-and-high-performing examinees, which is a definite sign that AIG distractors can still be improved and made even more effective.

The qualitative outcomes attained using expert review were partially consistent with the quantitative evaluation procedures that were conducted in this paper. Both points of view agree that the AIG-items were of high quality. However, the qualitative viewpoint used here went further, claiming that AIG-items reflected accepted item-writing principles and assigning slightly higher quality scores to automatically generated items over those manually written. This finding is interesting and merits additional investigation in future studies. It also serves as a potential starting point for a review of the AIG items' content validity. Since test developers have at least a general understanding of the object they intend to assess, the more clearly these objects are expressed, the more precisely content validity may be examined (Beck, 2020). By conducting this assessment, we are able to identify the most significant defects (in terms of item writing) that these items harbor and what the most important fixes ought to be. In this case, we realized that these defects consist, in particular, of the presence of a possible technical flaw, the possible absence of a vignette, or the absence of a single-answer format.

Finally, the validity of the AIG-items was evaluated using HLM. Since the research focused specifically on the validity of AIG-items is practically non-existent, it is not surprising that the validity of these items still falls into a murky area of the literature. However, along with the affording logistical benefits mentioned, AIG can indeed play a role in building a validity case (Colvin et al., 2016). Points in favor of this argument relate to possible evidence of content validity (Hommel et al., 2022) and construct validity (Harrison et al., 2017). In this paper we decided to go further. The method employed here for developing this argument is novel as we have gathered evidence supporting the validity of AIG in terms of response processes from a modern psychometrics standpoint (American Educational Research Association, 2018). We found that the RT for the manually written items as for the AIG-items varied by only 3 s. This brief time difference, in our opinion, illustrates how similar these item types are. Since RT provides an explicit, although indirect information about the complexity of the item (and, therefore, the amount or degree of processing involved) (Padilla & Benítez, 2014), we may partially claim that the response processes/processing mechanisms used to answer both item types are very similar, at least in terms of how much processing is done or how students interpret the questions (Deng et al., 2021). To materialize this evidence, the RT to the PT exam questions was analyzed considering both answer switching and item type. We discovered that students performed fairly similarly while responding to both item types utilized in the PT since there was no interaction between the number of responses altered and the item type. With the RT being equiprobable with the responses altered between both item types, these data complement one another and represent evidence for AIG-items' validity, strengthening any validity argument for this procedure.

4.2 Practical implications for FA: how should feedback be delivered?

AIG pledges to provide test items and sizable item banks rapidly. Since it is based on the cognitive processes that underlie task performance, it prevents item exposure, and

exhibits construct validity (Harrison et al., 2017). These characteristics appeal to SA but are also notably helpful to FA regarding the quantity of testing items available. A potential concern of using CFT with an AIG functionality may be based on how feedback can be delivered within a bank of test items. At first glance, it may come to mind that feedback delivered by these systems could only be given based on achievement standards.

Consequently, students using these platforms could learn by becoming familiar with key elements (through item repetition) or coming across more challenging items (Choi et al., 2018). Although useful, this approach seems simplistic and amenable to improvement, as students should receive real feedback on how to solve each testing problem (Gierl & Lai, 2018). The available literature offers some suggestions on how AIG can be used to deliver feedback. Gierl and Lai (2018) described a method for generating both the items and the rationales required to solve testing tasks within FA in medical education. According to the authors, rationale generation could be incorporated into the three-step AIG process by expanding the item model in the second step and identifying key features in the task required to solve the items. Xinxin (2019) presented a modified generation framework that employs a tree structure for cognitive modelling, an assembly mechanism, and a validation tool to support CFT within the context of HEI. In a process known as a tree traversal, elements that are related by nodes and edges can be automatically and logically searched for and merged, producing test cases and the corresponding feedback. These methods of providing feedback highlight AIG's adaptability and versatility. For more information on this matter, refer to the original papers.

4.3 Strengths, limitations and directions for future research

This paper's main contribution is a comprehensive psychometric/statistical analysis that compares the quality/validity of AIG-items versus those written manually within a PT of medicine in a HEI. It also reviews issues pertinent to educational assessment and explains how AIG can be used to complement CFT systems. The inventiveness of this research is another benefit. As far as we know, no research has been done that examines explicitly how AIG-items perform in FA assessments such as PT or the validity of these items using quantitative and qualitative procedures such as the ones used here.

However, this paper has some drawbacks. It is important to note that the sample of AIG-items used in the PT is notoriously limited compared to the number of manual written items. To compare the exam's questions more accurately, there should be a more evenly distributed distribution of AIG elements. However, due to school norms, adding more AIG-items to the PT was impossible, as this method is currently at an experimental stage. Additionally, one must consider that the number of respondents who took the test sample prevented us from using a more reliable IRT model that would have allowed us to evaluate other parameters of the AIG-items besides b_p , such as discrimination or guessing. Evaluating these parameters, one could obtain a thorough picture of how these items performed. (Gierl & Lai, 2018). Finally, one should note that only one test developer expert conducted the qualitative review of the items included in the PT. This constitutes a limitation of our work, as analyses of

this type provide solid robust results when more than one expert performs this evaluation, in order to assess the degree of agreement between raters. However the expert is a highly trained item writer and trainer in item writing and was blind to the the nature of items. Future avenues should work with more extensive assessment panels to qualitative evaluate/compare the quality of both AIG and manually written items.

The mass manufacturing, intelligent item calibration and management, learner-centered evaluation, and other elements of AIG are quite varied and promise to revolutionize educational measurement. However, a comprehensive validation of these assertions in educational settings such as FA has not yet been achieved (Choi et al., 2018). There is a clear need for creating feedback guidelines within AIG frameworks (Gierl & Lai, 2018; Xinxin, 2019). More applications with AIG-items in PTs should be run in order to get reliable results regarding the use of automatically generated items in such tests. Research in this area would be beneficial for spreading AIG in the context of FA.

5 Conclusion

AIG-items represent suitable material for evaluating student's knowledge, even on FA modalities such as PT. Despite being computer-generated, these items are valid, present psychometrical quality and are most advantageous in terms of production speed and quantity. Additionally, they have a superior quality as assess by item-writing experts. These capabilities are expected to ease the item development burden, resulting in significant cost savings for educational institutions when developing test items.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10639-023-12014-x>.

Acknowledgements The authors acknowledge the Portuguese Foundation for Science and Technology (FCT) and Icoignitus4All—It Solutions, Lda for providing a doctoral fellowship to FF (Grant no. PD/BDE/150621/2020), through the PhDiHES program.

Funding Open access funding provided by FCT|FCCN (b-on). This study was funded by National funds, through the FCT—project UIDB/50026/2020 and UIDP/50026/2020.

Data availability The dataset analysed during the current study is not publicly available but is available from the corresponding author on reasonable request.

Declarations

Conflict of interest None.

Ethics approval We complied with the privacy policy of the University of Minho's School of Medicine's. All data were anonymised and treated with the utmost care.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative

Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Albanese, M., & Case, S. M. (2016). Progress testing: Critical analysis and suggested practices. *Advances in Health Sciences Education*, 21(1), 221–234. <https://doi.org/10.1007/s10459-015-9587-z>.
- American Educational Research Association. (2018). *Standards for educational and psychological testing*. American Educational Research Association.
- Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences*, 22(1), 112–117. <https://doi.org/10.1016/j.lindif.2011.11.005>.
- Beck, K. (2020). Ensuring content validity of Psychological and Educational tests—the role of experts. *Frontline Learning Research*, 8(6), 1–37.
- Bejar, I. I. (2012). Item generation: Implications for a validity argument. In T. M. H. Mark J Gierl (Ed.), *Automatic item generation: Theory and practice* (pp. 50–66). Routledge. <https://doi.org/10.1163/ej.9789004172067.i-752.40>.
- Bijol, V., Byrne-dugan, C. J., & Hoenig, M. P. (2015). Medical student web-based formative assessment tool for renal pathology. *Medical Education Online*, 20(1), 1–7.
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment Evaluation and Accountability*, 21(1), 5–31. <https://doi.org/10.1007/s11092-008-9068-5>.
- Boston, C. (2002). The concept of formative assessment. *Practical Assessment, Research, and Evaluation*, 8(1), 9.
- Campbell, I. (2007). Chi-squared and Fisher–Irwin tests of two-by-two tables with small sample recommendations. *Statistics in Medicine*, 26(19), 3661–3675.
- Cecilio-Fernandes, D., Kerdijk, W., Bremers, A. J., Aalders, W., & Tio, R. A. (2018). Comparison of the level of cognitive processing between case-based items and non-case-based items on the Interuniversity Progress Test of Medicine in the Netherlands. *Journal of Educational Evaluation for Health Professions*, 15.
- Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48, <https://doi.org/10.18637/jss.v048.i06>.
- Choi, J., & Zhang, X. (2019). Computerized item modeling Practices using computer adaptive formative Assessment Automatic Item Generation System: A Tutorial. *The Quantitative Methods for Psychology*, 15(3), 214–225. <https://doi.org/10.20982/tqmp.15.3.p214>.
- 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>.
- Chowdhury, R. R., & Kalu, G. (2004). Learning to give feedback in medical education. *The Obstetrician & Gynaecologist*, 6(4), 243–247. <https://doi.org/10.1576/toag.6.4.243.27023>.
- Colvin, K. F., Keller, L. A., & Robin, F. (2016). Effect of Imprecise Parameter estimates on ability estimates in a multistage test in an automatic item generation context. *Journal of Computerized Adaptive Testing*, 4(1).
- Couto, L. B., Durand, M. T., Wolff, A. C. D., Restini, C. B. A., Faria, M., Romão, G. S., & Bestetti, R. B. (2019). Formative assessment scores in tutorial sessions correlates with OSCE and progress testing scores in a PBL medical curriculum. *Medical Education Online*, 24(1), <https://doi.org/10.1080/108010872981.2018.1560862>.
- De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44(1), 109–117.
- Demuth, R. H., Gold, J. G., Mavis, B. E., & Wagner, D. P. (2018). Progress on a New Kind of Progress Test: Assessing medical students' clinical skills. *Academic Medicine*, 93(5), 724–728. <https://doi.org/10.1097/ACM.0000000000001982>.

- Deng, J. M., Streja, N., & Flynn, A. B. (2021). Response process validity evidence in chemistry education research. *Journal of Chemical Education*, 98(12), 3656–3666.
- Desjardins, C., & Bulut, O. (2018). *Handbook of Educational Measurement and Psychometrics using R*. CRC Press.
- Dijksterhuis, M. G. K., Scheele, F., Schuwirth, L. W. T., Essed, G. G. M., Nijhuis, J. G., & Braat, D. D. M. (2009). Progress testing in postgraduate medical education. *Medical Teacher*, 31(10), <https://doi.org/10.3109/01421590902849545>.
- Dorans, N. J., & Holland, P. W. (1992). DIF detection and description: Mantel-Haenszel and standardization 1, 2. *ETS Research Report Series*, 1992(1), i–40.
- Falcão, F., Costa, P., & Pêgo, J. M. (2022). Feasibility assurance: A review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 1–21.
- Falcão, F., Pereira, D. M., Gonçalves, N., De Champlain, A., Costa, P., & Pêgo, J. M. (2023). A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation. *Advances in Health Sciences Education*. <https://doi.org/10.1007/s10459-023-10225-y>.
- Farrell, G., & Leung, Y. K. (2004). Innovative Online Assessment using confidence measurement. *Education and Information Technologies*, 9(1), 5–19. <https://doi.org/10.1023/b:eait.0000024258.29560.3c>.
- Gierl, M., & 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. (2013). 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., & Lai, H. (2015). Using automated processes to generate test items and their associated solutions and rationales to support formative feedback. *Interaction Design and Architecture(s)*, 25(1), 9–20.
- Gierl, M., & 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., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice test items. *Medical Education*, 46(8), 757–765. <https://doi.org/10.1111/j.1365-2923.2012.04289.x>.
- Gierl, M. J., Bulut, O., Guo, Q., & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87(6), 1082–1116.
- Görlich, D., & Friederichs, H. (2021). Using longitudinal progress test data to determine the effect size of learning in undergraduate medical education—a retrospective, single-center, mixed model analysis of progress testing results. *Medical Education Online*, 26(1), <https://doi.org/10.1080/10872981.2021.1972505>.
- 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>.
- Harrison, P. M. C., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, 7(1), 1–18. <https://doi.org/10.1038/s41598-017-03586-z>.
- Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement*, 71(4), 732–746. <https://doi.org/10.1177/0013164410390032>.
- Hommel, B. E., Wollang, F. J. M., Kotova, V., Zacher, H., & Schmukle, S. C. (2022). Transformer-based deep neural Language modeling for Construct-Specific Automatic Item Generation. *Psychometrika*, 87(2), 749–772. <https://doi.org/10.1007/s11336-021-09823-9>.
- Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.
- Huble, A. M. (2021). *Response processes validity evidence: Understanding the meaning of scores from psychological measures*.
- Irons, A. (2007). Enhancing learning through formative assessment and feedback. In *Enhancing Learning through Formative Assessment and Feedback*. <https://doi.org/10.4324/9780203934333>.
- Jendryczko, D., Berkemeyer, L., & Holling, H. (2020). Introducing a Computerized Figural Memory Test based on Automatic Item Generation: An analysis with the Rasch Poisson counts Model. *Frontiers in Psychology*, 11(June), 1–16. <https://doi.org/10.3389/fpsyg.2020.00945>.

- Joyce, P. (2018). The effectiveness of online and paper-based formative Assessment in the learning of English as a Second Language. *PASAA: Journal of Language Teaching and Learning in Thailand*, 55(June), 126–146.
- Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, 77(2), 156–161. <https://doi.org/10.1097/00001888-200202000-00016>.
- Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., & Baumert, J. (2008). Engagement and emotional exhaustion in teachers: Does the school context make a difference? *Applied Psychology*, 57, 127–151.
- Knight, P. T. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education*, 27(3), 275–286. <https://doi.org/10.1080/03075070220000662>.
- Koşan, M., Koç, N., Elhan, A., & Öztuna, D. (2019). Developing an Item Bank for Progress tests and application of computerized adaptive testing by Simulation in Medical Education. *International Journal of Assessment Tools in Education*, 6(4), 656–669. <https://doi.org/10.21449/ijate.635675>.
- 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>.
- Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, 2(3), 151–160.
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package ‘lmerTest’. *R Package Version*, 2(0), 734.
- 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., Philippe, & de Champlain (2016b). André. 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>.
- Leenknecht, M., Wijnja, L., Köhler, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: The role of students’ motivation. *Assessment and Evaluation in Higher Education*, 46(2), 236–255. <https://doi.org/10.1080/02602938.2020.1765228>.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 50–60.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748.
- 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>.
- Mitra, N. K., & Barua, A. (2015). Effect of online formative assessment on summative performance in integrated musculoskeletal system module. *BMC Medical Education*, 15(1), 1–7. <https://doi.org/10.1186/s12909-015-0318-1>.
- Nwafor, A., & Onyenwe, E. (2021). An automated multiple-choice question generation using Natural Language Processing techniques. *International Journal on Natural Language Computing*, 10(02), 1–10. <https://doi.org/10.5121/ijnlc.2021.10201>.
- Olson, B. L., & McDonald, J. L. (2004). Influence of Online Formative Assessment upon Student Learning in Biomedical Science Courses. *Journal of Dental Education*, 68(6), 656–659. <https://doi.org/10.1002/j.0022-0337.2004.68.6.tb03783.x>.
- Öztürk, Z. (2013). *A Bayesian Approach for Item Response Theory in Assessing the Progress Test in Medical Students*.
- Padilla, J. L., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema*, 26(1), 136–144. <https://doi.org/10.7334/psicothema2013.259>.
- Patrick, A., Hatzinger, R., Maier, M. J., Rusch, T., & Mair, M. P. (2018). *Package ‘eRm’*.
- Prasetyo, S. E., Adjti, T. B., & Hidayah, I. (2020). Automated Item Generation: Model and Development Technique. *7th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2020 - Proceedings*, 64–69. <https://doi.org/10.1109/ICITACEE50144.2020.9239243>.
- Prashanti, E., & Ramnarayan, K. (2019). Ten maxims of formative assessment. *Advances in Physiology Education*, 43(2), 99–102. <https://doi.org/10.1152/advan.00173.2018>.
- Preston, R., Gratani, M., Owens, K., Roche, P., Zimanyi, M., & Malau-Aduli, B. (2020). Exploring the impact of Assessment on Medical Students’ Learning. *Assessment and Evaluation in Higher Education*, 45(1), 109–124. <https://doi.org/10.1080/02602938.2019.1614145>.

- 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>.
- Rasch, G. (1960). Probabilistic models for some intelligence and attainment test. *Studies in mathematical psychology* (1 vol.). Danmarks Paedagogiske Institut.
- Revelle, W., & Revelle, M. W. (2015). Package ‘psych.’ In *The comprehensive R archive network* (pp. 337–338.).
- 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.
- Russell, L. B., & Hubley, A. M. (2017). Some thoughts on gathering response processes validity evidence in the context of online measurement and the digital revolution. *Understanding and investigating response processes in validation research* (pp. 229–249). Springer.
- Rzasa, S. E. (2002). *Item analysis on a developmental rating scale using both statistical and qualitative methods*. Pennsylvania State University.
- Schüttpelz-Brauns, K., Karay, Y., Gehlhar, K., Arias, J., & Zupanic, M. (2020). Comparison of the evaluation of formative assessment at two medical faculties with different conditions of undergraduate training, assessment and feedback. *GMS Journal for Medical Education*, 37(4), 1–23. <https://doi.org/10.3205/zma001334>.
- Shea, T., Cooper, B. K., de Cieri, H., & Sheehan, C. (2012). Evaluation of a perceived organisational performance scale using Rasch model analysis. *Australian Journal of Management*, 37(3), 507–522. <https://doi.org/10.1177/0312896212443921>.
- Silvia, P. J., Rodriguez, R. M., Beaty, R. E., Frith, E., Kaufman, J. C., Loprinzi, P., & Reiter-Palmon, R. (2021). Measuring everyday creativity: A Rasch model analysis of the Biographical Inventory of Creative Behaviors (BICB) scale. *Thinking Skills and Creativity*, 39, 100797.
- Socha, A., DeMars, C. E., Zilberberg, A., & Phan, H. (2015). Differential Item Functioning Detection with the Mantel-Haenszel Procedure: The Effects of matching types and other factors. *International Journal of Testing*, 15(3), 193–215. <https://doi.org/10.1080/15305058.2014.984066>.
- Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using Stata, R, Mplus, and SPSS. *International Review of Social Psychology*, 30, 203–218.
- Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, 57(8), 1358–1362.
- Teresi, J. A., & Fleishman, J. A. (2007). Differential item functioning and health assessment. *Quality of Life Research*, 16(SUPPL. 1), 33–42. <https://doi.org/10.1007/s11136-007-9184-6>.
- Tor, E., & Steketee, C. (2011). Rasch analysis on OSCE data: An illustrative example. *Australasian Medical Journal*, 4(6), 339–345. <https://doi.org/10.4066/AMJ.2011.75>.
- von Davier, A. A., Mislevy, R., & Hao, J. (2021). *Computational psychometrics: New methodologies for a new generation of digital learning and assessment: With examples in R and Python*, (10Doi, 10 vol., pp. 973–978). Berlin: Springer.
- Watling, C. J., & Ginsburg, S. (2019). Assessment, feedback and the alchemy of learning. *Medical Education*, 53(1), 76–85. <https://doi.org/10.1111/medu.13645>.
- Wind, S. A., Alemdar, M., Lingle, J. A., Moore, R., & Asilkalkan, A. (2019). Exploring student understanding of the engineering design process using distractor analysis. *International Journal of STEM Education*, 6(1), <https://doi.org/10.1186/s40594-018-0156-x>.
- Xinxin, Z. (2019). *Using Automatic Item Generation to Create Content for Computerized Formative Assessment (Doctoral dissertation)*. University of Alberta, Edmonton, Canada.

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

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Filipe Manuel Vidal Falcão^{1,2,3} · **Daniela S.M. Pereira**^{1,2,3} · **José Miguel Pêgo**^{1,2,3} · **Patrício Costa**^{1,2}

✉ Filipe Manuel Vidal Falcão
fmvf92@gmail.com

¹ Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Largo do Paço, Braga 4770-075, Portugal

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

³ iCognitus4All – IT Solutions, Braga, Portugal