

Universidade do Minho
Escola de Medicina

Filipe Manuel Vidal Falcão

Unveiling Automatic Item Generation: Efficiency, Formation, Proficiency and Differential Item Functioning in Medical Education Assessment

Março de 2024

Unveiling Automatic Item Generation: Efficiency, Formation,Proficiency and Differential Item Functioning in Medical Education Assessment

Filipe Manuel Vidal Falcão

UMinho | 2024

Universidade do Minho
Escola de Medicina

Filipe Manuel Vidal Falcão

Unveiling Automatic Item Generation: Efficiency, Formation, Proficiency and Differential Item Functioning in Medical Education Assessment

Tese de Doutoramento
Doutoramento em Ciências da Saúde

Trabalho efetuado sob a orientação do
Professor Doutor Patrício Costa
e do
Professor Doutor José Miguel Pêgo

Março de 2024

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.
Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.
Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho



Atribuição-NãoComercial-SemDerivações

CC BY-NC-ND

https://creativecommons.org/licenses/by-nc-nd/4.0/

ACKNOWLEDGEMENTS / AGRADECIMENTOS

"Mountains, according to the angle of view, the season, the time of day, the beholder's frame of mind, or any one thing, can effectively change their appearance. Thus, it is essential to recognize that we can never know more than one side, one small aspect of a mountain."

— **Haruki Murakami**, A Wild Sheep Chase.

I can only thank everyone who helped me reach the summit of this mountain.

To **my parents**, for their unwavering belief in me. Even if the world ever doubts me, I know I can always come home; To **Beatriz Santos**, for all your care and love. As long as the music of life plays, I want to dance with you forever; To **Professor Patrício Costa**, for the life lessons, friendship and for changing classes at the last minute; To **Professor José Miguel Pêgo**, for all the support throughout my project; To **Dra. Carla Santos**, for lifting the lid on a deep, dark well; To **Daniela Pereira**, for sailing with me through the turbulent waters of this nebulous journey (we've finally reached land, my friend); To my former neighbor **Miguel Ribeiro**, for all the Champions League nights (maybe one day you'll understand the feeling, old friend); To **Andy Puddicombe**, for reminding me that there's always a blue sky; To Eve, Kessonga, Dora, Sam and the rest of the **Headspace** team. Even from a distance, thank you for keeping me company and for "reminding" me to pause and breathe; To **all my friends** (the old ones and those who have joined me on this journey), I hope we continue to create memories together.

The work presented in this thesis was performed in the Life and Health Sciences Research Institute (ICVS), at the School of Medicine, University of Minho. Financial support was secured by a PhD Grant (PD/BDE/150621/2020 to Filipe Manuel Vidal Falcão) provided by the Foundation for Science and Technology (FCT) and Icognitus4All—IT Solutions through the PhDiHES program. The work was supported by National funds, through the FCT—project UIDB/50026/2020 and UIDP/50026/2020.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Revelando a Geração Automática de Itens: Eficiência, Formação, Proficiência e Funcionamento Diferencial dos Itens na Avaliação da Educação Médica.

Resumo

A avaliação de conhecimentos está a mudar rapidamente, sobretudo na educação médica, onde testes realizados em computador com perguntas de escolha múltipla são cada vez mais frequentes. Embora esta transição ofereça benefícios em termos de eficiência e acessibilidade, também levanta preocupações quanto à segurança dos testes devido ao risco de exposição das perguntas. Consequentemente, as perguntas dos testes devem ser elaboradas mais rapidamente e em maiores quantidades. Neste contexto, a Geração Automática de Itens (AIG) surge como uma solução promissora. Esta envolve o uso de algoritmos de computador e modelos cognitivos para gerar automaticamente perguntas de teste. Para além de oferecer uma abordagem sistemática e económica para o desenvolvimento de itens, a AIG permite a criação rápida de bancos substanciais de itens. Apesar do seu potencial, a AIG continua a ser subutilizada em muitas escolas médicas, o que realça a necessidade de mais investigação. A presente Tese procura esclarecer as complexidades do processo da AIG no contexto da avaliação da educação médica, com foco em vários aspetos-chave. Estes incluem a qualidade dos itens gerados e as suas propriedades psicométricas, a facilidade de integração nas práticas de avaliação, a validade da AIG e a sua compatibilidade com diferentes modalidades de avaliação. Ao longo desta Tese, apresentamos evidência empírica que destaca a rapidez, qualidade, validade e facilidade de uso deste procedimento. Ao oferecer exemplos e perspetivas específicas, esperamos demonstrar como a AIG pode melhorar significativamente as práticas de avaliação na educação médica, ao mesmo tempo que proporciona economias substanciais de recursos às instituições.

Palavras-chave: Avaliação, Escrita de itens, Geração Automática de Itens, Qualidade dos itens; Validade

Unveiling Automatic Item Generation: Efficiency, Formation, Proficiency and Differential Item Functioning in Medical Education Assessment.

Abstract

Rapid changes are occurring in knowledge assessment, particularly in medical education, where computer-based exams with multiple-choice questions are becoming more frequent. While this shift provides advantages in terms of efficiency and accessibility, it also raises issues regarding test security due to the danger of items being exposed. As a result, test questions must be developed more swiftly and in larger quantities. Automatic Item Generation (AIG) emerges as a promising solution in this setting. It involves the use of computer algorithms and cognitive models to automatically generate test questions. Aside from providing a systematic and economical approach to item development, AIG enables the rapid creation of tangible item banks. However, despite its potential, AIG is still underutilized at many medical schools, highlighting the need for additional research. The purpose of this thesis is to explore the intricacies of the AIG process within the context of medical education assessment, focusing on several key aspects. These include the quality of generated items and their psychometric properties, the ease of integrating AIG into current assessment practices, the validity of AIG, and its compatibility with different assessment modalities. Throughout this Thesis, we present empirical evidence that underscores the procedure's speed, quality, validity, and ease of use. By offering specific examples and insights, we hope to demonstrate how AIG can significantly enhance evaluation practices in medical education while also delivering substantial cost savings to institutions.

Keywords: Automatic Item Generation, Assessment, Item writing, Item quality, Validity

TABLE OF CONTENTS

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros	ii
Acknowledgments	iii
Statement Of Integrity	iv
Resumo	v
Abstract	vi
Table Of Contents	vii
List Of Abbreviations	xiii
List Of Figures	xiii
List of Tables	xv
Thesis Planning	xvi
CHAPTER I: UNVEILING AUTOMATIC ITEM GENERATION	1
Introduction: A new educational order	2
Assessment in Medical Education and the Item Development Burden	2
A Next-Gen Augmented Intelligence Approach: Automatic Item Generation	4
Practical Example	7
Advantages and Limitations: AIG vs Traditional Item Development Methods	10
Different assessment systems, same problem?	12
Statement of the problem and thesis aims	16
References	17
CHAPTER II: THE ROLE OF TECHNOLOGY IN ASSESSMENT QUALITY ASSURANCE PROCESSES	24
Orientation to the chapter	25
Introduction—The Increasing Role of Technology in Medicine and Medical Education	25
Computer-Based Testing	26
Automated Item Generation	28
How Technology Is Impacting Foundational Tasks in Assessment	29
Test Design	29
Test Form Assembly	33

Scoring Of Traditional and Novel Items 34
How Technology Is Impacting Operations 35
Testing Modality (Adaptive Testing Models) 35
Technology In Assessment—Prognostications and Confabulations 38
Conclusion 40
References 40

CHAPTER III: FEASIBILITY ASSURANCE: A REVIEW OF AUTOMATIC ITEM GENERATION IN MEDICAL ASSESSMENT 44

Abstract 45
Literature Review 46
Educational Measurement Is Evolving: The Medical Education Case 46
A Review of a Next-Gen Assessment Theory: Automatic Item Generation 47
Method 48
Quality Assessment of The Review 48
Sources Of Information 48
Data Extraction 49
Selection Criteria 49
Exclusion Criteria 49
Screening Phase 49
Quality Assessment of Included Studies 50
Validity Framework 50
Results 53
Description of Included Studies 53
Data Synthesis 53
Data Synthesis: AIG's Feasibility 53
Time-spent 53
Number of Items Generated 56
Ease of Learning 56
Data Synthesis: Quality of MCQs Generated by AIG 56
AIG vs Traditional Item Development Methods 56
Psychometric Analysis of Items Generated by AIG 57

Revisions of the Cognitive Model 57
Validity Assessment 57
Discussion 60
Directions for Future Research 62
Conclusions 63
References 63

CHAPTER IV: A SUGGESTIVE APPROACH FOR ASSESSING ITEM QUALITY, USABILITYAND VALIDITY OF AUTOMATIC ITEM GENERATION 66

Abstract 67
Introduction 68
Background 68
The Role of Multiple-Choice Questions in Medical Education 68
Automatic Item Generation as a Solution 69
AIG versus Manual Item Writing 70
A Top-down, Strong Theory Approach 72
Methods 72
General Procedure 72
Software 73
Study I 73
Sample 73
Procedure 73
Statistical Analysis 74
Results 75
Study II 76
Data Collection and Sample 76
Measure 77
Procedure 77
Analytical Approach 77
Results 78
Discussion 79
Item Quality 79

Usability 80
Validity 80
Implications in the Context of Education, Assessment and Test Development 81
Strengths and Limitations 82
Directions for Future Research 82
Conclusion 83
Appendix A: Learnability Analysis 83
Appendix B: IRT Assumptions – Strategies and Results 84
Appendix C: Model-Data Fit 85
Appendix D: RM Parameter Estimates 85
Appendix E: Item Information Curves 86
References 88

CHAPTER V: ITEM RESPONSE THEORY IN PSYCHOMETRIC RESEARCH: HOW TO ASSESS THE RELATIONSHIP BETWEEN ITEM PROPERTIES, RESPONDENTS, AND LATENT TRAITS 92

Abstract 94
Project Overview and Context 94
Research Design 95
Overview 95
Why Use IRT to Conduct Validation of Assessment Tests? 96
Step 1. Model Selection 101
Step 2. IRT Assumptions 101
Step 3. Model, Person and Item Diagnostics 102
Step 4. Item Calibration 102
Step 5. Reliability Analysis 102
Step 6. Ability Score Estimation 102
Research Practicalities 103
Where Should We Conduct Such an Analysis? 103
How do you Test Assumptions of an IRT Analysis? 103
What About Model Diagnostics? 104
Estimation of Item Parameters: Which Values Should We Pay Attention to? 105

In Need of a Visual Guide: How to Interpret the Graphical Outputs from IRT 105
Ability Score Estimation 113
Method in Action 114
Practical Lessons Learned 115
Keep your Code Clean 115
Importance of Testing Assumptions 115
Item Performance 116
Know How to Visually Interpret your Graphs 116
Conclusion 117
Further Reading 121
Web Resources 122
References 122

CHAPTER VI: PROGRESS IS IMPOSSIBLE WITHOUT CHANGE: IMPLEMENTING AUTOMATIC ITEM GENERATION IN MEDICAL KNOWLEDGE PROGRESS TESTING. 126

Abstract 127
Introduction 128
Background 128
Assessment of Learning or Assessment for Learning? 128
Making Progress: Formative Assessment Within Medical Education 129
The Underlying Need for a Next-Gen Assessment Theory: Automatic Item Generation 130
Implementing AIG in Online PT 131
Statistical and Psychometric Methods 132
Study Design 132
Data Collection and Sample 132
Procedure 132
Psychometric Approach 132
Qualitative Assessment (Expert Review) 134
Validity Assessment (Hierarchical Linear Modeling) 134
Software 135
Results 135

Item Calibration: Student and Question Comparison 135
Reliability 137
Quantifying DIF 139
Distractor Analysis 140
Qualitative Assessment (Expert Review) 140
Validity Assessment (HLM) 140
Discussion 143
Overview 143
Practical Implications for FA: How Should Feedback be Delivered? 145
Strengths, Limitations and Directions for Future Research 146
Conclusion 147
References 148

CHAPTER VII: GENERAL DISCUSSION 153

Overview 154
Theoretical Approach - Act 1: Technology-based assessment procedures in medical education 154
Theoretical Approach - Act 2: State of things 155
Empirical Testing 157
Compatibility with different assessment methods 158
Future Perspectives and Practical Implications 160
Conclusion 164
References 164

LIST OF ABBREVIATIONS

$	United States Dollar
€	Euro
AIG	Automatic Item Generation
CAT	Computerized Adaptive Testing
CBT	Computer-Based Test
COVID	Coronavirus Disease
GMAT	Graduate Management Achievement Test
IRT	Item Response Theory
MCQ	Multiple-Choice Question
MCCQE I	Medical Council of Canada Qualifying Exam Part I
NCLEX-RN	National Council Licensure Examination for Registered Nurses
PT	Progress Test
SME	Subject Matter Experts

LIST OF FIGURES

CHAPTER I: UNVEILING AUTOMATIC ITEM GENERATION

Figure 1. Three-stage process for generating medical MCQs based on AIG 6

Figure 2. Cognitive model structure 8

Figure 3. Item model structure. 10

CHAPTER II: THE ROLE OF TECHNOLOGY IN ASSESSMENT QUALITY ASSURANCE PROCESSES

Figure 7-1. Top: Example of a Cognitive Model and Respective Item Model. Bottom: Two Examples of Items Derived from the Cognitive Model 39

CHAPTER III: FEASIBILITY ASSURANCE: A REVIEW OF AUTOMATIC ITEM GENERATION IN MEDICAL ASSESSMENT

Figure 1. AIG Three-Step Process for Generating Medical MCQs 47

Figure 2. Flow Chart of the Included Studies 49

CHAPTER IV: A SUGGESTIVE APPROACH FOR ASSESSING ITEM QUALITY, USABILITY AND VALIDITY OF AUTOMATIC ITEM GENERATION

Figure 1. Three-Step Process for Generating Medical MCQs based on AIG 69

Figure 2. Sample of Items Used for Analysis 74

Figure 3. Test Information Function with Conditional Standard Error of Measurement 78

Figure 4. ICCs for the Twenty-Seven Items Composing the General and Vascular Surgery Strand 84

Figure 5. Person Fit Histogram 85

Figure 6. Item Information Curves Plot 87

CHAPTER V: ITEM RESPONSE THEORY IN PSYCHOMETRIC RESEARCH: HOW TO ASSESS THE RELATIONSHIP BETWEEN ITEM PROPERTIES, RESPONDENTS, AND LATENT TRAITS

Figure 1. Example of an Item Characteristic Curve in IRT 97

Figure 2. Roadmap for an IRT Analysis 100

Figure 3. Rasch Model Item Characteristic Curve for Item M1 107

Figure 4. Graded Response Model Item Category Characteristic Curves for Item 6 108

Figure 5. Item Information Function for Item M1 for the Rasch Model 110

Figure 6. Test Information Function and Conditional Standard Error of Measurement Plot for the Rasch Model 112

Figure 7. Test Information Function and Conditional Standard Error of Measurement Plot for the Graded Response Model 113

CHAPTER VI: PROGRESS IS IMPOSSIBLE WITHOUT CHANGE: IMPLEMENTING AUTOMATIC ITEM GENERATION IN MEDICAL KNOWLEDGE PROGRESS TESTING

Figure 1. Three-step Process for Generating Medical MCQs Based on AIG 131

Figure 2. Map of the Difficulty Level of the Questions from the RM 136

Figure 3. Bi Values for the PT AIG-items 136

Figure 4. ICCs for the PT AIG-items 137

Figure 5. IIFs for the PT-items 138

Figure 6. TIF for the PT 139

LIST OF TABLES

CHAPTER I: UNVEILING AUTOMATIC ITEM GENERATION

Table 1. Types of rationales for an AIG-item 15

CHAPTER III: FEASIBILITY ASSURANCE: A REVIEW OF AUTOMATIC ITEM GENERATION IN MEDICAL ASSESSMENT

Table 1. Quality Assessment of Included Studies 51

Table 2. Data Synthesis 54

Table 3. AIG Validity Assessment 58

CHAPTER IV: A SUGGESTIVE APPROACH FOR ASSESSING ITEM QUALITY, USABILITY AND VALIDITY OF AUTOMATIC ITEM GENERATION

Table 1. ANOVA Results Using Time Spent Developing the Cognitive Model and Quantity of Items Generated as the Criterion 83

Table 2. Rasch Model Item Parameter Estimates and Fit Statistics for the Twenty-Seven Exam Questions 86

CHAPTER VI: PROGRESS IS IMPOSSIBLE WITHOUT CHANGE: IMPLEMENTING AUTOMATIC ITEM GENERATION IN MEDICAL KNOWLEDGE PROGRESS TESTING.

Table 1. MH Procedure for Measuring and Detecting Differential item functioning 139

Table 2. Distractor Analysis 141

Table 3. Hierarchical Linear Model Analysis 142

THESIS PLANNING

Chapter I outlines how medical education assessments are conducted and how medical test items are written. It then introduces Automatic Item Generation (AIG) and its underlying mechanisms. It describes how this procedure might overcome the underlying issues of traditional item development methods and improve medical education assessment. Rationale and major aims of this Thesis are presented in this chapter.

Chapter II approaches the role of technology in assessment, how technology impacts foundational tasks in assessment and its increasing use in medical education assessment. It then proposes AIG as a next-generation solution to the issues imposed by standard assessment practices. This chapter is part of the book "*Understanding Assessment in Medical Education through Quality Assurance*", published in 2021 (Falcão et al., 2021).

Chapter III comprises the research work on AIG and provides a narrative review of this procedure concerning its feasibility, validity and item quality in medical education assessment. This chapter is an original paper published in the *Advances in Health Sciences Education* journal in 2022 (Falcão et al., 2022).

Chapter IV displays the research work on AIG and depicts a strong theoretical approach to evaluate this procedure in terms of item quality, usability, and validity in medical education assessment. This chapter is presented as an original paper published in the *Advances in Health Sciences Education* journal in 2023 (Falcão, Pereira, Gonçalves, et al., 2023).

Chapter V builds on the previous chapter's experience to demonstrate how an Item Response Theory (IRT) analysis is conducted with datasets containing automatically generated items and manually written items. This paper considered common issues and practicalities in the type of analysis conducted, including model selection, model diagnostics, item calibration, reliability, and ability estimation. This chapter is presented as an original case published in *Sage Research Methods Cases Part 1* in 2023 (Falcão, Pereira, Pêgo, et al., 2023).

Chapter VI comprises research on how AIG-items perform in medical education formative assessment modalities, such as progress tests (PTs), compared to manually written items. In conjunction with a comprehensive psychometric approach, the AIG-items and the manually written items were subjected to qualitative and validity assessment procedures. This chapter is presented as an original paper published in the *Education and Information Technologies* journal in 2023 (Falcão, Pereira, Pêgo, et al., 2023)

Chapter VII contains a general discussion, in which the major findings of the present work are discussed considering the most recent relevant literature, as well as the important limitations found. Future avenues on how the field may evolve are also briefly discussed.

CHAPTER I

UNVEILING AUTOMATIC ITEM GENERATION

Introduction: A new educational order

Knowledge assessment is one of the most significant contributions of cognitive and behavioural sciences to our society as it provides important data regarding individuals and groups (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). In recent decades, educational testing has witnessed significant changes, particularly in how tests are administered (Gierl et al., 2021). As electronic devices with internet access are commonplace in today's modern digital era, educational tests are increasingly being conducted using computers (Lynch, 2022). This is mainly due to the increasing use of computers in everyday life and the time and resource-intensive nature of printing, scoring, and reporting paper-based exams, which makes paper-based testing no longer feasible in modern educational environments (Gierl et al., 2021). As a result, tests are now frequently administered online, with students answering to digital exams containing text, tables, diagrams, sound, and even video (Gierl & Haladyna, 2012).

The tendency toward computer-based tests (CBT) is witnessed across all academic fields, both in individual classroom settings and at state-level standardized exams (McClelland & Cuevas, 2020). As of today, numerous popular and well-known exams in the field of certification and licensing testing such as the Graduate Management Achievement Test (GMAT), the Medical Council of Canada Qualifying Exam Part I (MCCQE I), and the National Council Licensure Examination for Registered Nurses (NCLEX-RN) are already being administered on computers (Gierl & Lai, 2016a). The unexpected arrival of the Coronavirus Disease (COVID) pandemic in 2020 only strengthened this paradigm shift, as school closures forced educators to take action to adapt to the circumstances and promote education remotely (Zhao & Watterston, 2021). CBT provides major economic benefits for testing, eliminating the need for paper-based test production/distribution and classification, as well as allowing tests to be taken on demand and corrected promptly, which benefits both students and teachers (Gierl et al., 2021). Given all these advantages, it is no wonder that the transition from paper-based tests to CBT is already well underway, as we prepare to enter a new educational order (Gierl et al., 2021; Yang & Xin, 2022).

Assessment in Medical Education and the Item Development Burden

Assessment has always been a major issue in medical education. This is evidenced by the establishment of institutions (such as national boards or boards of medical examiners), the

ubiquitous presence of qualified psychometricians, and increasing concerns regarding the high stakes involved in ensuring that medical students possess the necessary competencies to advance to the next level of training (Yudkowsky et al., 2019). In the era of information technology, medical education assessment is bound to evolve rapidly, with an eye toward a more technologically driven curriculum (McLean, 2016; Remtulla, 2020). Indeed, over the last decade, medical schools have been attempting to change classical pedagogy by reducing lecture time and using technology to encourage learning, personalised instruction, and interprofessional education (Gierl & Lai, 2013; Rose, 2020). Therefore, new evaluation practices had to be implemented (Holmboe et al., 2011). The shift from paper-based tests to CBT is particularly evident in medical education assessment, with exams traditionally administered on paper now being delivered on computers with Internet connectivity, enabling educators to use complex item types that can gauge medical students' performance and a wide range of skills (Gierl & Lai, 2013).

As mentioned, the medical curriculum has undergone a significant transformation, particularly in terms of assessment and teaching methods, with the guiding principle of replacing subjectivity with objectivity (Javaeed, 2018). Exams with multiple-choice questions (MCQs) have become the most common testing method (Douthit et al., 2021). MCQs generate quantitative examinations capable of testing students' knowledge more objectively, reducing the logistical difficulties and the time required to evaluate multiple content while retaining high levels of validity (Epstein, 2007; Kurtz et al., 2020; Salam et al., 2020). However, their inclusion in CBT also created new difficulties, particularly in item development (Gierl & Haladyna, 2012). As most tests are administered regularly, items exposure becomes an issue over time. Consequently, large numbers (thousands, if not millions) of high-quality MCQs are therefore necessary to develop secure item banks and reduce exposure hazards (Gierl et al., 2021; Sinharay et al., 2003).

The conventional method of producing test items is costly and time-consuming (Gierl & Lai, 2015). For example, in a high-stakes testing program, an operational item may cost anywhere between $1500 and $2000 (1374€ to 1832€, approximately) to produce using traditional item development methods (Gierl & Haladyna, 2012). Since specialists must outline and reproduce cognitive problem-solving skills, writing one single MCQ is a challenging process in which subject matter experts (SME) must pool their knowledge and collaborate with other professionals to manage this operation (National Board of Medical Examiners, 2021). There are at least five stages involved in manual (or traditional) item writing: (i) recruiting item writers to expand the methods

already used by content specialists; (ii) training item writers to produce new items; (iii) detailed review of the items created; and (iv) empirical try-outs to calibrate item parameters (e.g., de Chiusole et al., 2018; Embretson & Kingston, 2018). This intricate combination of procedures only lengthens and increases the cost of the item development process, restricting the supply of items required for proper assessment practices (Gierl & Haladyna, 2012). Although this approach is practical for in-class assessments, it is challenging for large-scale assessments such as medical licensure exams (Gierl et al., 2012). For this reason, modern item-writing methods are required to close the gaps in conventional item-writing.

A Next-Gen Augmented Intelligence Approach: Automatic Item Generation

Augmented intelligence focuses on how computers mimic/enhance human cognitive capacities, thus improving human task performance (Gierl et al., 2021). It refers to any process/system able to blend computational power with human judgement by implementing computational analysis and data storage to solve complex and unstructured problems (Gierl et al., 2021; Shin et al., 2019). The autonomous production of test items by computer algorithms is becoming increasingly popular these days to ease the burden of item development without sacrificing item quality (Sinharay & Johnson, 2012). This usage of algorithms in assessment is novel and promises to revolutionize the educational setting (von Davier, 2018). Automatic Item Generation (AIG), the idea of using computer modules to produce testing items, emerges as one promising solution to overcome the item development burden that plagues medical assessment (Gierl et al., 2012, 2021; Gierl & Lai, 2016b, 2018). More specifically, it refers to a next-gen augmented intelligence approach where the development of test items is governed by cognitive and psychometric modelling techniques with the aid of computer technology, which scales up the item development process (Duvvuri et al., 2023; Gierl, Shin, et al., 2022).

AIG has never had a guiding principle throughout its short history. It is not yet a fully integrated science of item development (Haladyna, 2012). Nonetheless, there is evidence that the field is thriving, notably over the past ten years (Gierl et al., 2021; Gierl & Lai, 2016a; Haladyna, 2012). It is a ground-breaking topic distinct from what we typically call the science of item development, promising valuable solutions for the matter (Gierl & Haladyna, 2012; Haladyna & Gierl, 2012). Its goal is to standardize test item design to measure academic performance, which may lead to better administration, regulation, and accurate prediction of item properties (Leighton, 2012). By fusing human experience with computer algorithms, AIG pledges to rapidly generate

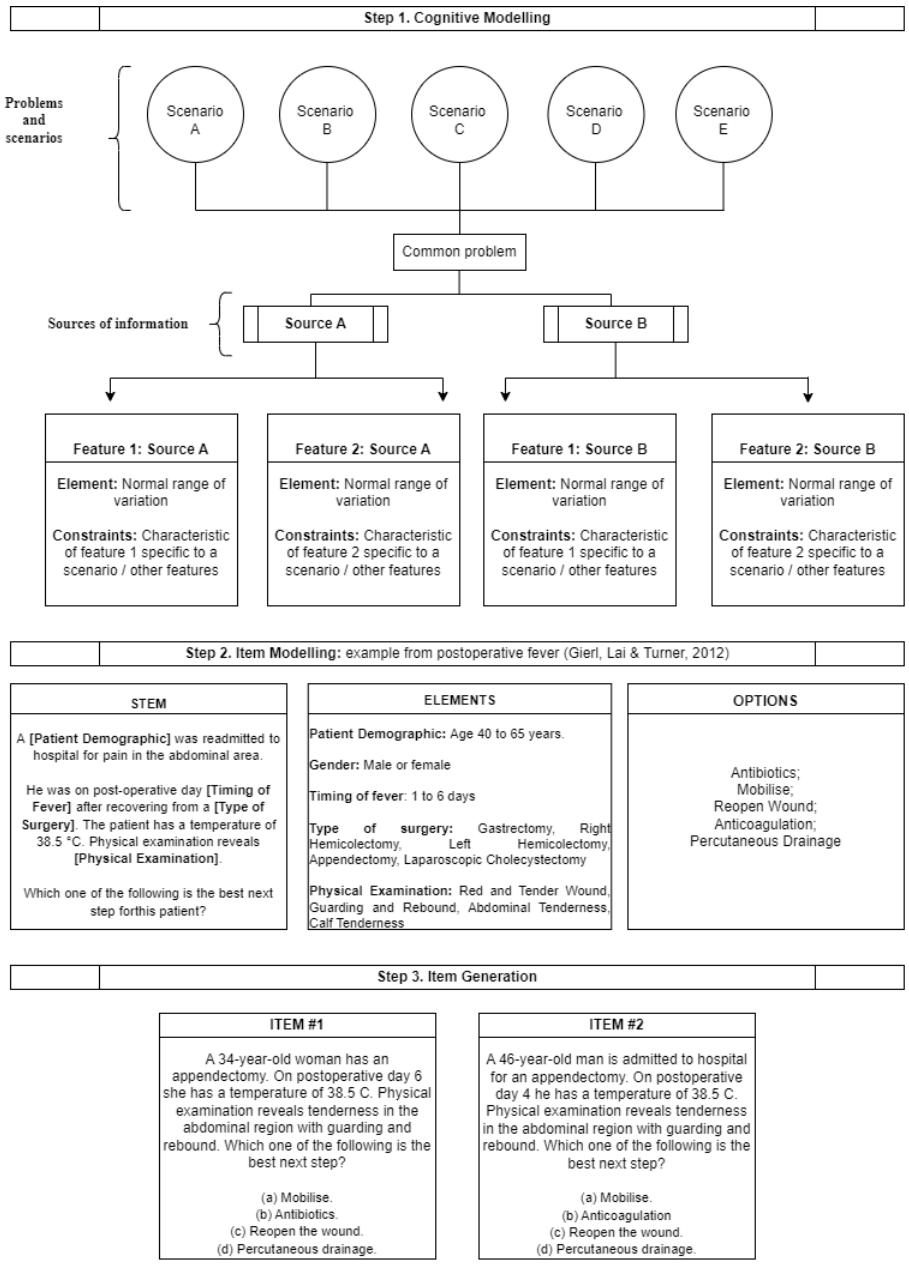
large sets of new and operational test items from just one single model (Choi, 2020; Gierl, Shin, et al., 2022). It can be differentiated from conventional item development methods in two key aspects: (i) first, in the definition of what an item is. Typical educational testing literature usually define items as "black boxes" where both inputs and outputs are presented without describing the mechanisms that transform the inputs into the outputs. Within AIG, items are defined by specifying the input as a set of parameters, constraints, and directives; (ii) second, in the workflow for writing an item. While conventional item development methods rely on a thorough process where SMEs collaborate to arrange and evaluate the contents according to their expertise, AIG blends the SMEs knowledge with computer technology to develop new functional test items at low cost and in a short period of time (Gierl et al., 2021).

The generation of medical MCQs using AIG is predicated on a scalable content development method with a theory-driven approach in which a three-stage workflow (Gierl et al., 2012; Gierl & Lai, 2016b) generates test items by considering clinical circumstances (**Cf. Figure 1**). In the first stage, SMEs design a *cognitive model* to decide on the content that will be used for item generation. The organizational structure of the cognitive model mirrors the typical approach doctors use to conduct a medical diagnosis by focusing on the knowledge and skills required to address one medical problem (Gierl, Swygert, et al., 2022; Gierl & Lai, 2016b). It is based on a coherent sequence that lists both cognitive and content-specific information, establishing relationships between content and sources of information through higher-order vs. lower-order forms of reasoning needed to solve problems (Gierl, Shin, et al., 2022; Gierl & Lai, 2016b). Once the cognitive models have been established, the stimulus features are determined according to the model (Yang et al., 2022). Therefore, it becomes possible to explain why students select incorrect options following the structural conception of the cognitive model (Gierl et al., 2021).

In the procedure's second stage, SMEs define an *item model* that specifies where the contents from the first step must go to generate new items (Gierl, Shin, et al., 2022). Item models work as a template that highlights the features of an item that must be manipulated to produce new items. It specifies the stem (the part of the item model that compiles all the information required to solve the problem), the response options available (all answers possible, including distractors and the correct option), the lead-in-question of the item and other ancillary information (e.g., graphs, tables, figures, etc.) that may compose the MCQ (Gierl, Shin, et al., 2022).

Lastly, in the third stage, computer algorithms incorporate the information from the cognitive model structure established in the first stage into the item model described in the second stage, assembling all possible value combinations, and removing any that do not comply with the restrictions previously established (Gierl, Shin, et al., 2022). Taken together, this three-step workflow serves as an item production framework that can be used to transform a model of thinking, reasoning, and problem-solving into large sets of newly generated test items (Gierl et al., 2021; Gierl, Swygert, et al., 2022).

Figure 1. Three-stage process for generating medical MCQs based on AIG.



Practical Example

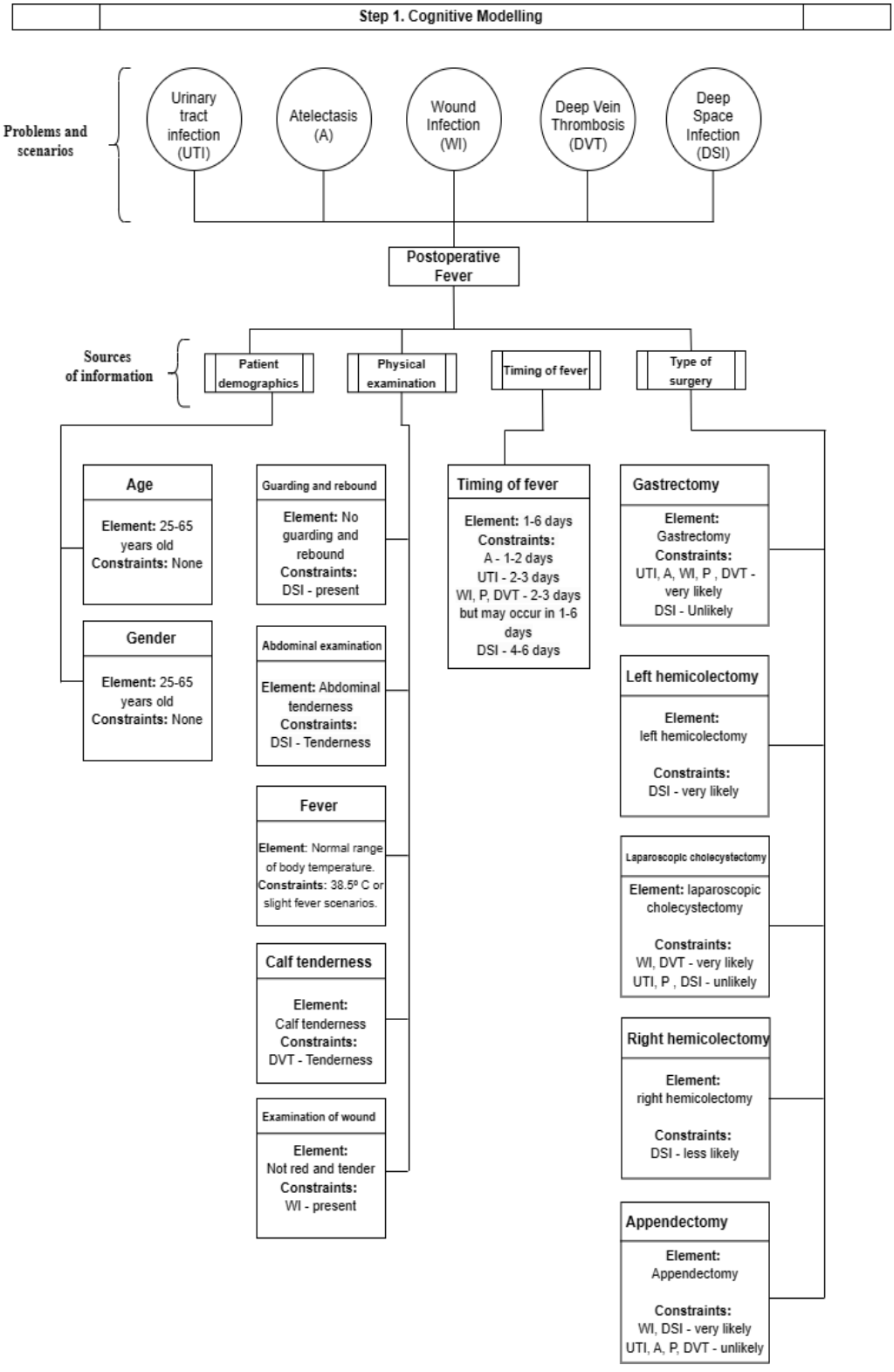
To illustrate the above stated workflow, we use a case based on the work of Gierl et al. (2012) to depict a practical example of the application of AIG to generate MCQs designed to evaluate students' understanding of complications related to postoperative fever. To solve the problems, medical students are required to understand key features for diagnosing complications related to this issue. However, it is worth mentioning that the approach described below is not limited to this topic and can be used in other medical conditions for item generation purposes (Gierl et al., 2012):

A) Cognitive Modelling

The process of building a cognitive model begins when SMEs select one parent item and consider the potential variations that can be generated from it (Gierl et al., 2021). Parent items refer to top-notch existing MCQs that generate reliable item analysis results when applied to operational tests (Gierl & Lai, 2012, 2018). They highlight the underlying structure of the cognitive model and provide a good starting point for the item generation workflow (Gierl & Lai, 2018). SMEs identify parent items at this stage by combining design principles and guidelines learned from experience, theory, and research. Cognitive theories of task performance can also be used to help select parent items and build the cognitive model. With this approach, cognitive features are identified in sufficient detail so that the item properties that predict test performance can be specified and controlled without extensive pilot testing (Drasgow, 2015; Gierl et al., 2012; Gierl & Lai, 2012).

Once the parent item has been established, SMEs identify and describe the key information required to solve the item. They do this by outlining the knowledge, skills, and content they would use to produce the right answer (**Cf. Figure 2**). In the present example, the SMEs began by describing the content domain of the item (i.e., postoperative fever) and the different scenarios modelling the problem (i.e., urinary tract infection, atelectasis, wound infection, deep vein thrombosis and deep space infection) at the top panel of the cognitive model. These issues are all considered in the model since they all gauge the problem of postoperative fever and display different features using the same symptoms (Gierl et al., 2012, 2021).

Figure 2. Cognitive model structure. Adapted from "Using automatic item generation to create multiple-choice test items" by Gierl, Lai, & Turner, 2012, *Medical education*, 46, p. 761, copyright 2012 by Blackwell Publishing. Adapted with permission from the original author.



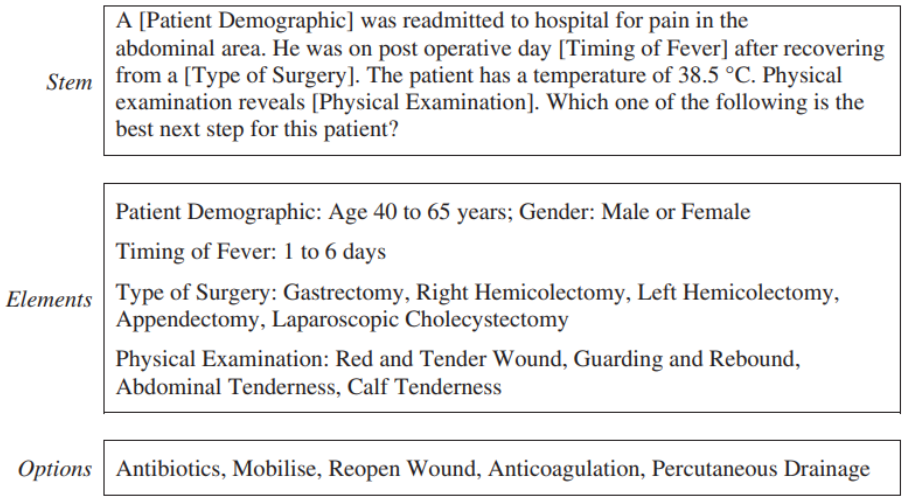
Using the set of related scenarios with the same underlying problem, SMEs then identify the sources of information required to diagnose the scenario in the middle panel of the cognitive model. Generally, these sources of data are classified as either generic (i.e., patient demographics, physical examination), or case-specific (i.e., timing of fever, type of surgery). Finally, once the features within each source of information in the cognitive model are determined (e.g., gastrectomy, left hemicolectomy, laparoscopic cholecystectomy, right hemicolectomy and appendectomy are features of the type of surgery), the possible variations (elements) can then be defined. Each feature contains manipulable elements that are unique to it and constraints that limit the elements by the scenarios relevant to the problem to ensure that only meaningful items are generated during the content assembly task in the third stage of the AIG workflow (Gierl & Lai, 2018). For instance, for the fever feature of the physical examination, the normal range of body temperature serves as the element while 38.5°C or slight fever serve as constraints (Gierl et al., 2012, 2021). Each value in the feature set should be associated with the correct option.

B) Item Modelling

Once the cognitive model has been specified, one can move forward to the second stage of the AIG workflow and format the contents defined in the cognitive model to cast new MCQs. **Figure 3** displays a sample item model where the patient demographics, the timing of fever, the type of surgery, and the physical examination represent different sources of information. These sources of data represent different variables in the item model. Here, the elements (subject to the constraints) described in the first stage are specified and manipulated to generate items with the respective options. The correct answer can be easily made for each stem when creating multiple-choice items. However, the incorrect options (i.e., the distractors) are frequently much harder to define. As AIG produces many items, there is a need for many distractors as well (Gierl & Lai, 2018). The process for creating the distractors follows a similar approach to the one used to create the correct options, with the cognitive model also including data related to the wrong options, which requires the addition of clinical features capable of distinguishing the correct from the incorrect options. This requires new restrictions that ensure that the possibilities offered in the distractors are plausible, but incorrect (Gierl & Lai, 2018; Lai et al., 2016). One method often mentioned in the literature for generating distractors is called Systematic Distractor Generation (Lai et al., 2016). Following this approach, each distractor is coded with the presentations and features shared by the generated item stem, the correct option, and other distractors. These codes are then used to

establish ranks that yield the best possible set of distractors for each item, ensuring that they are displayed whenever the correct option's features are needed, but not always in their entirety (Lai et al., 2016).

Figure 3. Item model structure. Reprinted from "Using automatic item generation to create multiple-choice test items" by Gierl, Lai, & Turner, 2012, *Medical education*, 46, p. 762, copyright 2012 by Blackwell Publishing. Reprinted with permission from the original author.



C) Item Generation

In the third stage of the workflow, dedicated computer modules assemble all possible combinations of elements to produce new test items. More specifically, during this stage, the content of the cognitive model is placed into the item model, producing hundreds/thousands of new test items (Gierl et al., 2021; Gierl & Lai, 2018). It is important to mention, however, that a single cognitive model may not be enough to produce the contents needed for an entire test as the cognitive model will exclusively produce items related to the content specified in the cognitive model's top panel. Therefore, one cognitive model would be sufficient to provide content for such a test if it aimed to measure one problem across a limited number of scenarios. Yet, in practice, tests of this nature are rarely developed. Therefore, different cognitive models may be needed to reach the test objectives (Gierl et al., 2021).

Advantages and Limitations: AIG vs Traditional Item Development Methods

AIG supports the trend towards innovation in educational testing, as it can be used to overcome the item development burden that affects medical assessment (Gierl, Shin, et al., 2022). It offers several advantages over conventional item-writing approaches, allowing the effective and affordable production of large sets of high-quality test items (Gierl, Swygert, et al., 2022). First, as it relies on the model as the analytical unit, it is much more efficient than more conventional item development methods. Classic item development approaches treat the item as the unit of analysis, thus creating each item from scratch. Conversely, AIG uses a single cognitive model to generate new items (Gierl, Shin, et al., 2022). The cognitive model can map various contexts/scenarios involving different assessment topics due to its sophisticated representation of both the content and the skills needed to solve items, which facilitates the creation of new items (Gierl et al., 2016). After that, newly functional test items are produced for different assessment topics, which makes assessment much more efficient (Gierl et al., 2021; Gierl, Swygert, et al., 2022). Second, it is more dynamic for creating new content. By using the model as the unit of analysis, medical knowledge can be easily updated to accommodate changes by modifying the content that is already present in the model. The items are then generated using the updated model version, reflecting the most current modifications to the test standards and medical knowledge, which is much more feasible than reviewing each item individually (Gierl et al., 2021). Thirdly, each model can be continually re-used, avoiding common errors in item writing that frequently occur when creating each item individually (Gierl, Swygert, et al., 2022). Fourthly, it can improve the safety of tests. Using large item banks, the item exposure rate naturally decreases. Additionally, the item assembly task in the AIG workflow also modulates the contents of the models, making it more difficult for examinees to memorize and reproduce items (Gierl et al., 2021). For all these reasons, AIG can, *de facto*, be considered a scalable item development method. As the human labor required to produce a limited number of models is reduced, item development expenses become more reasonable (Gierl et al., 2012, 2021; Gierl, Swygert, et al., 2022). Furthermore, it becomes easier to meet the needs of current medical licensure testing programs, which constantly need to update the models, specifications, and substance of the examinations (Gierl, Swygert, et al., 2022).

However, AIG also presents noteworthy issues, notably in terms of item bank management. Managing item banks containing large numbers of items is complex, which may lead to problems related to storage restrictions, search criteria, and content revision. The shift in the analytical unit from the item to the model can also bring new challenges. Medical schools may be familiar with developing and organizing items individually. Still, when using AIG, they will need to add extra steps

to their current procedures to create, manage, and maintain the models (Gierl, Swygert, et al., 2022). Additionally, one should mention that AIG is more rigorous regarding its development and revision methods. Since authors must establish values for substitutions in the item models and provide examples of stems and distractors to assure the proper generation of new items, AIG may need more time and expertise than manual item writing for initial item development. Lastly, it's also important to consider the technical environment where item generation will occur. Since AIG generates items using computer modules, this may involve more work from software engineers and subsequent quality checks than for manually written items (Kosh et al., 2019; Pugh et al., 2016).

Despite these drawbacks, AIG still offers significant advantages for medical education assessment. Modern medical licensure testing programs are continuous, extensive, and high-stakes, requiring multiple forms to ensure test security. As AIG combines SMEs skills and knowledge with computer technology rather than solely relying on human expertise, it can be applied to scale the item development process and produce the large numbers of items required to support contemporary medical education assessment programs (Gierl, Shin, et al., 2022; Gierl, Swygert, et al., 2022).

Different Assessment Systems, Same Problem?

Using test formats that enable assessments to produce both summative and formative judgments is a major development in modern educational testing (Choi & McClenen, 2020; Gierl et al., 2021). Indeed, assessment can now be leveraged to integrate classroom learning experience and offer personalized instruction (Marchisio et al., 2020). Parallel to this trend, medical education needs to respond to changes in health systems and clinical practice constantly. As a result, it requires more inventive, accurate, and adaptable methods for evaluating medical knowledge and clinical competencies (Carney et al., 2018). Traditionally, medical schools have focused on summative assessment (Knight, 2002; Terry et al., 2017). Summative examinations produce credit points which are used to calculate students' final grades (Raupach et al., 2013). These are usually administered to students in planned standardized settings, producing results that influence judgments that are frequently referred to as "high-stakes" since they predict future performance and ensure the reliability of high-stakes decisions such as progression and licensing (Gierl & Lai, 2018; Terry et al., 2017). Nowadays, we are witnessing a shift in focus in medical education assessment, with learner-centered assessments gaining interest over assessments that are

decision-maker-centered to promote better clinical reasoning, monitor students learning and develop better clinical competencies (Choi et al., 2018; Schüttpelz-Brauns et al., 2020). At the core of these learner-centered systems are contemporary computerized assessment tools that evaluate students' knowledge. Computerized Formative Testing (CFT) systems represent a trend present in modern medical education assessment. However, the advent of this trend also raised daunting challenges in item development (Gierl & Lai, 2018; Luecht, 2012; Xinxin, 2019). The possibility of using AIG to bolster the adoption of these contemporary assessment systems in medical education becomes apparent.

AIG and Formative Assessment

Formative assessment is an evaluation strategy employed by teachers and students during instruction that uses feedback to adapt teaching and improve student learning so that they can accomplish desired results more efficiently (Gierl & Lai, 2018). Generally, it can refer to any strategy that improves students' learning in response to feedback (Irons, 2007). Feedback is most valuable when it produces data that can guide educators accelerate student's acquisition of knowledge and skills (Gierl & Lai, 2018). Notwithstanding the advantages of formative assessment, it is impossible to expect all teachers to design precise tests, evaluate data quickly, and provide students with timely feedback, which advocates for new modelling support information and communication technology-based systems so that teachers can use assessment items as needed (Choi et al., 2018). CFT is an enhanced assessment system that uses computer technology to provide students feedback on their performance and help them monitor their educational progress (Gierl & Lai, 2018). They consist of on demand assessment systems that allows students to evaluate themselves by providing immediate feedback with the solutions and rationales required to solve each test item. Additionally, they give students access to many subject-specific test items designed to gauge their understanding across a variety of topics, allowing them to adjust their learning strategies (Gierl & Lai, 2018).

There are, however, two core challenges when implementing CFT systems. As mentioned before, the first lies in developing secure item banks for continuous test administration. The second lies in creating feedback within the item bank itself (Choi et al., 2018; Gierl et al., 2021; Gierl & Lai, 2018). AIG may offer a potential way to address these issues (Choi et al., 2018). It can provide immediately feedback and generate large quantities of new items while ensuring that the item bank has an adequate number of test items (Choi et al., 2018; Gierl & Lai, 2018). As feedback in CFT

systems should be focused and elaborated on the specific tasks, students should receive feedback on the contents, facts, or surface information (i.e., How well have the tasks been understood? Is the result correct or incorrect?) (Gierl et al., 2021; Wisniewski et al., 2020). Therefore, a rationale (presented here as a concise and straightforward piece of text) should be provided when giving feedback (Gierl et al., 2021). Considering this, it becomes important to set up mechanisms ready to swiftly produce rationales based on the features of the test items (Gierl & Lai, 2018).

Feedback can be created as part of the AIG workflow since the reasoning used to assemble the correct/incorrect options can also be used to generate the rationales of the problems to solve (Gierl et al., 2021). By extending the item modeling procedure in the second stage of the item development workflow and by focusing on the key features in each task, it should be possible for the SMEs to generate rationales using AIG (Gierl & Lai, 2018). Three distinct types of rationales can then be generated: (i) The first type (*Key Features*) is the simplest and is based on a list of all key features used in the stem of the item model. The second type of rationale (*Key Features Set*) is based on a set of key features in the stem of the item model. It includes elements from the key features list and at least one new set of key features, allowing for the emphasis to be placed on certain combinations of key features; Finally, the third type (*Key Features Set with Distracter*) is based on a set of key features in the stem of the item model and on a rationale for the distracters. This rationale contains elements that justify why distractors produce an incorrect answer and may contain elements from the key features list and elements from the key features set. It offers students the most thorough feedback since it includes two explanations: one for the correct options and another for the incorrect ones (Gierl et al., 2021; Gierl & Lai, 2018). In a CFT system, students usually get one type of rationale in the form of feedback for each item (Gierl & Lai, 2018). An example of each rationale type is provided in **Table 1**.

An AIG-enabled CFT system may offer an accurate and trustworthy diagnosis of medical student's strengths and limitations (Choi et al., 2018). The automated generation of rationales offers significant benefits for CFT as it can be integrated into the AIG workflow. The elements that result in the correct and incorrect options can be clearly identified and then used as feedback. It does cause some difficulties, though, as one must consider the many forms of feedback that should be delivered to students and the efforts that SMEs still must waive to devise explanations for the generated statements (Choi et al., 2018; Gierl et al., 2021). Additionally, gathering sufficient data for the correct and incorrect options (and respective coding; assemble; constraint logic

programming) can add to the time required to generate items, which may slow down the process (Gierl et al., 2021; Gierl & Lai, 2018).

Table 1. Types of rationales for an AIG-item. Reprinted from "Using automatic item generation to create solutions and rationales for computerized formative testing" by Gierl & Lai, 2012, *Applied psychological measurement*, 42(1), 42-57, copyright 2018 by SAGE. Reprinted with permission from the original author.

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? (a) Option A (b) Option B (c) Option C **(d) Option D*** *** Correct Option**
Rationale 1 (Key Features List): [Option D] is the most correct option for this patient because her type of surgery was [*Type of Surgery Feature*], her timing of fever was [*Timing of Fever feature*] and physical examination revealed [*Physical Examination feature*]. These are the key features for [Option D] as the next step.
Rationale 2 (Key Features Set): [Option D] is the most correct option for this patient as the patient was subjected to a [*Type of Surgery Feature*], with a timing of fever of [*Timing of Fever feature*] and revealed [*Physical Examination feature*]. The key features in this diagnosis include [*Set of key features*].
Rationale 3 (Key Features Set w/ Distracter Rationale): [Option D] is the most correct option for this patient as the patient was subjected to a [*Type of Surgery Feature*], with a timing of fever of [*Timing of Fever feature*] and revealed [*Physical Examination feature*]. The key features in this diagnosis include [*Set of key features*]. The incorrect options are not plausible because [*justification for why the distracters produce an incorrect answer*]

Statement of the Problem and Thesis Aims

Despite the promising scenario that AIG presents, certain "grey areas" regarding this assessment procedure lack depth. As a result, academics should intensify their efforts in these areas to dispel any lingering questions regarding the AIG procedure. More specifically, there should be a focus on concerns related to the assumption that AIG-items may not be as high-quality as those produced using more traditional procedures (Pugh et al., 2020). This may be due to the fact that the psychometric properties of AIG-items (e.g., item difficulty, reliability, information, distractor performance, differential item functioning, etc.) are mostly undocumented in the available literature (Shappell et al., 2021). Similarly, the relationship between users and the tasks in an AIG-based item development context lacks further empirical evidence, particularly in terms of ease of learning. Studies on this subject are scarce (or even inexistent) to this point and do not apply to a wider context. This lack of knowledge may be preventing educators who lack knowledge/specialized training in AIG from using a modern and efficient procedure that may be accessible to everyone in terms of learning. In addition, it is critical to ensure that AIG is a valid procedure and will not jeopardize testing security. Since item development records the procedures and outcomes necessary to produce high-quality test content, it is necessary to assemble major evidence and provide a scientific basis for validating procedures such as the one used by AIG (Gierl, Swygert, et al., 2022). Lastly but not least, considering the developments unfolding throughout the world in this new age of testing and the recent paradigm shifts, it is relevant to understand whether AIG can be implemented in contemporary assessment processes that support both summative and formative inferences.

In this Thesis, our main aim was to provide new insights into these grey areas of the literature surrounding AIG. To this end, data from various research focusing on exams used in the context of medical education assessment were subjected to empirical, psychometric, and statistical processes. These tests were composed of both AIG and manually written items, which allowed the two item development methods to be compared with respect to several features. Apart from this objective, this Thesis proposes to disseminate knowledge regarding the AIG procedure, the psychometric properties of AIG-items, the necessary practices to employ such analysis and the assessment systems where AIG can be implemented to deepen future implementation processes. Therefore, considering the above-mentioned objectives, the detailed tasks performed in this Thesis are presented below:

1. Discuss the evolution of assessment in medical education and how the increasing need for high-stakes assessment opportunities has empowered computer-based assessment over traditional assessment formats; and how a next-generation item development method such as AIG could overcome the challenges imposed by conventional item writing methods (**Chapter II**).
2. Portray the state of the art regarding AIG through a narrative review focused on the feasibility, validity, and item quality of this procedure in medical education assessment (**Chapter III**).
3. Compare the quality of AIG-items and manually written items; measure the efficiency of the AIG procedure; explore its learnability properties; and search for sources of evidence that support AIG as a valid item development method in medical education assessment (**Chapter IV**).
4. Discuss the process of applying an Item Response Theory (IRT) analysis to evaluate the psychometric properties of AIG-items (**Chapter V**).
5. Compare the psychometric properties and validity of AIG-items versus manually written items in formative assessment modalities through a mixed-methods analysis (**Chapter VI**).

References

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
2. Carney, P., Mejicano, G., Bumsted, T., & Quirk, M. (2018). Assessing learning in the adaptive curriculum. *Medical Teacher, 40*(8), 813–819. https://doi.org/10.1080/0142159X.2018.1484083
3. Choi, J. (2020). Automatic item generation with machine learning techniques. In H. Jiao & R. Lissitz (Eds.), *Application of Artificial Intelligence to Assessment* (Vol. 1, pp. 189–210). Information Age Publishing. https://doi.org/10.3389/feduc.2023.858273
4. 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

5. Choi, Y., & McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. *Applied Sciences*, *10*(22), 8196. https://doi.org/10.3390/app10228196
6. de Chiusole, D., Stefanutti, L., Anselmi, P., & Robusto, E. (2018). Testing the actual equivalence of automatically generated items. *Behavior Research Methods, 50* (1), 39–56. https://doi.org/10.3758/s13428-017-1004-5
7. Douthit, N., Norcini, J., Mazuz, K., Alkan, M., Feuerstein, M., Clarfield, A., Dwolatzky, T., Solomonov, E., Waksman, I., & Biswas, S. (2021). Assessment of Global Health Education: The Role of Multiple-Choice Questions. *Frontiers in Public Health, 9*. https://doi.org/10.3389/fpubh.2021.640204
8. Drasgow, F. (Ed.). (2015). *Technology and Testing: Improving Educational and Psychological Measurement* (1st ed.). New York, NY: Routledge. https://doi.org/10.4324/9781315871493
9. Duvvuri, V., Lee, G., Hsu, Y., Makwana, A., & Morgan, C. (2023). Post Processing Selection of Automatic Item Generation in Testing to Ensure Human-Like Quality with Machine Learning. *Proceedings of the 2023 7th International Conference on Machine Learning and Soft Computing* (pp. 82–88). Association for Computing Machinery. https://doi.org/10.1145/3583788.3583800
10. Embretson, S., & Kingston, N. (2018). Automatic Item Generation: A More Efficient Process for Developing Mathematics Achievement Items? *Journal of Educational Measurement, 55*(1), 112–131. https://doi.org/10.1111/jedm.12166
11. Epstein, R. (2007). Assessment in medical education. *New England Journal of Medicine, 356*(4), 387–396. https://doi.org/10.1056/NEJMra054784
12. Gierl, M., & Haladyna, T. (2012). Automatic Item Generation: An Introduction. In M. Gierl & T. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 3–12). Routledge. https://doi.org/10.4324/9780203803912
13. 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
14. Gierl, M., & 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

15. Gier, 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*, 9–20. https://doi.org/10.55612/s-5002-025-001.
16. Gierl, M., & Lai, H. (2016a). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of Test Development* (2nd ed., pp. 410–429). Routledge.
17. Gierl, M., & Lai, H. (2016b). The role of cognitive models in automatic item generation. In Rupp, A., Leighton, J. (Eds.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications*, (pp. 124–145). John Wiley & Sons. https://doi.org/10.1002/9781118956588.ch6
18. 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
19. Gierl, M., 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
20. Gierl, M., Lai, H., & Tanygin, V. (2021). *Advanced Methods in Automatic Item Generation*. Routledge. https://doi.org/10.4324/9781003025634
21. 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
22. Gierl, M., Shin, J., Firoozi, T., & Lai, H. (2022). Using Content Coding and Automatic Item Generation to Improve Test Security. *Frontiers in Education*, 1–14. https://doi.org/10.3389/feduc.2022.853578/full
23. Gierl, M., Swygert, K., Matovinovic, D., Kulesher, A., & Lai, H. (2022). Three Sources of Validation Evidence Needed to Evaluate the Quality of Generated Test Items for Medical Licensure. *Teaching and Learning in Medicine*, 1–11. https://doi.org/10.1080/10401334.2022.2119569
24. Haladyna, T. (2012). Automatic item generation: A historical perspective. In M. Gierl & T. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 23–35). Routledge.

25. Haladyna, T., & Gierl, M. (2012). Obstacles for Automatic Item Generation. In M. Gierl & T. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 241–249). Routledge.
26. Holmboe, E., Ward, D., Reznick, R., Katsufrakis, P., Leslie, K., Patel, V., Ray, D., & Nelson, E. (2011). Faculty development in assessment: The missing link in competency-based medical education. *Academic Medicine, 86*(4), 460–467. https://doi.org/10.1097/ACM.0b013e31820cb2a7
27. Irons, A. (2007). *Enhancing Learning through Formative Assessment and Feedback* (2nd ed.). London: Routledge. https://doi.org/10.4324/9780203934333
28. Javaeed, A. (2018). Assessment of higher ordered thinking in medical education: multiple choice questions and modified essay questions. *MedEdPublish*, *7*(128), 128. https://doi.org/10.15694/mep.2018.0000128.1
29. Knight, P. (2002). Summative assessment in higher education: Practices in disarray. *Studies in Higher Education, 27*(3), 275–286. https://doi.org/10.1080/03075070220000662
30. Kosh, A., Simpson, M., 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
31. Kurtz, J., Holman, B., & Monrad, S. (2020). Training medical students to create and collaboratively review multiple-choice questions: a comprehensive workshop. *MedEdPORTAL, 16*, 10986. https://doi.org/10.15766/mep_2374-8265.10986
32. Lai, H., Gierl, M., Touchie, C., Pugh, D., Boulais, A-P., & De Champlain, A. (2016). Using Automatic Item Generation to Improve the Quality of MCQ Distractors. *Teaching and Learning in Medicine, 28*(2), 166–173. https://doi.org/10.1080/10401334.2016.1146608
33. Leighton, J. (2012). Learning sciences, cognitive models, and automatic item generation. In M. Gierl & T. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 131–145). Routledge.
34. Luecht, R. (2012). Automatic Item Generation for Computerized Adaptive Testing. In M. Gierl & T. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 196–216). Routledge.

35. Lynch, S. (2022). Adapting Paper-Based Tests for Computer Administration: Lessons Learned from 30 Years of Mode Effects Studies in Education. *Practical Assessment, Research, and Evaluation, 27* (22). https://scholarworks.umass.edu/pare/vol27/iss1/22
36. Marchisio, M., Fissore, C., & Barana, A. (2020). From standardized assessment to automatic formative assessment for adaptive teaching. *Proceedings of the 12th International Conference on Computer Supported Education, 1*, 285–296. https://doi.org/10.5220/0009577302850296
37. McClelland, T., & Cuevas, J. (2020). A comparison of computer based testing and paper and pencil testing in mathematics assessment. *The Online Journal of New Horizons in Education*, 10(2), 78–89. https://www.tojned.net/journals/tojned/articles/v10i02/v10i02-01
38. McLean, S. (2016). Case-based learning and its application in medical and health-care fields: a review of worldwide literature. *Journal of Medical Education and Curricular Development*, *3*, JMECD-S20377. https://doi.org/10.4137/JMECD.S20377
39. National Board of Medical Examiners. (2021). *NBME Item Writing Guide*. NBME. https://info.nbme.org/item-writing-guide.html
40. 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
41. 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
42. Raupach, T., Brown, J., Anders, S., Hasenfuss, G., & Harendza, S. (2013). Summative assessments are more powerful drivers of student learning than resource intensive teaching formats. *BMC Medicine*, *11*, 1–10. https://doi.org/10.1186/1741-7015-11-61
43. Remtulla, R. (2020). The present and future applications of technology in adapting medical education amidst the COVID-19 pandemic. *JMIR Medical Education, 6*(2), e20190. https://doi.org/10.2196/20190
44. Rose, S. (2020). Medical student education in the time of COVID-19. *JAMA*, *323*(21), 2131–2132. https://doi.org/10.1001/jama.2020.5227

45. Salam, A., Yousuf, R., & Bakar, S. (2020). Multiple choice questions in medical education: How to construct high quality questions. *International Journal of Human and Health Sciences, 4(2)*, 79. https://doi.org/10.31344/ijhhs.v4i2.180
46. 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
47. Shappell, E., Podolej, G., Ahn, J., Tekian, A., & Park, Y. (2021). Notes From the Field: Automatic Item Generation, Standard Setting, and Learner Performance in Mastery Multiple-Choice Tests. *Evaluation & the Health Professions, 44*(3), 315–318. https://doi.org/10.1177/0163278720908914
48. Shin, J., Guo, Q., & Gierl, M. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology, 10*, 825. https://doi.org/10.3389/fpsyg.2019.00825
49. Sinharay, S., & Johnson, M. (2012). Statistical modeling of automatically generated items. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and Practice* (pp. 193–205). Routledge.
50. Sinharay, S., Johnson, M., & Williamson, D. (2003). Calibrating Item Families and Summarizing the Results Using Family Expected Response Functions. *Journal of Educational and Behavioral Statistics, 28*(4), 295–313. https://doi.org/10.3102/10769986028004295
51. Terry, R., Hing, W., Orr, R., & Milne, N. (2017). Do coursework summative assessments predict clinical performance? A systematic review. *BMC Medical Education, 17*, 1–15. https://doi.org/10.1186/s12909-017-0878-3
52. 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
53. Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087
54. Xinxin, Z. (2019). *Using automatic item generation to create content for computerized formative assessment* [Doctoral Dissertation, University of Alberta]. University of Alberta. https://doi.org/10.7939/r3-vbt3-6598

55. Yang, L., & Xin, T. (2022). Changing Educational Assessments in the Post-COVID-19 Era: From Assessment of Learning (AoL) to Assessment as Learning (AaL). *Educational Measurement: Issues and Practice, 41*(1), 54–60. https://doi.org/10.1111/emip.12492
56. Yang, Y., Sanyal, D., Michelson, J., Ainooson, J., & Kunda, M. (2022). Automatic item generation of figural analogy problems: A review and outlook. *The Ninth Advances in Cognitive Systems Conference,* ArXiv:2201.08450. https://doi.org/10.48550/arXiv.2201.08450
57. Yudkowsky, R., Park, Y. S., & Downing, S. M. (2019). *Assessment in health professions education* (2nd ed.). New York: Routledge. https://doi.org/10.4324/9781138054394
58. Zhao, Y., & Watterston, J. (2021). The changes we need: Education post COVID-19. *Journal of Educational Change, 22(*1), 3–12. https://doi.org/10.1007/s10833-021-09417-3

CHAPTER II

THE ROLE OF TECHNOLOGY IN ASSESSMENT QUALITY ASSURANCE PROCESSES

This chapter is based on the following publication: **Falcão, F**., Costa, P. & Pêgo, J.M. (2021). The Role of Technology in Assessment Quality Assurance Processes. In Malau-Aduli, B. S., Hays, R., & Van Der Vleuten, C. P (Eds.). *Understanding Assessment in Medical Education through Quality Assurance* (pp. 171-189). ISBN: 9781260469653

The Role of Technology in Assessment Quality Assurance Processes

7

Filipe Falcão, Patrício Costa and José Miguel Pêgo

CHAPTER HIGHLIGHTS

- Technology is changing the landscape of assessment in medical education.
- Computer-based testing increases efficiency and flexibility of assessment.
- Machine learning adds realism and reliability to assessment.

ORIENTATION TO THE CHAPTER

In this chapter, we will address the impact of the latest technological developments in quality processes regarding assessment in medical education.

In the first part, we discuss the evolution of assessment in medical education and how the growing need for assessment opportunities has nudged traditional formats toward computer-based assessment. In the second part, we look at how technology is impacting foundational tasks in assessment in particular, test design, test assembly, and reporting. In the third part, we will analyze how technology is impacting operationally and improving efficiency and accuracy of assessments. Finally, we will elaborate on what we envision the future holds for assessment in medical education.

INTRODUCTION—THE INCREASING ROLE OF TECHNOLOGY IN MEDICINE AND MEDICAL EDUCATION

Profound global and economic changes are changing educational measurement [1]. Thanks to the unified effort of cognitive sciences, statistical theories of test scores, education psychology, educational technology,

Ch07.indd 171 29-05-2021 16:23:16

and computer science, educational measurement is evolving at a fast pace. This exponential evolution translates into clear changes in the assessment systems, the most notable of which is the transition from assessment through paper tests to computer tests [2]. This change reveals a shift in the paradigm of educational assessment, with paper-based tests being no longer a feasible method because of its resource-intensive process, which urged for solutions like computer-based testing (CBT) systems [3,4,5]. In fact, over the last few years, CBT has become a widely used form of evaluation and is now considered one of the most effective assessment methodologies in education [6].

Computer-Based Testing

CBT is changing educational measurement due to the merging of test-administration procedures with digital media and the Internet, which makes it possible to create new types of tests and test-ing resources. These tests are paperless and administered with a computer that uses a test model to implement a process of selecting and adminis-tering items, allowing the students' performances to be scored and cat-egorized by their ability in each item [2]. According to the authors, this approach has potential benefits such as shortened tests without loss of measurement precision; enhanced score reliability, particularly for low- and high-ability examinees; improved test security; testing on demand; and immediatest scoring and reporting.

The introduction of the Internet and the proliferation of Internet-based computerized assessment has further improved this scoring system, making it especially advantageous for both educators and stu-dents when compared with traditional testing methods [2,7]. Through the administration of tests over the network, educators not only are free of time-consuming processes usually associated with paper-based tests but also are able to use more diverse item formats that can measure a broader range of skills. Students also profit from this approach, sometimes being able to take exams when and where they want, as well as receive imme-diate feedback on their performance thanks to computerized algorithms that allow the immediate quotation of their tests [5,8].

Medical education stands out as one of the areas where this change in assessment methodologies has become evident, introducing new teaching and learning strategies that allow medical educators to measure complex performances and competencies [9,10]. Consequently, medical examinations once administered in paper format are now delivered through computers connected to the Internet. Computers allow the use of item types that help medical educators measure performances and a broad set

Ch07.indd 172 29-05-2021 16:23:16

of competencies, as well as provide continuous on-demand testing and instantaneous feedback [10].

Measurement in medical education is an extremely complex task because it involves a variety of competences and settings [11]. Medical educators are constantly challenged with teaching demands that reflect the continuous growth in scientific information and society's expectations of the universe of knowledge that medical doctors must master to be fit for the profession [11]. In consequence, there is a constant conflict between education and research related to funding, time, and promotional aspects [11]. Because of this expectation and the inadequacies of training models, medical education experienced great pressure to transform, which took medical educators to seek better evaluation tools like CBT [12]. This search for assessment tools that support learning and measure performance is also a reaction against old methods of assessment that had adverse effect on both learners and teachers, which culminated in new assessment strategies to meet these needs of innovation and quality improvement [9,11].

Assessment in medical education has three main goals: (1) optimize the capabilities of learners, directing them to future learning; (2) protect the public through the identification of doctors not fit for the job; and (3) provide a platform for choosing applicants for complex training [11]. According to the same authors, assessment in this area can be classified into one of two types: (1) *formative*, which provides guidance for future learn-ing, promotes reflections and reassurance, and shapes values; and (2) *summative*, which consists of an overall judgement regarding the competence or qualification of the learner.

According to Batalden et al. [11], one of the main evaluation methodologies in this area is the written exam with open-ended or multiple-choice questions (MCQs). The latter are the foundation for assessments used in medical education because of their capacity to provide broad sampling of knowledge in many areas and domains of specialization using larger sets of items, which makes other methodologies impractical [13,14]. In addition, they can be scored by computers, which can help administer them in an even shorter period [11]. They are also particularly advantageous for evaluating medical students due to their ability of testing a large number of candidates with reduced human intervention, being efficient and economic [13,15]. In the medical context, MCQs measure the breadth and ability of graduate and postgraduate students within a particular specialty, standing out for their objectivity, efficiency, increased quantity, and item statistics that help discern their psychometric qualities [14].

However, CBT brought new challenges, particularly in item development [2,16]. In fact, the growth of Internet-based

computerized testing warned of the needs for new testing methods and practices [7]. Since these tests are continually applied and consequently exposed, thousands of items are necessary to develop item banks—pools of test items containing information regarding their content and psychometric details such as difficulty—that need to be constantly replenished to minimize these risks [1,10]. In medical education, in addition to the demands of larger item banks, there are problems related to the implementation of MCQs such as the time and money spent on their construction [10,14]. Although the creation of MCQ follows guidelines that allow the development of items across different specializations, this process is complex because cognitive problem-solving skills and content knowledge should be expressed by specialists and replicated countless times [13]. These items are also not free of validity problems, which calls for more alternative methods of test development [10,15].

Because of this transition to computerized testing, the demand for new items is now greater than the supply in educational assessment [3, 17]. Educational items as they are currently produced are expensive and time-consuming [1]. In the traditional method, developers are hired to extend the approach of content experts, which consists in creating one item at a time [16]. This approach to the development of content items requires "experts to use test specifications, and item-writing guides to author each item" ([2], p. 5), which is an expensive process because of the need to hire experts.

Altough feasible for classroom assessments, this method is difficult to apply in large-scale examinations [13]. Extensive MCQ development is necessary to create items needed for medical licensure examinations [13]. In addition to the costs of hiring specialists, there are costs related to item quality control, which implies that all items must be field-tested before their use to measure their psychometric characteristics (sometimes resulting in the elimination of a large number of items). Because educators now need to create a large number of new items for computerized testing, this process of writing, revision, and editing shows us that this traditional approach does not seem to be the most promising way to deal with this problem [1].

Automated Item Generation

Automated item generation (AIG) is a next-generation assessment theory that promises to overcome the challenges imposed by traditional item development and to smooth the item development burden for medical examinations [8,13,18]. AIG is an item construction technology in applied psychometrics that has attracted interest because of its potential capacity to produce high-quality and content-specific test items rapidly

Ch07.indd 174 29-05-2021 16:23:17

and efficiently [3,19]. It regards an area where computer technology, cognitive theories, and psychometric practices gather to conduct a process that can produce assessment tasks [1]. More specifically, it consists in the "process of using cognitive models to generate items using computer technology," thus providing a way to systematically produce a large number of items focused on a specific content both quickly and efficiently ([5], p. 6). Guided by psychometric theory of test items and aided by computer algorithms, it requires human judgment and expertise to define the elements in the item model and computers to combine these elements [13,14,20]. This requires content specialists to design and develop meaningful item models. Next, computer technology systematically combines the elements of these item models to produce large numbers of test items in a short period of time, allowing the measurement of medical student's knowledge and problem-solving skills in different areas of medical education [2,13,20].

HOW TECHNOLOGY IS IMPACTING FOUNDATIONAL TASKS IN ASSESSMENT

Test Design

Test design includes all steps from the development of a blueprint, item development, test assembly, and test administration. In addition, postadministration tasks such as scoring and feedback provision depend on the purpose of the assessment but need to be thought about beforehand to understand its impact on the design of the test. All of these steps can be technologically enhanced, improving efficiency and the quality of the final design in some cases.

Item Development

As indicated in Chapter 3, item development is traditionally accomplished by an individual or group of content experts. College-sized institutions typically develop their assessments in-house, resorting to individual faculty members and groups in some cases. Less frequently are people specially dedicated to item development. On the other hand, licensing institutions such as certification boards and standard test developers frequently group content experts to develop a small number of high-quality items. Although in the first case items usually tend to be of low cost and of poor quality [21], certification-level items that are used in high-stakes examina-tions are expensive due to the extensive quality-control process [22].

The widespread use of computer-based assessment has helped ease this process by enforcing item templates and guiding authors toward the

best item shape and format. In addition, CBT provides item formats that are difficult or impossible to reproduce on paper such as multistep MCQs, hotspot, long-list items requiring video or audio recordings, and so on.

Recent developments have brought new avenues to item development and the final quality of the assessment that depends on computer technology such as automation of item development and the use of natural language processing and machine learning.

Automated Item Generation

To clearly understand the process of creating items through AIG, it is necessary to define some concepts—namely, item model, element, stem, option, and auxiliary information. According to Gierl and Lai [1], an item model is like a template that indicates the variables (elements) in the task that can be manipulated to create new items. More specifically, it is a general representation of the item to be generated, incorporating necessary properties of the items like the stem, options, and auxiliary information [23]. Gierl and Lai [1] claim that the stem is the part of the item model that contains the "context, content, [or] the question the student is required to answer" (p. 11)—that is, formulating the problem that the item addresses [23]. The options correspond to the possible answers to the item, where one is correct and one or more are incorrect. Finally, auxiliary information consists in item-related data that provide information regarding the item (e.g., image or tables) [1]. Further, the stem and options are divided into elements, which can be non-numeric or numeric (integers). Regarding MCQ item models, the stem and options are always required, whereas only the stem is necessary in open-ended response item models [7].

AIG usually requires two major steps. In the first, an item model is created derived from good-quality test items [23]. More specifically, a content specialist creates item models indicating which elements in the task should be manipulated [1]. In the second step, these variables in the item model are manipulated through computer algorithms to generate new items. More specifically, once the item model is created by content specialists, computer technology combines the elements specified by them in each model, producing a series of new items called *isomorphs*—that is, clone items [1,23]. With this two-step process, AIG creates a large quantity of new items from one item model, ensuring that there are enough items for the item banks, which reflects the full potential of this technique in the future of assessment [1,18].

Regarding the generation of medical multiple choice items using AIG, Gierl et al. [13] list three steps. Unlike the process used in most item- generation techniques, the first step in medical evaluation is to create a cognitive model structure. Here the content

Ch07.indd 176 29-05-2021 16:23:17

included in medical examinations is identified by specialists, who outline a framework that structures the knowledge, skills, and content necessary to create a medical diagnosis, forming a cognitive model structure for AIG—that is, the experts summarize how they usually approach a problem [22]. This model organizes cognitive and content-specific information, presenting the content relationships and sources of information used to formulate medical diagnoses ([13], p.47). The second step consists in the development of the item models. More specifically, in this step the content of the cognitive model structure is cast to an MCQ, including the necessary information to generate a group of items. Variables to be altered are specified, vignettes including the information needed to answer a question are written, questions and correct options are defined, and distractors are generated [22,24]. Finally, the third step corresponds to the generation of items using item models through computer algorithms. Taking this process into account, we can consider AIG as a process by which the representation of students' thinking processes (cognitive models) are used to generate items thanks to computer algorithms [25].

Through the manipulation of the elements, specialists can generate a high number of items from just one item model, which can consequently modify the difficulty of an item. If the generated items were created with the objective of measuring contents at the same levels of difficulty, we consider the items to be isomorphic. In this case, only the incidental elements (or surface features of an item) are manipulated. On the other hand, when the goal is to create items with different difficulties, the incidental elements of the item model can be manipulated together with one or more radical elements (features that change item difficulty and test characteristics) [7].

However, AIG is not a rule-free process. It must follow standards of quality and guidelines for item-development practices that are initially used to identify a parent item model [13]. A *parent item model* refers to a high-quality item that "produce[s] strong item analysis results drawn from existing operation tests" (Gierl & Lai, 2018, p. 46). A possible approach to identify a parent item model is through a cognitive theory of task performance. This approach is called *strong theory*, in which cognitive features are identified to control and predict test performance, which can help predict the difficulty of the generated items and, consequently, calibrate them without extensive pilot testing. However, these theories are rare. When a theory is not available, parent item models are identified through *weak theory*. In this approach, older items from other exams are reviewed "drawing on an inventory of existing test items" to identify an underlying structure that provides a reference to create alternative item models with features that can be changed to generate new items ([7], p. 278).

Ch07.indd 177 29-05-2021 16:23:17

Another important point to consider regards the taxonomy of the created item models. Item models can be (1) independent, (2) dependent, (3) mixed, or (4) fixed. This classification depends on the elements of the stem and the options of the model. In independent item models, the elements of the stem are unrelated, and a change in one element will not affect other elements. On the other hand, in dependent item models, those elements are causally related, whereas in mixed independent or dependent models at least one pair of the elements of the stem are related. Finally, fixed-item models are models where the stem elements have no variation. Regarding the options, they may be fixed (when distracters—plausible but incorrect alternatives [25]—and the right option are invariant in the item model), presented randomly, or in a constrained way (distracters are generated according to constraints). After the process of item model development and taxonomy is defined properly, AIG is ready to start. Through the aid of computer algorithms, item models are used to generate items and item banks [7].

The use of algorithms to generate items in education has the appeal that it may be able to produce unlimited resources for assessment develop-ment (von Davier, 2018). In addition to being a scalable process because it creates a large volume of items [3], AIG promises also to increase exam security through large sets of items with equal levels of difficulty (Cole et al., 2020), helping test developers identify tests that contains too many easy or difficult items. AIG can ease the development burden from the reuse of the item models instead of writing new items from scratch, avoiding errors that occur frequently during this phase (e.g., grammatical, punctuation, etc.) [3]. Because of these properties, we believe AIG can be used in medical schools, especially because their own item banks often have reduced numbers of items in certain areas of study [14].

Very Short Answer Questions

A very short answer question (VSAQ) is a novel type of item that can only be administered in CBT. It consists of a clinical vignette followed by a question (usually about diagnosis or management) that requires a candidate to generate a short response [26]. The computer records these free-text answers (usually between one and four words) in a list of unique answers that can then be processed in a semiautomated way. Item developers can prepopulate the scoring system with a list of possible correct choices, and the assessment software can select correct answers by natural language processing. The advantage over traditional MCQs is that the cueing effect is absent, which has been shown to be an excep-tionally reliable and discriminatory assessment method both in formative and summative examinations [26,27]. By eliminating cueing, it recreates

Ch07.indd 178 29-05-2021 16:23:17

real-life conditions of knowledge application and candidates are better challenged. On the other hand, the list of answers that VSAQs collect can be used as distractors to develop classical MCQs. The advantage of this approach is that these are more likely to be better distractors and increase the discrimination power of classical MCQs because they represent the most common mistakes and misconceptions of test takers in a particular subject.

Test Form Assembly

Test form assembly is the process by which items in a particular assessment are arranged and presented to the test taker. This step is based on the blueprint of the test and can be extraordinarily complex, depending on the number of dimensions that have been defined for the blueprint. These dimensions are usually conditioned by the content (e.g., geometry, anatomy) being assessed by the item but may involve additional characteristics outside the content being sampled such as cognitive tasks being elicited (e.g., Bloom's taxonomy) and many more characteristics of the item that collectively constitute metadata associated with the item. Test assembly should be careful to ensure that the content is being comprehensively sampled without much overlap or underrepresentation of both content and metadata because these are important sources of information and feedback for both assessors and test takers. Computer-based assessment has changed how test form assembly is done, making it easier to classify items and manage item banks. Metadata can be easily associated with the item content and can be extraordinarily detailed. With specialized assessment software, it is possible to identify items associated to a particular characteristic and select them for the final test form. What was once a laborious process of manually picking items stored in text files and spreadsheets can be done with relative ease and automated to the point of fully autonomous assembly. Software can decide the specific set of items being included based on a blueprint of any dimensions but also taking into consideration the psychometric proprieties of items or the number of previous times used. This automation provides a randomly chosen sample of items that maximizes representation and avoids overlaps.

Another advantage of automated test assembly is the development of unique test forms for each test taker. Randomization of item order presentation is a method commonly used to avoid fraud and cheating. Paper-based test may include a few versions (usually three to five) of the same test form. Although it adds security, it also increases the workload needed to develop the test. Most software solutions, however, allow for the deployment of as many test forms as test takers, which reduces the

Ch07.indd 179 29-05-2021 16:23:17

risk of collusion and fraud. The level of randomization that computers can generate reduces the risk of assembling a test form that congregates too many difficult or easy items at the beginning or any other section of the form that has an emotional impact on test takers.

Scoring of Traditional and Novel Items

Test items traditionally fall into two categories: choice items and open answer (input items). Although open-answer items are highly flexible and allow for the evaluation of complex cognitive tasks such as synthetic ability, reasoning, and critical thinking, calculus, or design, they are more prone to bias in the scoring process. In addition, answering open-answer items may be labor and time intensive, which limits the number of items that can be included in one test form. The gain in freedom and realism provided by this category of item is counterbalanced by the limitation in sampling of contents and may introduce a sense of unfairness in the test taker. Choice items have gained popularity over the years in higher education. They can be developed relatively quickly and allow for the testing of a broad range of subjects in a short period of time. On the other hand, choice items have been criticized for being less flexible, mostly testing knowledge (memory) or knowledge application, which are the lowest levels of cognitive functions. Most important, scoring choice items is straightforward, unbiased, and rapid even when done manually. This provides a rich source of reliable information. The psychometric proprieties of choice items are important indicators of the reliability of the measurement and the discriminating power of the tool. With the advancement of computer technology, this statistical analysis has evolved from simple arithmetic as found in classical test theory (CTT) to sophisticated mathematics as found in item response theory (IRT) [28]. Although CTT is a deterministic model that provides measurements of the test performance as a whole and of the difficulty and discrimination power of individual items, it depends on the cohort of test takers and is based on the premise that items' characteristics are similar across the spectrum of performance, which is not true. IRT is a probability model that was developed to correct this by developing metrics that are independent of the population taking the test. It provides information that is intrinsic to the item ability as well as providing information about sources of error outside the domain of knowledge such as guessing. IRT allows for the development of adaptative test modalities such as computerized adaptative testing, which uses population independent statistics to provide a test form that is tailored to the test taker to maximize measure-ment precision [28].

Ch07.indd 180 29-05-2021 16:23:17

Scoring test items is a critical step in the assessment process and is potentially a source of error. Traditional items have a relatively straightforward process that can be fully automated as in choice items or that need the intervention of a judge to determine the extent of correctness in choice items. The emergence of machine learning, however, has provided tools to bridge best-of-choice and open-answer items. VSAQs are a simplistic form of machine learning with a scoring system based on a list of potential correct answers that can be matched to the answer given by the test taker. Its main limitations are the number of words that can be accepted (usually fewer than five) and the need for human decision when the inserted answer does not match the incorrect or correct list of options. Recent developments have further expanded this potential by using fully automated marking of open-answer and small narrative items. Using dictionaries of words and natural language processing, these computer algorithms can mark small pieces of narrative text with better precision than a human judge [29].

Score Reporting and Feedback Provision

Using two case studies, we now address the utility of technology in developing score reporting and feedback provision.

HOW TECHNOLOGY IS IMPACTING OPERATIONS

Testing Modality (Adaptive Testing Models)

In computer-adaptive tests (CATs), an algorithm tailors a unique examination for each candidate to maximize the precision of measurement at his or her specific ability level with the fewest number of items possible. It does so by using IRT as a core model [28]. In one CAT scenario, the probability of a correct response is computed for an initial item of average difficulty using an **R** T model. The candidate's answer to this item determines whether an easier or more difficult question is subsequently administered—that is, the examination "adapts" to the candidate's ability until precision of measurement can no longer be improved. Complementary to the gains in measurement precision, CAT needs a shorter period of assessment and a smaller number of items to estimate the candidate's ability. This represents an obvious gain in time but also reduces the number of exposed items because only a small fraction of candidates sees the same item, and each candidate has a specific set of items to which he or she is exposed. In addition, CAT has been shown to improve learning and may be a future form of adaptative assessment and learning method [30]. Although

Case Study 7-1—Reports as Student Feedback and Remediation Guides

Problem:

The director of a course on clinical skills wishes to provide feedback to students on an objective structured clinical examination (OSCE) test regarding three dimensions: clinical interview, physical examination, and planning. These placements vary in duration, and the moment within the course. How might the director provide meaningful and accurate feedback to students?

Solution:

The clinical skills course assesses learners through a series of 6 OSCE stations in three moments during the course. The assessment blueprint needs to be carefully designed so that each of the three dimensions is adequately sampled in each placement. The format of the OSCE stations should be homogenous to avoid presentation bias. Each station checklist must include items that are related to all dimensions in the blueprint but are specific to station content. For example, if the assessment outcome is the assessment of family history in one station, the item would be "The student identifies mother had cancer?" while another station might be "The student identifies the father has diabetes?". This way the reporting of the general performance on family history taking can be reported without disclosing the checklist. This feedback can be used by students to improve specific dimensions in their clinical skills from one assessment moment to another.

Take Home message:

Careful planning of a blueprint and characterization of metadata is essential to ensure adequate sampling and data driven reporting to students. These reports are essential to allow students to reflect on specific aspects of their performance and search for remediation strategies.

CATs are attractive because they can yield precise measurements of abilities with shorter test lengths, especially large item banks are necessary to sustain such programs using sophisticated software packages [28].

Test Security

Test security is critical to ensure that the assessment process is fair and valid [31]. This is particularly important in high-stakes settings such as licensing

Ch07.indd 182 29-05-2021 16:23:17

Case Study 7-2—Digital Score Reporting

Problem:

The head of assessment at a medical school wishes to report on scores of the progress test taken by students of all 6 years of a medical course. These happen every quarter within the course. How might learners be reassured that their scores are equivalent to the same standard and achieve the same learning outcomes?

Solution:

The progress test assesses all areas of the medical course through a combination of four blueprint dimensions: discipline (e.g., anatomy, cardiology, etc.), organic system (e.g., pulmonary, urinary, etc.), physician task (e.g., diagnosis, treatment), and cognitive task (e.g., memory, decision, etc.). Because students at different levels of development have different performances, raw scores would not be comparable among groups. To overcome this difficulty, scores should be reported based on reference to a norm. Adequate reporting should include information regarding each learner's position in the cohort performance curve with a visual depiction of the group's performance. Central and dispersion measures should be reported to each dimension of the blueprint. This way students understand their relative performance toward their own cohort and understand the relative difficulty of each area within the blueprint without disclosing any items.

Take Home Message:

Informative score reporting is critical for learners to understand their areas of strengths and areas to improve. Visual representation of the student's position in the cohort's performance makes it more tangible and relevant.

examinations and testing for competitive processes such as career choice. Although this has been a core issue in the assessment field, it remains secretive; many institutions withhold information on security breaches. So, it is unclear to which extent test security is a small or big concern.

CBT testing has improved testing security by reducing the number of intermediaries involved in the preparation and administration of tests. Establishing a direct link between servers in the licensing bodies and the test taker's computer enables higher protection. In addition, the use of security facilities with a controlled environment greatly

Ch07.indd 183 29-05-2021 16:23:17

reduces the chances of collusion and person-in-the-middle type of interference. In addition, electronic assessment provides additional sources of information such as keystroke, mouse tracking, response times, and other parameters that relate to the test taker's behavior during the exam that can be used to analyze suspected security breaches in the postadministration period [32].

Item Review Processes

Item review is an essential step to ensure that test items have high quality and provide accurate and reliable measurements of performance [33]. The establishment of item review committees has long been advocated. Licensing bodies such as certification boards and professional item developers have dedicated teams to this task that thoroughly review items before administration. Yet, many small and middle-sized institutions struggle to implement them. This is partly because of the amount of time and resources that are needed to implement such procedures. Many software packages that are specialized in the assessment process now provide the possibility to develop item banks in a collaborative and hierarchical manner, thus providing an easier and less expensive way to implement item review committees. Moreover, the possibility to interact in a virtual nonsynchronous environment provides a chance for people from different geographical regions and time zones to collaborate in international projects, something that would demand significant amounts of money and resources when done on-site. These online solutions typically provide a digital fingerprint to the entire development cycle of an item that can be used in auditing processes.

TECHNOLOGY IN ASSESSMENT—PROGNOSTICATIONS AND CONFABULATIONS

Assessment has advanced greatly in the last few decades. Computational capacity that was once reserved for big technological companies can now be achieved with a personal computer. In addition, open-source code and a growing number of data scientists have taken the ultimate step by democratizing computer programs that provide the tools for advanced assessment techniques. Artificial intelligence (AI) tools will become providers of old and novel formats of assessment in a near future. Conversely, as we grow comfortable with the use of natural language processing and other forms of machine learning, we will be able to delegate traditionally human-dependent tasks such as scoring essays by computer. Eventually, this will free our time to gain efficiency—for example, by developing better AIG models and developing

Ch07.indd 184 29-05-2021 16:23:17

COGNITIVE MODEL

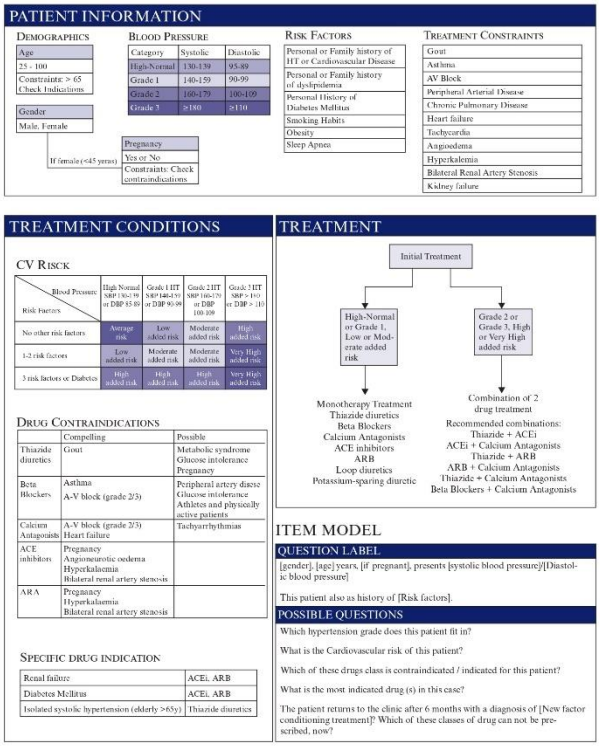


Figure 7-1 • Top: Example of a cognitive model and respective item model. Bottom: Two examples of items derived from the cognitive model.

novel forms of assessment such as a virtual patient encounter with real-life dialogues or computer-generated imagery for virtual patients. As data and knowledge grow, we will be able to produce better statistics and make better and faster decisions. It is likely that in a not-so-distant future, computers will be able to combine the automation of item generation and the application of adaptative processes on the fly. The exclusion of the

A 43-year-old man comes to the medical office. The patient has a history of tachyarrhythmia. Physical examination shows blood pressure of 131/87 mm Hg.
What is the absolute risk of this patient?
A. Average risk
B. Low added risk
C. Average added risk
D. High added risk

A 52-year-old man comes to the medical office. The patient has a history of home measured morning blood pressure of 145/95 mm Hg.
What degree of hypertension does this patient have?
A. Isolated systolic hypertension
B. Grade 1 hypertension
C. Grade 2 hypertension
D. Grade 3 hypertension

human factor in the scoring could represent the fairest way to assess one's performance. This would represent the ultimate personalized, precise, and fairest assessment tool.

CONCLUSION

Human knowledge grows exponentially each year and will be accelerat-ing with the dominance of AI. In the presence of *big data*, decision-making has become incredibly complex. It is impossible for professionals to cognitively contend with this complexity. The same holds true with assessment and measurement models. Current psychometric models do well with the kind of data we collect using traditional tools, but they do poorly with sparse and unstructured data. AI tools will potentially uncover data structures that humans simply cannot identify, which leads to new ways and systems for creating more authentic and complex learning and testing scenarios.

Computer technology is not the problem. It is the solution.

The use of diagnostic, management, data mining, and summarization algorithms is already drastically altering education and assessment. The professional's role will be even more critical in regard to the introduction, evaluation, and best use of these technologies in their role as "advisor and knowledge navigator." AI and assessment will be a "team sport" predicated on a set of new competencies (statistics, computer sciences, etc.)

REFERENCES

1. Gier MJ, Lai H. Using automated processes to generate test items and their associated solutions and rationales to support formative feedback. *Interaction Design and Architecture(S)*. 2015;25(1):9–20.

Ch07.indd 186 29-05-2021 16:23:19

2. Gierl MJ, Haladyna TM. Automatic item generation: An introduction. In: Gierl MJ, Haladyna, eds. *Automatic Item Generation: Theory and Practice*. Routledge; 2012. doi:10.4324/9780203803912.
3. Gierl M, Lai H, Hogan J, Matovinovic D. A Method for Generating Educational Test Items That Are Aligned to the Common Core State Standards. *Journal of Applied Testing Technology*. 2015;16(1):1–18.
4. Gierl MJ, Lai H. A Process for Reviewing and Evaluating Generated Test Items. *Educational Measurement: Issues and Practice*. 2016;35(4):6–20. Retrieved from https://doi.org/10.1111/emip.12129.
5. Gierl MJ, Lai H, Pugh D, et al. Evaluating the Psychometric Characteristics of Generated Multiple-Choice Test Items. *Applied Measurement in Education*. 2016;29(3):196–210. Retrieved from https://doi.org/10.1080/08957347.2016.1171768.
6. Medrano LA, Liporace MF, Pérez E. Computerized assessment system for academic satisfaction (ASAS) for first-year university student. *Electronic Journal of Research in Educational Psychology*. 2014;12(2):541–562.
7. Gierl MJ, Lai H. The Role of Item Models in Automatic Item Generation. *International Journal of Testing*. 2012;12(3):273–298. Retrieved from https://doi.org/10.1080/15305058.2011.635830.
8. Kosh AE, Simpson MA, Bickel L, Kellogg M, Sanford-Moore E. A Cost–Benefit Analysis of Automatic Item Generation. *Educational Measurement: Issues and Practice*. 2019;38(1):48–53. Retrieved from https://doi.org/10.1111/emip.12237.
9. David MFB, Davis MH, Harden RM, et al. AMEE Medical Education Guide No. 24: Portfolios as a Method of Student Assessment. *Medical Teacher*. 2001;23(6):535–551. Retrieved from https://doi.org/10.1080/01421590120090952.
10. Gierl MJ, Lai H. Evaluating the Quality of Medical Multiple-Choice Items Created with Automated Processes. *Medical Education*. 2013;47(7):726–733. Retrieved from https://doi.org/10.1111/medu.12202.
11. Batalden P, Leach D, Swing S, Dreyfus H, Dreyfus S. General Competencies and Accreditation in Graduate Medical Education. *Health Affairs*. 2002;21(5):103–111. Retrieved from https://doi.org/10.1377/hlthaff.21.5.103.
12. Holmboe ES, Ward DS, Reznick RK, et al. Faculty Development in Assessment: The Missing Link in Competency-Based Medical Education. *Academic Medicine*. 2011;86(4):460–467. Retrieved from https://doi.org/10.1097/ACM.0b013e31820cb2a7.
13. Gierl MJ, Lai H, Turner SR. Using Automatic Item Generation to Create Multiple-Choice Test Items. *Medical Education*. 2012;46(8):757–765. Retrieved from https://doi.org/10.1111/j.1365-2923.2012.04289.x.
14. Royal KD, Hedgpeth M-W, Jeon T, Colford CM. Automated Item Generation: The Future of Medical Education Assessment? *EMJ Innov*. 2018;2(1):88–93.
15. McCoubrie P. Improving the Fairness of Multiple-Choice Questions: A Literature Review. *Medical Teacher*. 2004;26(8):709–712. Retrieved from https://doi.org/10.1080/01421590400013495.
16. Sinharay S, Johnson MS, Williamson DM. Calibrating Item Families and Summarizing the Results Using Family Expected Response Functions. *Journal*

of Educational and Behavioral Statistics. 2003;28(4):295–313. Retrieved from https://doi.org/10.3102/10769986028004295.

17. Arendasy ME, Sommer M. Using Automatic Item Generation to Meet the Increasing Item Demands of High-Stakes Educational and Occupational Assessment. *Learning and Individual Differences*. 2012;22(1):112–117. Retrieved from https://doi.org/10.1016/j.lindif.2011.11.005.
18. Choi J, Kim H, Pak S. Evaluation of Automatic Item Generation Utilities in Formative Assessment Application for Korean High School Students. *Journal of Educational Issues*. 2018;4(1):68. Retrieved from https://doi.org/10.5296/jei.v4i1.12630.
19. Arendasy M, Sommer M. 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*. 2007;17(4):366–383. Retrieved from https://doi.org/10.1016/j.lindif.2007.03.005.+
20. Gierl MJ, Lai H. Instructional Topics in Educational Measurement (ITEMS) Module: Using Automated Processes to Generate Test Items. *Educational Measurement: Issues and Practice*. 2013b;32(3):36–50. Retrieved from https://doi.org/10.1111/emip.12018.
21. Jozefowicz RF, Koeppen BM, Case S, et al. The Quality of In-House Medical School Examinations. *Academic Medicine*. 2002;77(2):156–161. Retrieved from https://doi.org/10.1097/00001888-200202000-00016.
22. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Using Cognitive Models to Develop Quality Multiple-Choice Questions. *Medical Teacher*. 2016;38(8):838–843. Retrieved from https://doi.org/10.3109/0142159X.2016.1150989.
23. Blum D, Holling H. Automatic Generation of Figural Analogies with the IMak Package. *Frontiers in Psychology*. 2018;9:1–13. Retrieved from https://doi.org/10.3389/fpsyg.2018.01286.
24. Gunabushanam G, Taylor CR, Mathur M, Bokhari J, Scoutt LM. Automated Test-Item Generation System for Retrieval Practice in Radiology Education. *Academic Radiology*. 2019;26(6):851–859. Retrieved from https://doi.org/10.1016/j.acra.2018.09.017.
25. Lai H, Gierl MJ, Touchie C, et al. Using Automatic Item Generation to Improve the Quality of MCQ Distractors. *Teaching and Learning in Medicine*. 2016;28(2):166–173. Retrieved from https://doi.org/10.1080/10401334.2016.1146608.
26. Sam AH, Field SM, Collares CF, et al. Very-Short-Answer Questions: Reliability, Discrimination and Acceptability. *Medical Education*. 2018;52(4):447–455. Retrieved from https://doi.org/10.1111/medu.13504.
27. Sam AH, Peleva E, Fung CY, et al. Very Short Answer Questions: A Novel Approach to Summative Assessments in Pathology. *Advances in Medical Education and Practice*. 2019;10:943–948. Retrieved from https://doi.org/10.2147/amep.s197977.

Ch07.indd 188 29-05-2021 16:23:19

28. De Champlain AF. A Primer on Classical Test Theory and Item Response Theory for Assessments in Medical Education. *Medical Education*. 2010;44(1):109–117. Retrieved from https://doi.org/10.1111/j.1365-2923.2009.03425.x.
29. Latifi S, Gierl MJ, Boulais AP, De Champlain AF. Using Automated Scoring to Evaluate Written Responses in English and French on a High-Stakes Clinical Competency Examination. *Eval Health Prof*. 2016 Mar;39(1):100–113.
30. Martin AJ, Lazendic G. Computer-Adaptive Testing: Implications for Students' Achievement, Motivation, Engagement, and Subjective Test Experience. *Journal of Educational Psychology*. 2018;110(1):27–45.
31. American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association; 2014. ISBN-10: 0935302352, ISBN-13: 978-0935302-35-6. Retrieved from https://www.aera.net/Publications/Books/Standards-for-Educational-Psychological-Testing-2014-Edition.
32. McManus IC, Lissauer T, Williams SE. Detecting Cheating in Written Medical Examinations by Statistical Analysis of Similarity of Answers: Pilot Study. *BMJ*. 2005;330(7499):1064–1066. doi:10.1136/bmj.330.7499.1064.
33. Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB. Use of a Committee Review Process to Improve the Quality of Course Examinations. *Advances in Health Sciences Education*. 2006;11:61–68.

Ch07.indd 189 29-05-2021 16:23:19

CHAPTER III

FEASIBILITY ASSURANCE: A REVIEW OF AUTOMATIC ITEM GENERATION IN MEDICAL ASSESSMENT

This chapter is based on the following publication: **Falcão, F**., Costa, P., & Pêgo, J. M. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education, 27*(2), 405-425. DOI: 10.1007/s10459-022-10092-z

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
© The Author(s), under exclusive licence to Springer Nature B.V. 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

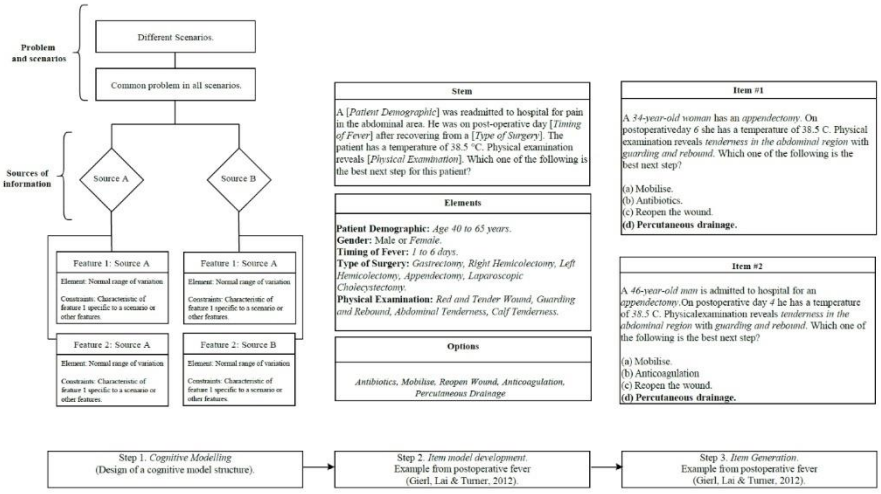


Fig. 1 AIG three-step process for generating medical MCQs

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

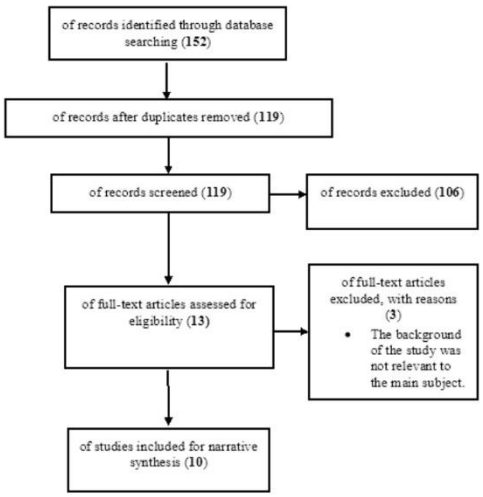


Fig. 2 Flow chart of the included studies

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
(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	1	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	0	2	2	0	2	2	2	2	0	2
(Pugh et al., 2016)	2	2	0	2	2	0	0	2	1	2	0	0
(Pugh et al., 2020)	2	2	0	2	2	0	0	2	2	2	0	2
(Shappell et al., 2020)	2	2	0	2	2	0	0	2	2	2	0	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.

CHAPTER IV

A SUGGESTIVE APPROACH FOR ASSESSING ITEM QUALITY, USABILITY AND VALIDITY OF AUTOMATIC ITEM GENERATION

This chapter is based on the following publication: **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*, 1-25. DOI: 10.1007/s10459-023-10225-y

Advances in Health Sciences Education
https://doi.org/10.1007/s10459-023-10225-y

RESEARCH

A suggestive approach for assessing item quality, usability and validity of Automatic Item Generation

Filipe Falcão[1,2,3] · Daniela Marques Pereira[1,2,3] · Nuno Gonçalves[2,3] · Andre De Champlain[4] · Patrício Costa[1,2] · José Miguel Pêgo[1,2,3]

Received: 6 September 2022 / Accepted: 19 March 2023
© The Author(s) 2023

Abstract

Automatic Item Generation (AIG) refers to the process of using cognitive models to generate test items using computer modules. It is a new but rapidly evolving research area where cognitive and psychometric theory are combined into digital framework. However, assessment of the item quality, usability and validity of AIG relative to traditional item development methods lacks clarification. This paper takes a top-down strong theory approach to evaluate AIG in medical education. Two studies were conducted: Study I—participants with different levels of clinical knowledge and item writing experience developed medical test items both manually and through AIG. Both item types were compared in terms of *quality* and *usability (efficiency and learnability)*; Study II—Automatically generated items were included in a summative exam in the content area of surgery. A psychometric analysis based on Item Response Theory inspected the validity and quality of the AIG-items. Items generated by AIG presented quality, evidences of validity and were adequate for testing student's knowledge. The time spent developing the contents for item generation (cognitive models) and the number of items generated did not vary considering the participants' item writing experience or clinical knowledge. AIG produces numerous high-quality items in a fast, economical and easy to learn process, even for inexperienced and without clinical training item writers. Medical schools may benefit from a substantial improvement in cost-efficiency in developing test items by using AIG. Item writing flaws can be significantly reduced thanks to the application of AIG's models, thus generating test items capable of accurately gauging students' knowledge.

Keywords Automatic Item Generation · Item quality · Item writing · Usability · Validity

✉ José Miguel Pêgo
jmpego@med.uminho.pt

[1] Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Largo Do Paço, 4710-057 Braga, Portugal

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

[3] iCognitus4All – IT Solutions, Braga, Portugal

[4] AstraZeneca, Gaithersburg, USA

Published online: 25 April 2023

Introduction

The majority of knowledge tests used in medical education, within various medical content areas, are based on multiple-choice questions (MCQ). While MCQs are effective in gauging student's knowledge, developing them demands in depth-comprehension of both the subject-matter and the purpose of the assessment, which makes them time-consuming and expensive to write. Automated Item Generation (AIG) is a model-based next-generation assessment method where computer algorithms generate items together with estimates of their psychometric parameters. It is a contemporary method that can scale the item development process for medical schools, producing large numbers of high-quality items both quickly and efficiency. Despite this promising framework, the item quality, usability, and validity of items generated by AIG relative to those manually written remains generally unexplored. Through a multimodal strategy based on the evaluation of medical test items developed by participants with varied levels of clinical expertise and item writing experience (ranging from no training to professional practice), as well as on the psychometric analysis of the item parameters of a medical summative exam, we aim to: (i) compare the quality of AIG-items vs manually written items (*item quality*); (ii) measure the efficiency of AIG and explore its learnability properties (*usability*) and; (iii) search for sources of evidence that support AIG as a valid item development method (*validity*). Our work contributes to the literature by assessing whether AIG can meet the quality, usability and validity standards expected of traditional testing methods that are used in educational assessment.

Background

The role of multiple-choice questions in medical education

The most popular tool for testing in medical education are exams comprised of multiple-choice questions (MCQs) (Douthit et al., 2021; Grainger et al., 2018). Besides allowing for instant scoring, MCQs are suitable for achievement testing since they cover different skills and enable the assessment of numerous candidates in a cost and resource efficient manner (Royal et al., 2018; Rust et al., 2020).Writing MCQs is complex since experts must outline and replicate cognitive problem-solving skills (Billings et al., 2020; Gierl et al., 2012). In order to plan this procedure, experts must share their knowledge and collaborate behind the scenes. Manual (or traditional) item writing involves at least five stages: (i) hiring item writers to broaden the approaches of content experts (Sinharay et al., 2003); (ii) training item writers to develop new items; (iii) detailed reviews; (iv) empirical try-outs to calibrate item parameters (de Chiusole et al., 2018; Embretson & Kingston, 2018) and; (v) the inclusion of new items in operational exams based on blueprint specifications and empirical properties (Embretson & Kingston, 2018).

This complex set of procedures adds expense and time to the item development process, limiting the availability of items needed for proper assessment (Gierl & Haladyna, 2012). Although they are feasible for lower-stakes exams, these approaches are less appealing for large-scale exams (Gierl et al., 2012). The limitations of this procedure are even more pronounced when large number of items are needed to build item banks—pools of items containing information about their content and psychometric details—or parallel test parcels, where each item is treated as an isolated entity (Gierl et al., 2022a; Gierl & Lai, 2013a). Since each item is designed, reviewed and formatted individually, unanticipated psychometric results may arise due to spurious elements that were missed throughout the

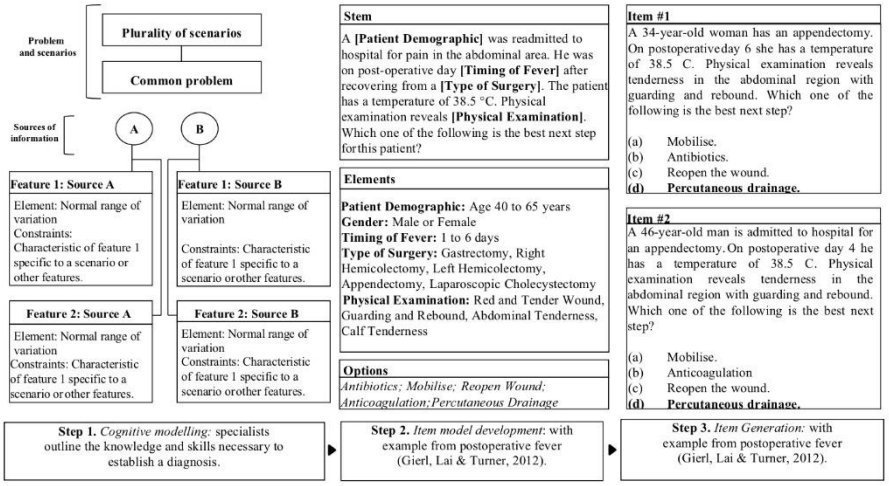


Fig. 1 Three-step process for generating medical MCQs based on AIG

development process (Mindyarto et al., 2018). Additionally, since exams are continuously administered, item exposure becomes a concern over time (Sinharay et al., 2003). Consequently, it would take thousands of MCQs to develop secure item banks (Gierl & Lai, 2013a). The restocking of new items is essential for assuring quality assessment, especially in high-stakes testing (Embretson & Kingston, 2018). As educators face pressure to develop large numbers of new items, alternative methods of item development become necessary (Gierl & Lai, 2013a; Jozefowicz et al., 2002).

Automatic item generation as a solution

In order to address the drawbacks of conventional item writing, modern psychometrics research explored the use of computer algorithms to construct test items (Falcão et al., 2022; Harrison et al., 2017). The use of algorithms is novel and might potentially yield an infinite number of items for assessment (von Davier, 2018). Among the approaches, *Automatic Item Generation* (AIG) refers to the process of employing cognitive models, i.e. structured content reflecting the sources of information that underly the cognitive process, to generate test items with the aid of computer technology (Gierl et al., 2022a). It represents an emerging cutting-edge assessment development method that mixes human experience with computer algorithms, promising to generate vast amounts of new items in a short time from a single model (Choi, 2020; Gierl et al., 2021; Gierl et al., 2022a). Simply stated, AIG may be conceptualized of as an approach where algorithms generate items along with estimates of their psychometric parameters due to the merging of cognitive and psychometric theory (Choi, 2020; Kosh et al., 2019; Mindyarto et al., 2018).

The generation of medical MCQs using AIG is predicated on a scalable content development method based on a strong theory-driven approach in which a three-stage process generates items considering clinical scenarios (Cf. Figure 1). In the first stage, the content for item generation is identified through a cognitive model (Gierl & Lai, 2016). Developing this structure requires the selection of a topic from the test blueprint and a content expert to summarize a typical approach to make a medical diagnosis (Gierl & Lai, 2016). During

this process, the cognitive and content specific informations needed to solve the problem are organized into a coherent structure that contains the content relationships and information sources necessary to formulate a medical diagnosis in a specific content area (Gierl et al., 2022a; Lai et al., 2016). These models list cognitive and content-specific information, characterize higher-order versus lower-order forms of thinking to solve problems, and provide a theoretical understanding of test performance (Gierl & Lai, 2016; Leighton & Gierl, 2011). Once the cognitive models are determined, the stimulus features will be determined in accordance with these models (Yang et al., 2022). In the second stage, an item model is designed to outline where the content from the cognitive model should be placed in order to generate new items (Gierl et al., 2022a). This model will specify which parts of the item can be manipulated, including the context, content, and the question that examinees are required to answer (Gierl et al., 2012). For a typical MCQ, it includes the stem (part of the item model that compiles the information needed to solve the problem), response options (all answers, including distractors and correct options), the lead-in-question and other ancillary information (Gierl et al., 2022a). Finally, in the third stage, algorithms integrate the content from the cognitive model outlined in step 1 into the item model developed in step 2 (Gierl et al., 2022a). Computer algorithms are used to conduct this process because the assembly often involves challenging combinatorial operations (Gierl et al., 2022a). The generation procedure can be viewed as an iterator that permutes over all possible value combinations, removing any that do not adhere to the restrictions previously established in the cognitive model (Gierl et al., 2022a).

Unlike conventional item development methods, where each item is treated as the unit of analysis and is created individually, AIG treats the model as the primary unit of analysis where a single model is utilized to generate numerous items (Gierl et al.,). Consequently, item development is linked to the number of available models rather than the number of experts (Gierl et al., 2022a). Through this approach, it is possible to specify and manage item features (i.e., the elements in the assessment task that must be manipulated) that predict test performance since cognitive features are defined in great detail. As a result, the generated items can potentially be calibrated without the need for extensive field or pilot testing because it is possible to specify the criteria that influences item difficulty (Gierl & Lai, 2012; Leighton & Gierl, 2011). Importantly, the cognitive models integrate aspects of active learning methologies and theoretical constructs that are thought to underpin the thought process of clinical reasoning, such as illness scripts (Schmidt & Mamede, 2015). Additionally, AIG may significantly help faculty members avoiding common item construtions errors. Due to the fact that the majority of faculty members have little to no formal training in item construction, adopting standardized templates such as the ones used within cognitive model framework that AIG uses for item development can also help them improve their item writing skills and generate more precise assessments tools of student's knowledge (Royal et al., 2018).

AIG versus manual item writing

AIG differs from more traditional item development methodologies in several aspects. First, AIG is more intensive in regard to its development and revision processes. Since authors must set values for substitutions in item models and provide instances of item stems and distractors to ensure proper generation of new items, it stands to reason that AIG may require more time/skills than manual item writing for initial item development. Second, the technological environment must also be considered. Since AIG employs algorithms to

generate items, this may require more effort from software engineers and subsequent quality evaluations than manually written items. However, the main difference between both processes lies in the fact that traditional item writing is usually based on "single instances", often without reference to a broader model. AIG, in fact, requires a more naturalistic approach to a clinical problem that is carefully defined based on a theoretical strong framework (Kosh et al., 2019; Pugh et al., 2016), evolving from the symptom/presentation to the clinical manifestations that characterize the conditions that are compatible with it.

Even with these limitations, AIG yields significantly more benefits than manual item writing. Its main asset lies in the generation of large sets of brand-new items, avoiding time-consuming processes typical of manual item writing and promising great resources for item banks (Harrison et al., 2017; Jendryczko et al., 2020). Nevertheless, research into the item quality, usability, and validity of AIG in comparison to manual item development approaches is still in its infancy. To date, only a small number of operational testing scenarios have used items generated using cognitive modeling procedures (Gierl et al., 2016). It will be challenging to defend the implementation of AIG in medical assessment until this subject is thoroughly investigated (Gierl & Lai, 2013a). Therefore, it is crucial to evaluate both item development approaches in light of the following characteristics:

Item quality. There is considerable worry that automatically generated items may not be as high-quality as those created using conventional techniques (Pugh et al., 2020). Furthermore, the quality and the psychometric properties of generated items to multiple-choice testing are mostly undocumented (Gierl et al., 2016; Shappell et al., 2021). Generated items need to be of a high caliber for AIG to be truly useful (Gierl et al., 2022b). The few studies that have compared manually and AIG written items generally support the use of the latter. However, the results provided are not consistent and need further evidence. Gierl and Lai (2013a) submitted AIG-items and manually written items to a panel of medical experts, who blindly reviewed the items for quality. The findings suggested that both item types were comparable on almost all quality indicators. However, the authors discovered that AIG-items could be distinguished from manually written items in terms of quality of the distractors, which were rated more poorly for the AIG-items (Gierl & Lai, 2013a). On the other hand, Pugh and colleagues (2020) found that AIG-items, including distractors, were not perceived as different from items created by manual processes. The primary reason behind this discrepancy in findings can be traced back to the fact that post-2013 AIG-engines were enhanced to generate better distractor relationships (Lai et al., 2016). However, this discrepancy yet reveals a lack of clarity regarding the quality of both item types and suggest a need for research. Furthermore, these studies only compared highly trained item writers that tend to produce high quality items. Such expertise, typical of licensing bodies, does not necessarily generalize at the medical school level.

Usability (or feasibility of implementation). Since AIG is based on a human–computer interaction, it is important to compare it to manual item writing methods in terms of *usability* (Jeng, 2005). Usability pertains to the relationships between users, tasks, and environments and is measured in terms of *efficiency* and *ease of use (learnability)* (Jeng, 2005; Lewis, 2016). Studies comparing both methodologies with respect to usability are scarce and do not generalize to a broader context. Gierl and Lai (2013a) report that their participants were able to generate 9496 items after learning the principles of AIG. Prior to learning, the same participants were able to manually develop 25 items. Similarly, Pugh and colleagues (2016) report that a group of inexperienced item writers was able to develop a complete cognitive model within about 2 h, resulting in a set of 5–10 automatically generated items. In short, the available literature seems to give AIG the advantage in regard to usability (Gierl & Lai, 2013a; Harrison et al.,

2017). Although AIG has indeed improved the efficiency of item development (Gierl et al., 2012; Pugh et al., 2020), its learnability properties (i.e., the quantity of items that an individual can generate in a particular period of time considering their item writing experience/knowledge) has not been studied, which is why research is needed.

Validity. While technology makes it easier to administer test contents to students, it also raises questions regarding test security and validity (Patel, 2021).Validation processes are key to ensuring the validity of assessment tools and consists of assembling evidence to provide a scientific basis for the interpretation of scores (American Educational Research Association, 2018). Item development is essential to validity assessments because it records the procedures and results required to produce high-quality test content (Gierl et al., 2022b). A fundamental issue for AIG research regards the validation of its processes (Gierl et al., 2022b; Shin, 2021). To date, few research has been devoted to collecting evidence to support the validity of AIG (Falcão et al., 2022; Gierl et al., 2022b; Rafatbakhsh et al., 2020). Since incorporating cognitive models into test design and development is required to support validity arguments for test-based inferences, we have reason to suppose that AIG incorporates validity evidence in its methods (Gierl et al., 2022b; Leighton & Gierl, 2011). Therefore, by obtaining evidences of validity, we would be taking a critical step to support greater use of AIG in educational assessment.

A top-down, strong theory approach

AIG promises to ease the item development burden (Arendasy & Sommer, 2007; Gierl & Lai, 2012). However, it is critical to understand its benefits and drawbacks (Bejar et al., 2003). This paper employs a strong theoretical approach to evaluate AIG in medical assessment. Over the course of two complementary studies, we evaluated AIG in terms of quality, usability and validity. Our objectives were as follows: (i) to compare the quality of AIG-items vs manually written items (*item quality*); (ii) to measure the efficiency of AIG and explore its learnability properties (*usability*); (iii) to search for sources of evidence supporting AIG as a valid item development approach (*validity*). Our paper contributes to the literature by demonstrating whether AIG can meet the quality, usability and validity standards expected of conventional item development frameworks that are currently used in medical assessment.

Methods

General procedure

In the first study, a group of participants developed items in different areas of surgery through AIG and manual writing methods. Both item development approaches were evaluated in terms of *quality* and *usability*. In the second study, an item sample composed of manually written and AIG-items was included in a summative exam of a general surgery curricular unit of the medical course at the School of Medicine of the University of Minho (EMUM). Both item types were evaluated in terms of *item quality* and *validity*.

Software

R software (version 4.0.5) and R packages "*Psych*" (Revelle & Revelle, 2015), "*mirt*" (Chalmers, 2012), and "*eRm*"(Patrick et al., 2018) were used to conduct statistical analyses. JASP (version 0.15) and R package "*Ggplot2*" (Wickham et al., 2016) were used for graphical representations of the data.

Study I

Sample

Twenty participants (N = 20) with varied levels of clinical knowledge and item writing experience volunteered to take part in the study. Participants were clinical teachers (n = 4), non-clinical teachers (n = 2), non-teacher clinicians (n = 2), members of the academic staff (n = 5) and students (n = 7) from the EMUM. Six participants (three students, one staff member, one clinical teacher, and one clinician) claimed to be aware of AIG prior to the study.

Procedure

The item quality and usability of AIG were compared to manually written items using a multi-methods approach. The investigation was conducted through seminars held by the authors in group sessions of no more than five people. Each participant attended one session. Participants were required to complete four tasks during the course of the seminars:

Task 1: Manual item development. Participants manually developed one MCQ in a particular domain of surgery (*jaundice; abdominal pain; shoulder pain; knee pain,* or *hematemesis*). They had access to the internet and other sources of information (e.g., MSD manuals) during this process. The time spent developing the item was recorded. The number of characters in the stem, lead-in-question, and option-set of the item, was tallied.

Task 2: Cognitive model and item model development. Participants learned the basics of AIG in a workshop taught by the authors of this paper. After learning the logic and instructions to operationalize the cognitive models, participants were asked to develop one cognitive model in a different surgical domain than the one used in the first task. Time spent developing the cognitive model was recorded.

Task 3: Item generation. Participants employed algorithms in the cognitive models to generate new items using comercial proprietary software developed for the purpose of this study (AIG module, *QuizOne*®). Item generation time; the number of items generated; and the number of characters in the stem, lead-in-question and in the option-set of each item were counted.

Task 4: Survey. Participants answered a survey related to their *confidence* in AIG and its *usefulness*. The survey assessed different facets of respondents' perceptions regarding their confidence in AIG (software used; use of AIG for their own assessments; developing a cognitive model) and its usefulness (measuring knowledge; understanding clinical reasoning; improving item quality). Participants evaluated the different facets on a scale between one ("*not confident at all";"not useful at all"*) and five ("*totally confident"; "very useful"*). The survey received responses from all respondents.

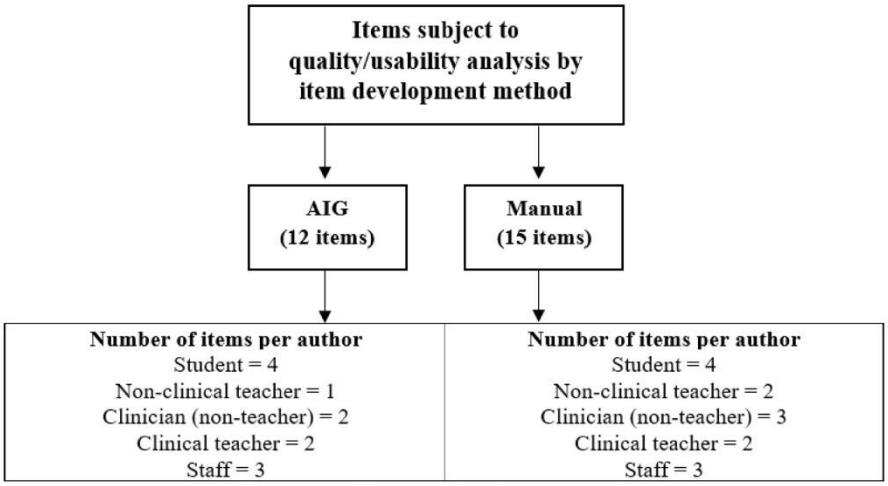


Fig. 2 Sample of items used for analysis

After the seminars, a random sample of twenty-seven items (AIG = 12; Manual = 15) developed during the seminars was subjected to a quality and usability analysis. The cognitive models that generated the AIG-items were identified. To ensure that there was no overlap or implausible items during the generation and sampling, all selected items were blind checked for content similarity and plausibility before inclusion in the test form. The method of item development and the author of each item that made up this sample are detailed in Fig. 2. Each of these items was blindly reviewed by a test development expert with experience in reviewing and evaluating MCQs. Judgments from the specialist evaluated the quality of the items (Gierl et al., 2016). The expert did not attend any of the seminars and was not informed about the purpose of the study.

Statistical analysis

Preliminary analysis. A series of t-tests was performed to determine whether the sample items were comparable in terms of format. This analysis allowed us to verify whether both item types were aesthetically distinct (overall structure) in terms of number of characters in the stem, option-set and lead-in-question, which could bias our results. Independent groups Student t-tests were conducted to test for differences between both item types in regard to the number of characters in the stem. Mann–Whitney tests were computed to test for differences between both item types with respect to the number of characters in the option-set and lead-in-question.

Item quality. Items may be linked to evidence that informs judgements of their quality (Gierl et al., 2022b). We used Jozefowicz's and colleagues (Jozefowicz et al., 2002) item quality rating scheme as a method to measure the *quality* of each question. The rating scheme was designed by its authors to reflect accepted item-writing principles. Each item was rated on a scale from one (*"the item tested recall only and was technically flawed")* to five (*"the item used a clinical or laboratory vignette, required reasoning to answer, and was free of technical flaws*"). Descriptive statistics were calculated to summarize our data.

Mann–Whitney tests were used to see whether there were statistically significant differences in the quality ratings between the two item types.

Usability (efficiency and learnability). Efficiency occurs when outputs are produced with low levels of resources (Johnes et al., 2017). An independent groups Student t-test was used to assess differences in the time spent developing a single item manually vs a cognitive model for AIG. Mann–Whitney tests were employed to assess differences in the generation time and in the number of items developed by both development methods. Learnability was measured using analysis of variance (ANOVAs). As most faculty members in medical schools are required to create their own testing items (Royal et al., 2018; Schmeiser & Welch, 2006), and given that the knowledge and item writing experience of the item writers makes each item unique during its development process (Gierl et al., 2022b), it is relevant to evaluate the learnability of AIG in light of the participants' item writing experience and clinical knowledge. Considering only the AIG-items, two ANOVAs were conducted: (i) a two-way ANOVA to test for differences regarding the time spent developing a cognitive model, considering the item writing experience (*ItemXP*) and clinical experience/knowledge (*ClinicXP*) of the item author[1]; and (ii) a two-way ANOVA to test for differences in the number of items generated from the models considering the ItemXP and ClinicXP of the author. We retained a nominal type I error rate of 0.05 for all analyses. Data from the surveys were also used to inspect AIG's learnability. Descriptive statistics were used to summarize the data.

Results

Preliminary Analyses. No statistically significant differences were found between the number of characters in the stem, t (20.7) = 0.196, p = 0.846, d = 0.072, or in the lead-in-question, U = 109, p = 0.338, r = −0.184, of both item types. However, statistically significant differences were noted regarding the number of characters in the option-set, U = 156, p < 0.001, r = −0.620, with manually written items presenting more characters (Mdn = 119; IQR = 283) than AIG-items (Mdn = 82.5; IQR = 28.75). Considering that both item types were comparable in two out of three components of the selected MCQs, we believe that both item types were comparable and hardly distinguishable.

Item quality. The majority of AIG-items yielded high quality ratings (M = 4.92; SD = 0.289). Only one AIG-item was evaluated with a rating of four. The reason for this score was related to the detection of one item flaw (an implausible distractor). Ratings of manually written items were more variable (M = 2.93; SD = 1.53). Statistically significant differences in the quality ratings of both item types were found, U = 16, p < 0.001, r = -0.746, with higher quality ratings associated with AIG-items (Mdn = 5; IQR = 0) vs manually written items (Mdn = 4; IQR = 3).

Efficiency. Statistically significant differences were observed in regard to time spent developing items between both methods, t(25) = 6.39, p < 0.001, d = 2.44. The average time (in

[1] ItemXP of the participants was classified into three different levels (0: "inexperienced"; 1: "Novice"; 2: "experienced"). Members of the staff were classified as inexperienced (n = 3). Clinicians and students were considered as novices (n = 6) since they had knowledge of the MCQs used in tests. Non-clinical teachers and clinical teachers were considered as experienced (n = 3). ClinicXP was treated as a factor with two levels (0: "non-clinic"; 1:"clinic"). Clinicians, clinical-teachers and students from the fifth or sixth year of the medical course at the EMUM were considered as "clinic" (n = 7). Non-clinical teachers, members of the staff and students from the third year were considered as "non-clinic" (n = 5).

minutes) to develop a cognitive model for AIG was higher ($M=74.3$; $SD=19.5$) than when manually developing a single new item (M=31.5; SD=15.4). Furthermore, when applying a ratio comparison for the two times, we found that the time spent developing a cognitive model was 2.36 times as long as the time devoted to manually writing a single item.

Next, we compared the time algorithms spent generating items from the models vs the time spent by participants manually writing a single item. The Mann–Whitney test revealed statistically significant differences, $U=180$, $p<0.001$, $r=0.846$, with the algorithms producing new items faster ($Mdn=0.310$; $IQR=0.48$) than when manually writing a new item ($Mdn=27$; $IQR=27$). The time spent manually writing an item was 87.1 times as long as the time spent by the algorithm to generate new items. In conclusion, though it takes more resources to develop a cognitive model than manually writing a single item, the item yield was far more advantageous when using AIG since large sets of items are produced in a relatively short time. Therefore, one should consider AIG as a more efficient item development method.

Learnability. A two-way ANOVA was performed to analyse the effect of ItemXP and ClinicXP on the time spent developing a cognitive model. Neither ItemXP nor ClinicXP had a statistically significant effect on the time spent developing a cognitive model ($p_{ItemXP}=0.144$; $p_{clinicalXP}=0.216$). There was no significant interaction between the effects of ItemXP and ClinicXP ($F(4,7)=0.002$ $p=0.966$, $\eta_p^2=0$). The time spent developing the cognitive models did not vary as a function of the participants' ItemXP or ClinicXP. This is a point in favour of AIG's learnability, as inexperienced participants were able to build cognitive models and took approximately the same time as experienced participants to design them.

A second two-way ANOVA analysed the effects of ItemXP and ClinicXP on the number of items generated from the cognitive models. The main effect of ItemXP was significant, F (2, 7)=12.3, $p<0.05$, ηp2=0.779, suggesting that the number of items generated by novice item writers (M=393; SD=439) was greater than experienced (M=188; SD=184) or inexperienced item writers (M=313; SD=94.4). The main effect of ClinicXP was also significant, F (1, 7)=13.8, $p<0.05$, ηp2=0.663, suggesting that the number of items generated by non-clinics (M=469; SD=443) was greater than clinicians (M=217; SD=164). These main effects were qualified by a significant interaction between ItemXP and ClinicXP on the numbers of items generated, F (1, 7)=16.3, $p<0.05$, ηp2=0.700, suggesting that the effect of ItemXP was greater for non-clinics than clinics. These results again point to the learnability of AIG. ANOVAs results are available in Appendix A.

Participants largely expressed confidence in AIG, according to the survey results. The results support substantial levels of confidence (M=4.05; SD=0.759) in the software used to generate new items and moderate levels of confidence (M=3.89; SD.99) in the use of AIG on their own assessments. However, individuals report having less confidence while creating cognitive models (M=3.47; SD=1.02). Participants' views on the usefulness of AIG were also positive. Participants thought AIG to be a useful tool to measure knowledge (M=4.42; SD=0.69); understand clinical processes (M=4.53; SD=0.61); and enhance item quality (M=4.47; SD=0.70).

Study II

Data collection and sample

Responses to a summative exam in the content area of surgery taken by 132 fifth-year medical students of the EMUM were used as our data. The majority of students were females (73.5%) with ages ranging from 21 to 40, with a mean age of 23 years ($SD=3.40$).

Measure

The exam consisted of 100 dichotomously-scored single best-answer five-option MCQs designed to measure students' knowledge in surgery. The exam covered multiple surgery content domains (strands): *general and vascular surgery; neurosurgery; urology; otorhinolaryngology* and *ophthalmology*. For the sake of analytic purposes, we only retained the questions from the *general and vascular surgery* strand (28 questions evenly distributed: *Manual* = 15; *AIG* = 13). Final scores were calculated by summing up item scores and dividing by the total number of items.

Procedure

Students completed the 2-h exam through a virtual testing platform (*QuizOne*®) under online supervision from their teachers. After finishing the exam, students submitted their responses and the platform closed the respective session.

Analytic approach

Attributes of exam items presented as parameter estimates from psychometric models entail evidence regarding the theory on which item construction is based(Bejar, 2012; Zegota et al., 2022). Comparisons between manually written and AIG-items can be used as a validation procedure to ensure that the two item development methods produce comparable outcomes (Gierl et al., 2021).We used the internal structure (i.e., the psychometric characteristics) and performance indices of the items used in the exam as sources of evidence to assess AIG's validity and item quality (Gierl et al., 2022b; Zegota et al., 2022). Item response theory (IRT) methods were employed to estimate the psychometric properties of the exam items—for a brief overview of IRT, see De Champlain's work (De Champlain, 2010). We used the Rasch model (RM) (Rasch, 1960) to conduct our analysis due to our sample size, its determinants of item response (respondent's ability and item difficulty) and its relevance for achievement tests with dichotomous scoring (Hohensinn & Kubinger, 2011).

Prior to formally running the RM, we assessed whether key assumptions were met, namely: (a) *unidimensionality* (i.e., one dimension explaining the covariance among items); (b) *local independence* (i.e., absence of systematic conditional covariance among items); and (c) *monotonicity between latent trait levels (θ) and true scores* (i.e., the need of the probability of endorsing an item to increase as the latent ability being measured levels increases). Strategies for evaluating RM assumptions and results are provided in appendix B. After ensuring appropriate model-data fit (Cf. Appendix C) we undertook a calibration process to estimate item properties and help in gauging the quality of the items. Finally, item reliability was assessed using item information function (IIF) plots whereas exam reliability was examined using test information function (TIF) plots and the item reliability index. Statistical procedures were based on the full sample. The significance level was set to $\alpha = 0.01$ for all analyses.

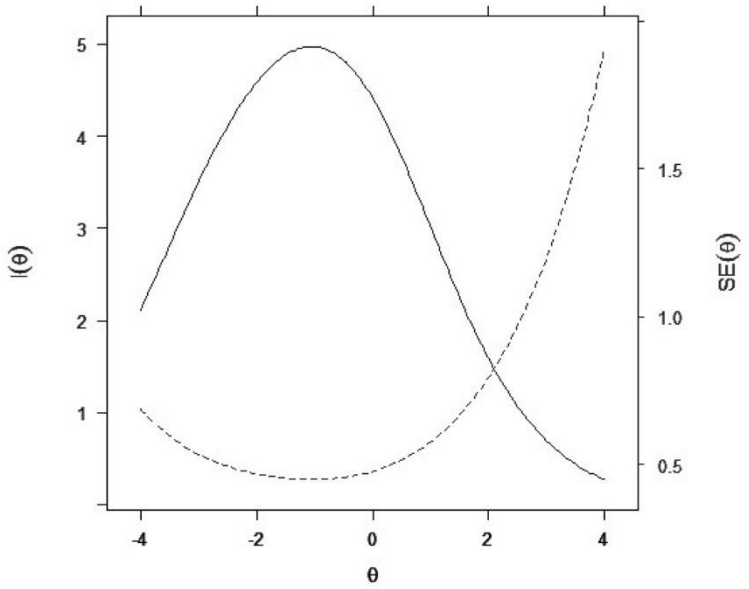


Fig. 3 Test information function with conditional standard error of measurement

Results

Item calibration. Parameter estimates from the fully specified RM are available in Appendix D. Global difficulty estimates of the questions ranged from −2.11 to 2.02 logits (the RM centers on a mean item difficulty value of 0). Item locations did not reveal evenly spaced parameters along the continuum, which may indicate the absence of discrimination between students. This finding is likely due to the fact that we were dealing with medical students, who are known for their high study skills and good academic performances (i.e. restriction of range in candidate ability estimates). The range of difficulty for AIG-items ranged from −2.11 (AIG21) to 1.85 (AIG16) logits, while for manually written items these difficulty parameter estimates ranged from −2.11 (M3) to 2.02 (M10) logits. On average, AIG-items yielded Rasch calibration difficulty levels similar to manually written items, which supports the validity and quality of the AIG process.

Item and test reliability. IIFs revealed that most items conveyed high information values, namely in the middle of the θ distribution (Cf. Appendix E). Since both item types presented similar amounts of information, findings suggest that AIG-items are of high quality and provide precise measurement of candidate ability levels. The TIF was computed together with the conditional standard error of measurement (CSEMs) to assess the level of precision with which the entire exam measured various θ values along the continuum (Cf. Figure 3). The CSEM increased at the high-end of the ability scale, as expected, given the paucity of data located at the extreme ranges of a score distribution. The peak of the TIF (and lowest amount of measurement error or CSEM) was located close to the centre of the ability distribution. Concretely, this suggests that the exam allowed us to measure low-and-middle-ability candidates with the highest level of reliability. The item reliability index

was of 0.96, reflecting replicability of the test if applied to another group of comparable subjects.

Note. The undashed line represents the TIF. The dashed line represents the conditional standard error of measurement function.

Discussion

Despite the fact that most of the work on AIG has been done in theoretical, non-operational contexts, test developers are increasingly employing AIG to augment item banks on practical tests. Naturally, this transition from theory to practice raises unforeseen implementation challenges (Kosh, 2021). Over the course of two studies, we completed a series of comparisons between AIG-items and manually written items to support the implementation of AIG in educational assessment.

Item quality

Designing quality test items can be challenging given the high standards that they must meet for inclusion into a form, especially in high-stakes testing(Yaneva et al., 2020). Our findings suggest that AIG-items meet these standards, and as such can help to address some of the difficulties associated with item development. We showed that the structure of AIG-items is comparable to that of manually developed items. Manually written items, nonetheless, seem to have a larger number of characters. This may be because cognitive models and algorithms, for the sake of computational power, use a less complex dialogue to allow for the generation of a greater number of items. Literature on the topic (Gierl et al., 2016; Gierl & Lai, 2013a) supports our findings. The results of the item quality assessments conducted so far reveal that, in general, AIG-items are considered to be of similar or superior quality to manually written items.

One concern expressed in the literature around AIG pertains to the quality of distractors (Pugh et al., 2016). Developing distractors is arduous since they must be continually adapted to fit with correct responses based on a rationale within the stem. Users of AIG have thus expressed concerns regarding the quality of distractors generated by AIG (Lai et al., 2016). In our first study, only one AIG-item was flagged due to the detection of an implausible distractor. The remaining distractors were found to be of high quality. Since no differences were found concerning other quality indicators, our results reflect those reported in previous studies, which support the high-quality of items and distractors generated by AIG. Psychometric properties for both AIG and manually written items were also comparable. Difficulty estimates support the use of AIG-items as these properties were similar to those of manually written items. These findings are consistent with earlier research (e.g.,) that have addressed the psychometric characteristics of AIG-items and manually written items. Statistical evidence found revealed that, similarly to manually written items, AIG-items are also capable of producing a wide range of difficulty values. These outcomes highlight the benefit of using a strong theory approach: if cognitive features are identified and the items generated

adequate for testing, item features that predict test performance may be specified and controlled (Gierl & Lai, 2012). Consequently, these results encourage the use of of items generated using cognitive modeling processes in operational administrations (Gierl et al., 2016a, 2016b).

Usability

AIG is not intended to replace human expertise (Gierl et al., 2022b). This is the main reason why we proposed to study the usability of this approach, considering not only its efficiency but also its learnability. Regarding efficiency, not surprisingly, the average time to develop a cognitive model exceeded the average time to write an item manually. However, this difference in time was negated by the rapid generation of hundreds of items from just one cognitive model. As expected, our results confirm past literature in that AIG appears to be an effective alternative for generating hundreds of items based on a single cognitive model in a short period of time (Gierl & Lai, 2013b; Lai et al., 2016). With respect to learnability, time spent developing cognitive models and number of items generated did not vary as a function of the participants' item writing experience or clinical knowledge. Our results are reflective of past literature on this topic, supporting that AIG is an easy to learn technique that does not require much training. Based on the framework used in the design of cognitive models, unexperienced users learned to generate MCQ. This is especially appealing to medical schools and other institutions that may lack the professional training and resources typically at hand with professional licensing bodies.

Validity

In contrast to the sources of evidence normally employed to evaluate standard item development methods, different types of evidence are needed to evaluate AIG (Gierl et al., 2022b). In the present paper, we used the statistical properties of generated items as a source of evidence of the validity of AIG. The results obtained were promising, as evidence of validity was found. First, ensuring unidimensionality of the item response data was fundamental to assure construct validity. In the same vein, as internal validity concerns item-level psychometric issues such as the relationship of items to a latent dimension and item difficulty, proving that the items were appropriate for both the theoretical concept and the applied psychometric model ensured validity at the item level (Grimm & Widaman, 2012; Shono et al., 2016).

Our work is in line with (Gierl et al., 2022b), as we focused on item statistics to provide evidence that AIG-items can indeed measure the intended knowledge, skills, and competencies of medical students. The evidence found supports AIG as a valid item development method. We believe that this evidence is supported by the strong theory approach that is used for the development of medical MCQs, since the items are designed based on the knowledge, skills and competencies that are the focus of the tests, providing support for the test item content, test item format and possible inferences about the students' domain of interest.

Implications in the context of education, assessment and test development

School and licensing/certification exams still use a single instance based item writing process that capitalizes on manual, committee based exercises (Albano & Rodriguez, 2018). However, there is a significant shift taking place in the way tests are administered, why they are designed, and who writes them (Gierl et al., 2021). AIG tries to benefit from that expertise in addition to harnessing the power offered by computers for more productive and cost-effective development of test items (Gierl et al., 2021). The use of this methodology has the potential to alter not only the paradigm for how institutions develop test items but also how they manage them. Whereas standard item banks act as an electronic repository for storing, maintaining, and managing information on each item, AIG may employ models rather than items to serve as the unit of analysis in an AIG model bank. Each model, which is written, evaluated, revised, edited, and banked, can be used to produce a large number of questions (Gierl et al., 2022a). Furthermore, the knowledge of content experts is utilised in AIG via the development of cognitive maps whereas the assembly task is done by computers.

The present work contributes two novel sources of evidence to the discussion on AIG. The first point is that while AIG psychometric proprieties have been previously characterized and published in other works, the latter are alsmot uniquevocally based on cognitive maps developed by professional test developers or licensing bodies. In these settings, the quality of items is usually very high even for traditionally developed items. Therefore, AIG may increase efficiency of the process, but the qualitative improvement of items per se is probably limited. The work we completed was in the context of a medical school (MS) that, like many MS, has limited resources. The quality of traditionally developed MCQs is therefore, unsurprisingly, typically lower. We believe that the templates that are used for the development of cognitive models help low-resource institutions to improve not only efficiency but more importantly, they provide a qualitative increment that is much more relevant to this context. A second contribution lies in the fact that we involved participants with very different degrees of expertise in the cognitive model development process. The most "naïve" were neither academicians nor medical students. Rather, some were administrative staff, biomedical researchers, and staff from the communication department. Our belief is that the process underlying the development of cognitive models is transparent and comprehensive, thus allowing participants with little expertise in item writing and limited understanding of clinical sciences to generate high-quality items with ease. In other words, the cognitive model templates impose a format and style of writing that promotes a quick "learning curve" and results in high quality writing. From our review, this issue has very rarely, if ever, been explored in the AIG literature.

Also, not to be neglected, the savings associated with AIG can be significant in comparison to manual item writing. On average, when considering the costs of editing, field-testing and calibrations procedures, the development of a single MCQ for high-stakes testing can range from US$1500 to US$2000 (Kosh et al., 2019; Lai et al., 2009; Rudner, 2010). Given this cost, it is easy to see why item writing is a costly process for test development in the twenty-first century (Lai et al., 2009). Since AIG is an easy technique to learn, institutions will not need to constantly call on professionals to develop test items. Consequently, institutions will save a significant amount of money in development costs, while promoting better assessment.

Strengths and limitations

The main contribution of our paper lies in the extensive framework that guided the comparisons between AIG and manually written items. These comparisons allowed us to examine whether AIG-items could be applied in the same way as manually written items. The multimodal approach that we adopted is another strong asset, as individual assessment of the items was made based on evaluators' judgments as well as item parameters. Gathering participants' perspective, as we did in our work, is also worth highlighting, as this provides evidence to support a confident application of AIG. In sum, the several sources of evidence collected to support the validity of the AIG process is the main contribution of our work and helps to fill a gap identified in the literature.

However, this work has limitations that should be considered when generalizing our findings to the broader AIG context. First, the samples used in both studies were limited and may lack inferential power. Future research should therefore focus on collecting larger samples of participants with different levels of ClinicXP and ItemXP to support a better generalization of our results. Another limitation of our work rests with the use of only one expert to evaluate the items developed in study I. Using a panel of experts with specific guidelines would provide a more complete and diversified assessment of the items. Similarly, it would be advisable to calibrate our exam items by administering it to a larger group of students. A larger sample size would also allow the application of more robust IRT models capable of estimating additional item parameters that the RM is unable to compute, such as item discrimination and a lower asymptote parameter ("pseudo-guessing"). Finally, we need to underscore that AIG was only evaluated in the content area of surgery. This focus on a single domain is in and of itself a limitation, and applications to other content areas are necessary.

Directions for future research

AIG still remains a relatively unknown item writing process in practice. For the reasons highlighted in this paper (cost, item quality, etc.), there is a need to explore more innovative and effective means of generating test items (Kurdi et al., 2020). Future studies should focus on assessing the psychometric properties of AIG-items and in particular, on expanding the cognitive model structure to include more complex item types beyond MCQs (Gierl & Lai, 2018). Studies focusing on evaluating the process of learning AIG, the generation of distractors, and how AIG can provide feedback for students while preventing test security breaches would also significantly contribute to the literature. The validation of AIG's processes is another point that can be strengthened with future studies. The item definition, the item development process, and the item quality review also represent useful sources of validity evidence that can be deepened with studies devoted to this topic (Gierl et al., 2022b). The application of AIG in different forms of testing (e.g., computerized adaptive testing) with items with different degrees of complexity (and difficulty) that could be generated on-the-fly would also be worth investigating(Gierl et al., 2021). Finally, it is worth mentioning that cognitive models are still not widely used in large-scale assessments nowadays (Ferrara &

DeMauro, 2006; Leighton & Gierl, 2011). As of right now, there are no cognitive models that have a direct relationship to large-scale educational assessments. The dearth of cognitive models might restrain the spread of AIG and support the ongoing use of antiquated methods of development that are becoming more and more out of step with the demands of contemporary evaluation. This entails the need to adapt cognitive models for large-scale assessment and methods to describe them diagrammatically, as well as systematic records of cognitive models of science learning and research in various academic fields (Leighton & Gierl, 2011). In order to encourage the usage of cognitive models and the subsequent application of AIG, future directions should concentrate on the development of these models, particularly in terms of validity arguments.

Conclusion

Testing in schools is changing at a never-before-seen pace. AIG is a next-generation assessment process capable of producing numerous high-quality items in a fast and cost-effective manner. Our findings suggest AIG-items are indistinguishable from items manually written by experts, both in terms of structure and psychometric properties. AIG can also allow non-clinical and non-professional item writers to develop items that are comparable to those written by experts. Our findings provide evidence of validity in a number of areas, including construct and internal validity. By using AIG, educational institutions can also reap substantial cost savings without compromising quality. The broader implementation of AIG at medical schools can potentially improve the work of teachers and content experts as well as promote better assessment.

Appendix A: Learnability analysis

See Table 1.

Table 1 ANOVA results using time spent developing the cognitive model (top quadrant) and quantity of items generated (lower quadrant) as the criterion

Source	Sum of Squares	*df*	Mean Square	*F*	*p*	Partial ηp2
(Intercept)	54,159	1	54,159	171	< .001	0.961
ItemXP	1637	2	818	2.59	0.144	0.425
ClinicXP	583	1	583	1.85	0.216	0.209
ItemXP * ClinicXP	0.626	1	0.626	0.002	0.966	< .001
Error	2211	7	315			
Intercept	1,442,226	1	1,442,226	56.1	< .001	.889
ItemXP	634,873	2	317,436	12.3	.005	.779
ClinicXP	354,905	1	354,905	13.8	.008	.663
ItemXP * ClinicXP	419,216	1	419,216	16.3	.005	.700
Error	180,058	7	25,722			

Appendix B: IRT assumptions—strategies and results

Unidimensionality. Unidimensionality was evaluated based on two approaches: first, since item response data are rarely strictly unidimensional, it was necessary to consider the existence of additional "minor" abilities along with the main construct of interest (i.e., *knowledge in general and vascular surgery*) (Bonifay et al., 2015). This corresponds to the concept of *essential unidimensionality*, which assumes that a scale is measuring a predominant construct while also tapping into additional factors that are less relevant (Raykov & Pohl, 2013). This hypothesis was assessed using the DETECT index in our study (Stout et al., 1996). The DETECT index is the result of the DETECT, ASSI and RATIO indices, which were calculated for the options unweighted (covariance of item pairs was equally weighted) and weighted (covariance of item pairs was weighted by the sample size of item pairs)(Zhang, 2007). The indices obtained for the weighted and unweighted options were as follows: DETECT = −0.489; ASSI = −0.365; RATIO = −0.498. These values suggest no departure from the assumption of essential unidimensionality (Robitzsch & Robitzsch, 2021).

Second, since we're analyzing items developed by two different processes, we used two partition criteria of the sample to evaluate if the item parameters were invariant across subgroups splits based on *score* and *gender*. This was evaluated following a consecutive unidimensional approach, more specifically, using the Andersen's LRT test. For unidimensionality to be met, item parameters based on subgroup fits should be approximately the same. The Andersen LRT test gave a non-significant result for both partition criteria, *score*: $\chi^2 = 39.1$; df = 26; $p = 0.047$/*gender*: $\chi^2 = 24.6$; df = 26; $p = 0.539$. No violations of essential unidimensionality were found, providing evidence to support the expect construct in our work. However, one item (AIG 28) was excluded from further analysis due to inappropriate response patterns between person subgroups. The remaining items revealed good fit, and as such provide evidence of internal validity.

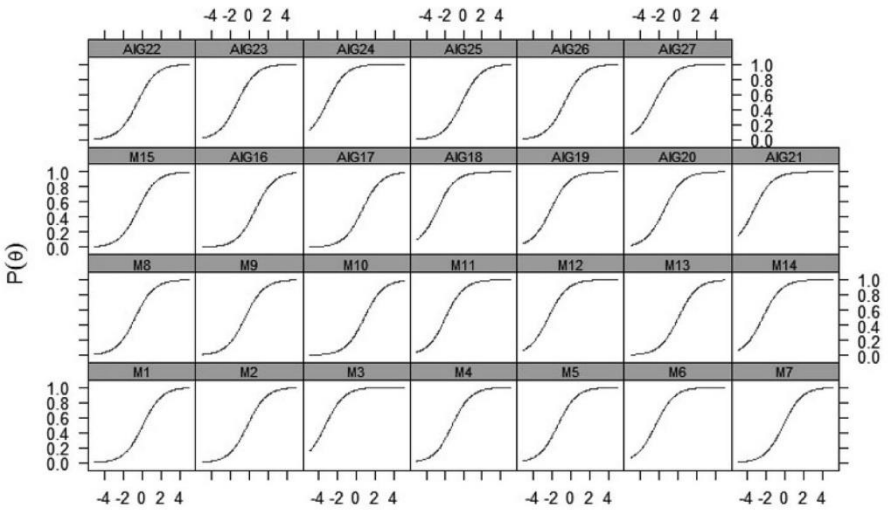


Fig. 4 ICCs for the twenty-seven items composing the general and vascular surgery strand

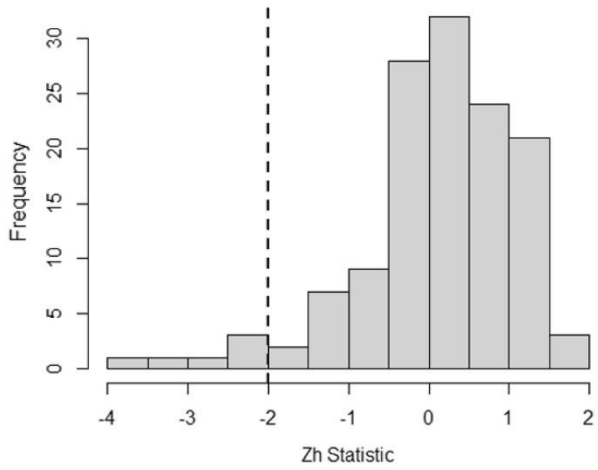


Fig. 5 Person fit histogram. *Note*. Large negative values (e.g., Zh < −2) indicate person misfit. Large positive values (e.g., Zh > 2) claim that the likelihood of the pattern of the responses is higher than the predictive likelihood of the RM, disturbing latent estimation

Local independence. Local independence at the global level was tested through the T11-test. Results suggested that strict local independence did not hold at a global level ($p < 0.01$), meaning that item responses may be correlated or dependent after controlling for one factor. However, these results should not call into question the legitimacy of our analysis. One possible reason for this result may be that the items in the study belong to the same strand in an exam with multiple strands. Furthermore, in most cases, local independence proves to be a difficult assumption to ensure and, for this reason, is often omitted in practice (Mair, 2018). Thus, this assumption is usually subsumed under the unidimensionality assumption(Edelen & Reeve, 2007).

Monotonicity. Monotonicity was examined using plots of item characteristic curves (ICCs)—plots of the probability for a test item to be answered correctly, considering the student's ability and the latent trait being measured (Cf. Figure 4). The probability of responding to an item correctly in this strand was a monotonically increasing function of θ for most items in our exam, corroborating this assumption.

Appendix C: Model-data fit

We evaluated if the item parameters were invariant across subgroup splits based on score. The Andersen LRT test was non-statistically significant for the score partition criterion ($\chi^2 = 39.1$; df = 26; p = 0.047), ensuring model fit. Infit and outfit mean square values were close to 1, ensuring item fit. These results represent evidence of internal validity for AIG, as none of its items presented poor fit. The Zh statistic was larger than −2 for most of the respondents, ensuring person fit (Cf. Figure 5).

Appendix D: RM parameter estimates

See Table 2.

Table 2 Rasch Model item parameter estimates and fit statistics for the twenty-seven exam questions

Item	b	SE	$\chi 2$	Df	p-value	Outfit MSQ	Infit MSQ
M1	1.29	0.18	129	131	0.52	0.98	0.98
M2	1.01	0.18	152	131	0.10	1.16	1.14
M3	−2.11	0.44	258	131	0.00	1.96	0.94
M4	0.02	0.21	125	131	0.61	0.95	0.94
M5	−0.16	0.21	109	131	0.91	0.83	0.89
M6	−1.17	0.30	114	131	0.85	0.87	0.94
M7	1.01	0.18	126	131	0.58	0.96	0.96
M8	0.48	0.19	161	131	0.04	1.22	1.16
M9	0.75	0.18	130	131	0.50	0.99	1.01
M10	2.02	0.19	116	131	0.82	0.88	0.90
M11	−0.71	0.25	112	131	0.87	0.85	0.92
M12	−1.17	0.30	110	131	0.90	0.84	0.92
M13	1.26	0.18	139	131	0.29	1.06	1.05
M14	−1.17	0.30	113	131	0.86	0.86	0.88
M15	0.75	0.18	143	131	0.22	1.09	1.06
AIG16	1.85	0.19	123	131	0.66	0.94	0.96
AIG17	1.81	0.19	134	131	0.40	1.02	1.01
AIG18	−1.61	0.36	133	131	0.42	1.01	0.96
AIG19	−0.92	0.27	91.6	131	1.00	0.69	0.86
AIG20	−0.26	0.22	150	131	0.12	1.14	1.08
AIG21	−2.11	0.44	130	131	0.51	0.99	0.96
AIG22	0.72	0.18	132	131	0.44	1.01	0.97
AIG23	−0.07	0.21	147	131	0.16	1.11	1.03
AIG24	−1.92	0.41	86.4	131	1.00	0.65	0.90
AIG25	1.07	0.18	112	131	0.88	0.85	0.87
AIG26	0.58	0.19	129	131	0.53	0.98	0.99
AIG27	−1.27	0.31	109	131	0.91	0.83	0.93

b: item difficulty; *SE*: standard error; $\chi 2$: Chi-Square item fit statistic; *Df*: degrees of freedom; MSQ: Mean square

Appendix E: Item information curves

See Fig. 6.

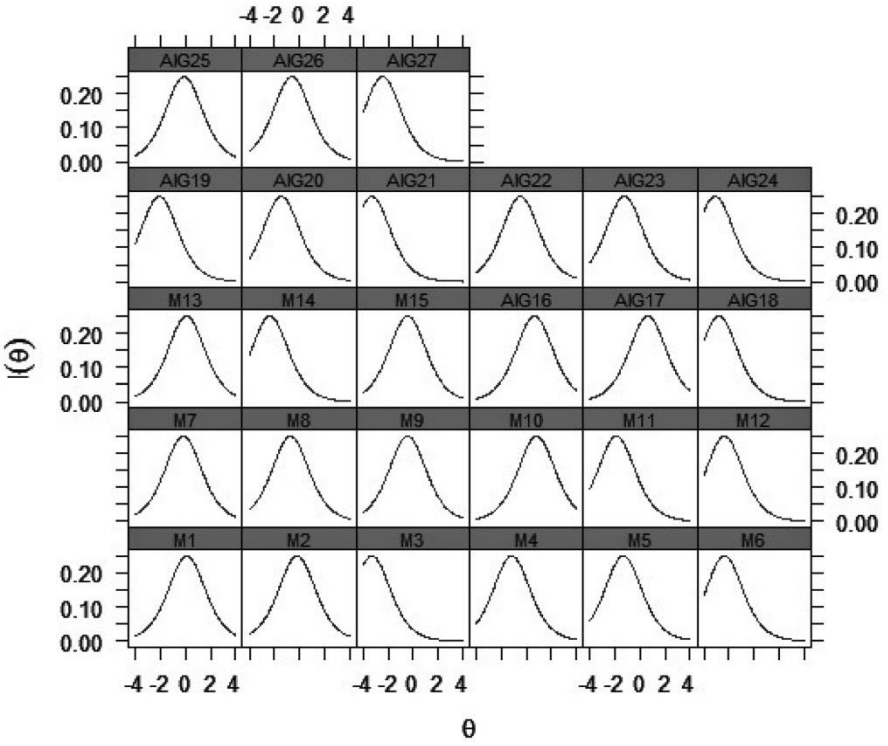


Fig. 6 Item information curves plot

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.

Author contributions FF conceived the presented idea and developed the theory.All authors gathered the data.FF and PC performed the statistical analysis.AC, PC and JMP verified the analytical methods and supervised the findings of this work.All authors discussed the results and contributed to the final manuscript.

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. The authors thank all study participants.

Data availability statement The data that support the findings of this study are available from the corresponding author upon request.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval Both studies in this paper were approved by the Ethics Committee for Research in Life and Health Sciences of the University of Minho. Written informed consent was obtained. Exam data were 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

Albano, A. D., & Rodriguez, M. C. (2018). Item development research and practice. In *Handbook of accessible instruction and testing practices* (pp. 181–198). Springer.

American Educational Research Association. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

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

Bejar, I. I. (2012). Item generation: Implications for a validity argument. In Mark J Gierl T. M. H. (Ed.), *Automatic item generation: Theory and practice.* (pp. 50–66). Routledge. https://doi.org/10.1163/ej.9789004172067.i-752.40

Bejar, I. I., Lawless, R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3), 1–30. https://doi.org/10.1002/j.2333-8504.2002.tb01890.x

Billings, M. S., Deruchie, K., Hussie, K., Services, M., Kulesher, A., Editor, M., Merrell, J., Swygert, K. A., Tyson, J., Case, S. M., Haist, S., & Swanson, D. B. (2020). *Constructing written test questions for the health sciences contributors from NBME* (6th ed., Issue November). National Board of Medical Examiners.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling, 22*(4), 504–516. https://doi.org/10.1080/10705511.2014.938596

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*. https://doi.org/10.1863/jss.v048.i06

Choi, J. (2020). Automatic item generation with machine learning techniques. In Jiao, H & Lissitz, R. (Eds.), *Application of artificial intelligence to assessment* (Vol. 1, pp. 189–210). Information Age Publishing.

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. https://doi.org/10.1111/j.1365-2923.2009.03425.x

de Chiusole, D., Stefanutti, L., Anselmi, P., & Robusto, E. (2018). Testing the actual equivalence of automatically generated items. *Behavior Research Methods, 50*(1), 39–56. https://doi.org/10.3758/s13428-017-1004-5

Douthit, N. T., Norcini, J., Mazuz, K., Alkan, M., Feuerstein, M. T., Clarfield, A. M., Dwolatzky, T., Solomonov, E., Waksman, I., & Biswas, S. (2021). Assessment of global health education: The role of multiple-choice questions. *Frontiers in Public Health*. https://doi.org/10.3389/fpubh.2021.640204

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(SUPPL. 1), 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Kingston, N. M. (2018). Automatic Item Generation: A more efficient process for developing mathematics achievement items? *Journal of Educational Measurement, 55*(1), 112–131. https://doi.org/10.1111/jedm.12166

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.

Ferrara, S., & DeMauro, G. E. (2006). Standardized assessment of individual achievement in K-12. *Educational Measurement*, *January 2006*, 579–622.

Gierl, M., & Haladyna, T. M. (2012). Automatic Item Generation: An introduction. In T. M. H. Mark J Gierl (Ed.), *Automatic item generation: Theory and practice.* (pp. 3–12). Routledge.

Gierl, M., Latifi, S. F., Lai, H., Matovinovic, D., & Boughton, K. A. (2016b). Using automated procedures to generate test items that measure junior high science achievement. In *Handbook of research on technology tools for real-world skill development* (pp. 590–610). IGI Global.

Gierl, M., & Lai, H. (2016). The Role of Cognitive Models in Automatic Item Generation. In Rupp, A. & Leighton, J. (Eds.), *The Handbook of Cognition and Assessment* (pp. 124–145). Wiley-Blackwell. https://doi.org/10.1002/9781118956588.ch6

Gierl, M., Shin, J., Firoozi, T., & Lai, H. (2022a). Using content coding and automatic item generation to improve test security. *Frontiers in Education*. https://doi.org/10.3389/feduc.2022.853578/full

Gierl, M., Swygert, K., Matovinovic, D., Kulesher, A., & Lai, H. (2022b). Three sources of validation evidence needed to evaluate the quality of generated test items for medical licensure. *Teaching and Learning in Medicine*, 1–11.

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., & 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., & 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., & 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., Pugh, D., Touchie, C., Boulais, A.-P., & de Champlain, A. (2016a). Evaluating the psychometric characteristics of generated multiple-choice test items. *Applied Measurement in Education, 29*(3), 196–210.

Gierl, M., Lai, H., & Tanygin, V. (2021). Advanced methods in automatic item generation. In *Advanced methods in automatic item generation*. https://doi.org/10.4324/9781003025634

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

Grainger, R., Dai, W., Osborne, E., & Kenwright, D. (2018). Medical students create multiple-choice questions for learning in pathology education: A pilot study. *BMC Medical Education, 18*(1), 1–8.

Grimm, K. J., & Widaman, K. F. (2012). Construct validity. In *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.* (pp. 621–642). American Psychological Association. https://doi.org/10.1037/13619-033

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

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

Jeng, J. (2005). Usability assessment of academic digital libraries: Effectiveness, efficiency, satisfaction, and learnability. *Libri, 55*(2–3), 96–121. https://doi.org/10.1515/LIBR.2005.96

Johnes, J., Portela, M., & Thanassoulis, E. (2017). Efficiency in education. *Journal of the Operational Research Society, 68*(4), 331–338. https://doi.org/10.1057/s41274-016-0109-z

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

Kosh, A. E. (2021). Distractor suites: A method for developing answer choices in automatically generated multiple-choice items. In *Journal of Applied Testing Technology* (Vol. 22, Issue 1). www.jattjournal.com

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

Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. (2020). A systematic review of automatic question generation for educational purposes. *International Journal of Artificial Intelligence in Education, 30*(1), 121–204. https://doi.org/10.1007/s40593-019-00186-y

Lai, H., Alves, C., & Gierl, M. J. (2009). Using Automatic Item Generation to address item demands for CAT. *GMAC Conference on Computerized Adaptive Testing*, 1–16.

Lai, H., Gierl, M., Touchie, C., Pugh, D., Boulais, A. P., & de Champlain, A. (2016). Using Automatic Item Generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine, 28*(2), 166–173. https://doi.org/10.1080/10401334.2016.1146608

Leighton, J. P., & Gierl, M. J. (2011). *The learning sciences in educational assessment: The role of cognitive models*. Cambridge University Press.

Lewis, J. (2016). Usability testing. In Salvendy, G. (Ed.), *Handbook of human factors and ergonomics.* (3rd ed., Vol. 12, Issue 1, pp. 1267–1305). John Wiley and Sons Inc. https://doi.org/10.1002/0470048204.ch49

Mair, P. (2018). Modern psychometrics with R. In S. I. Publishing (Ed.), *technometrics* (Vol. 62, Issue 1). https://doi.org/10.1080/00401706.2019.1708675

Mindyarto, B. N., Nugroho, S. E., & Linuwih, S. (2018). Applying automatic item generation to create cohesive physics testlets. *Journal of Physics: Conference Series*. https://doi.org/10.1088/1742-6596/983/1/012054

Patel, S. (2021). *Exploring the effect of occlusion on a computerized mental-rotation test: Implications for automatic item generation*. Louisiana Tech University.

Patrick, A., Hatzinger, R., Maier, M. J., Rusch, T., & Mair, M. P. (2018). *Package ' eRm .'*

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*. https://doi.org/10.1186/s41039-020-00134-8

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

Rafatbakhsh, E., Ahmadi, A., Moloodi, A., & Mehrpour, S. (2020). Development and validation of an Automatic Item Generation system for english idioms. *Educational Measurement: Issues and Practice, 40*(2), 1–11. https://doi.org/10.1111/emip.12401

Rasch, G. (1960). Probabilistic models for some intelligence and attainment test. In *Studies in mathematical psychology* (Vol. 1). Danmarks Paedagogiske Institut.

Raykov, T., & Pohl, S. (2013). Essential unidimensionality examination for multicomponent scales: An interrelationship decomposition approach. *Educational and Psychological Measurement, 73*(4), 581–600. https://doi.org/10.1177/0013164412470451

Revelle, W., & Revelle, M. W. (2015). Package ' psych .' In *The comprehensive R archive network* (pp. 337–338.).

Robitzsch, A., & Robitzsch, M. A. (2021). *Package 'sirt.'*

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

Rudner, L. (2010). Elements of adaptive testing. *Elements of Adaptive Testing, May.* https://doi.org/10.1007/978-0-387-85461-8

Rust, J., Kosinski, M., & Stillwell, D. (2020). Modern Psychometrics. In: *Modern psychometrics*. https://doi.org/10.4324/9781315637686

Schmeiser, C., & Welch, C. (2006). Test Development. In *Educational measurement* (4th ed., pp. 307–353). National Council on Measurement in Education and American Council on Education.

Schmidt, H. G., & Mamede, S. (2015). How to improve the teaching of clinical reasoning: A narrative review and a proposal. *Medical Education, 49*(10), 961–973.

Shappell, E., Podolej, G., Ahn, J., Tekian, A., & Park, Y. S. (2021). Notes from the field: Automatic item generation, standard setting, and learner performance in mastery multiple-choice tests. *Evaluation & the Health Professions, 44*(3), 315–318.

Shin, E. (2021). *Automated Item Generation by combining the non-template and template-based approaches to generate reading inference test items* [University of Alberta]. https://doi.org/10.7939/r3-75wr-hc80

Shono, Y., Ames, S. L., & Stacy, A. W. (2016). Evaluation of internal validity using modern test theory: Application to word association. *Psychological Assessment, 28*(2), 194.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics, 28*(4), 295–313. https://doi.org/10.3102/10769986028004295

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20*(4), 331–354.

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

Wickham, H., Chang, W., & Wickham, M. H. (2016). *Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics.* (2.1; Vol. 2, Issue 1, pp. 1–189).

Yaneva, V., Ha, L. A., Baldwin, P., & Mee, J. (2020). Predicting item survival for multiple choice questions in a high-stakes medical exam. *LREC 2020 - 12th international conference on language resources and evaluation, conference proceedings*, *May*, 6812–6818.

Yang, Y., Sanyal, D., Michelson, J., Ainooson, J., & Kunda, M. (2022). Automatic Item Generation of figural analogy problems: A review and outlook. *ArXiv Preprint* https://arxiv.org/abs/2201.08450.

Zegota, S., Becker, T., Hagmayer, Y., & Raupach, T. (2022). Using item response theory to appraise key feature examinations for clinical reasoning. *Medical Teacher, 44*, 1–7.

Zhang, J. (2007). Conditional covariance theory and detect for polytomous items. *Psychometrika, 72*(1), 69–91.

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

CHAPTER V

ITEM RESPONSE THEORY IN PSYCHOMETRIC RESEARCH: HOW TO ASSESS THE RELATIONSHIP BETWEEN ITEM PROPERTIES, RESPONDENTS, AND LATENT TRAITS

This chapter is based on the following publication: **Falcão, F**., Pereira, D., Pêgo, J. M., Costa, P. (2023). Item response theory in psychometric research: how to assess the relationship between item properties, respondents, and latent traits. In *Sage Research Methods Cases Part 1*. SAGE Publications, Ltd. DOI: 10.4135/9781529627459

SAGE Research Methods Cases Part 1

Item Response Theory in Psychometric Research: How to Assess the Relationship Between Item Properties, Respondents, and Latent Traits

Author: Filipe Falcão, Daniela Pereira, José M. Pêgo, Patrício Costa

Pub. Date: 2023

Product: SAGE Research Methods Cases Part 1

DOI: https://dx.doi.org/10.4135/9781529627459

Methods: Item response theory, Confirmatory factor analysis, Latent trait analysis

Keywords: traits, property, parameters, instruments, attitudes

Disciplines: Psychology

Access Date: February 6, 2023

Publishing Company: SAGE Publications Ltd

City: London

Online ISBN: 9781529627459

© 2023 SAGE Publications Ltd All Rights Reserved.

© 2023 SAGE Publications Ltd

Abstract

In this methodological case, we draw on the experience from two previous studies to discuss the process of applying an Item Response Theory (IRT) analysis. In both studies, the psychometric properties of two measurement instruments were characterized based on IRT. In this paper, we consider common issues in this type of analysis, including model selection, assumptions testing, model diagnostics, item calibration, reliability, and score estimation. The practicalities of applying IRT are examined, and solutions to overcome common obstacles are offered.

Learning Outcomes

By the end of this case study, students should be able to:

- Understand the processes underlying an IRT analysis.
- Reflect on the methodological decisions inherent to an IRT analysis.
- Understand and apply strategies for testing the assumptions underlying IRT.
- Understand and interpret, both analytically and graphically, the parameters used to calibrate the items of an instrument and the common methods of score estimation used.
- Make autonomous and conscious decisions regarding model diagnosis, ability estimation, item calibration, evaluation, and test refinement.

Project Overview and Context

Academics face several challenges when analyzing the psychometric features of measurement instruments, especially those based on Item Response Theory (IRT). IRT comprises a collection of modeling techniques for analyzing item-level data to measure interindividual variation (Edelen & Reeve, 2007). It is a widely used approach for typical procedures in test development (e.g., test scaling, item calibration, construction of item banks, etc.) not only by scholars but also by testing companies in the United States and Europe (Paek & Cole, 2020). This research methods case is intended to raise scholars' awareness of the potential roadblocks in this type of analysis and help them complete their tasks. The knowledge produced here is based on the experience gathered from two studies conducted by the authors that are being published. In both studies, the

© 2023 SAGE Publications Ltd

data under inquiry was subjected to a psychometric evaluation under different IRT models: (1) the first study focused on the examination of a dichotomously scored, high education, medical summative exam in the content area of surgery, while the (2) second focused on the analysis of a polytomously scored, psychological instrument focused on the assessment of populist attitudes. IRT approaches were employed to evaluate the instrument's properties and estimate respondents' latent abilities/attitudes (*Study 1:* knowledge in surgery; *Study 2:* populist attitudes). Both investigations relied on different datasets originally compiled by the authors.

Various IRT publications are accessible to readers who want to learn more about IRT (e.g., De Ayala, 2013). This paper provides a roadmap for an IRT analysis—from validation of the assumptions to ability score estimations. However, one should consider that this is not a case study that delves into IRT's deep mathematical/theoretical/technological components. Instead, with this methods case, we offer an overview of IRT and propose a set of steps that we consider essential in applying such an analysis, providing a learning toolkit for those interested in it.

Section Summary

- IRT is a group of modeling tools for measuring interindividual variance in item-level data analysis.
- The information produced here is based on the experience gathered from two studies conducted by the authors, focusing on the assessment of the properties of two measurement instruments and the estimation of respondents' latent abilities based on item response data.
- This case study offers a step-by-step guide for doing an IRT analysis, from confirming the presumptions through estimating ability scores.

Research Design

Overview

Our research design relied on the psychometric analysis of different datasets under different IRT models. In the first study, we obtained the responses of 132 medical students to a surgery exam with different types

© 2023 SAGE Publications Ltd

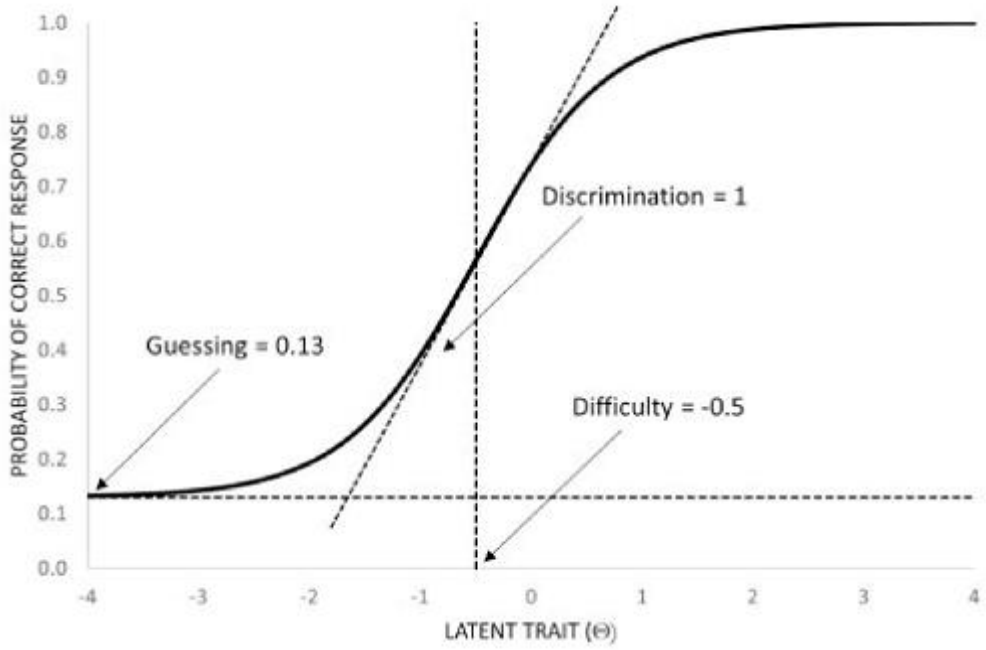
of items (items automatically created by algorithms vs. manually designed items). In the second study, we obtained data from 1343 Portuguese voters to a battery of populist attitudes. The first step in both analyses was to select an appropriate IRT model for each instrument—considering the nature of the data and the sample size. Next, we determined whether the models met the theory's assumptions and assessed several fit measures relating to the models, items, and respondents. Finally, we assessed the properties of the items, conducted reliability tests, and estimated the respondent's scores. This will be explained in detail in the next section.

Why Use IRT to Conduct Validation of Assessment Tests?

In the literature, we find two main measurement frameworks: IRT and Classical Test Theory (CTT). The aim of both theories is to establish a respondent along a latent trait continuum. CTT has been used most of the time by the measurement community. IRT, however, has gained popularity recently and is gradually replacing CTT. It is a more recent advancement in psychometrics and offers a more complex (although more reliable) method of assessing test validity and reliability. The central feature of this paradigm is the specification of a mathematical function known as the Item Characteristic Curve (ICC), which relates the probability of an individual's correct response on a test item to an underlying ability (e.g., Edelen & Reeve, 2007; Hambleton et al., 1991; Paek & Cole, 2020). This "ability," or latent trait, or simply "θ," represents an unobservable psychological attribute of a person (e.g., *knowledge in statistics*). However, since these qualities cannot be measured directly, a θ score is generated from a subject's responses to a series of test items. Figure 1 represents the graphic display of a typical ICC.

© 2023 SAGE Publications Ltd

Figure 1. Example of an item characteristic curve in IRT.



Source: From *Applying item response theory models to entrance examination for graduate studies: Practical issues and insights*, by O. Bulut, 2015, in *Journal of Measurement and Evaluation in Education and Psychology*,6(2), p.315.

IRT uses θ and item parameters to estimate the likelihood of responses**.** There are five unidimensional IRT models for test items with dichotomous scores: the Rasch Model (RM) (Rasch, 1960), the one-parameter model, two-parameter, three-parameter, and four-parameter IRT models**.** For items with polytomous scores, we use generalized versions of the above-mentioned models, such as the Partial Credit Model (Masters, 1982), the Generalized Partial Credit Model (Muraki, 1992), the Rating Scale Model (Andrich, 1978), the Graded Response Model (Samejima, 1969), the Nominal Response Model (Darrell Bock, 1972), or the Nested Logit Model (Suh & Bolt, 2010)—for an overview of these models, see Desjardins and Bulut (2018) work**.** Each item is characterized within these models, and the test characteristics are derived from its items**.** These

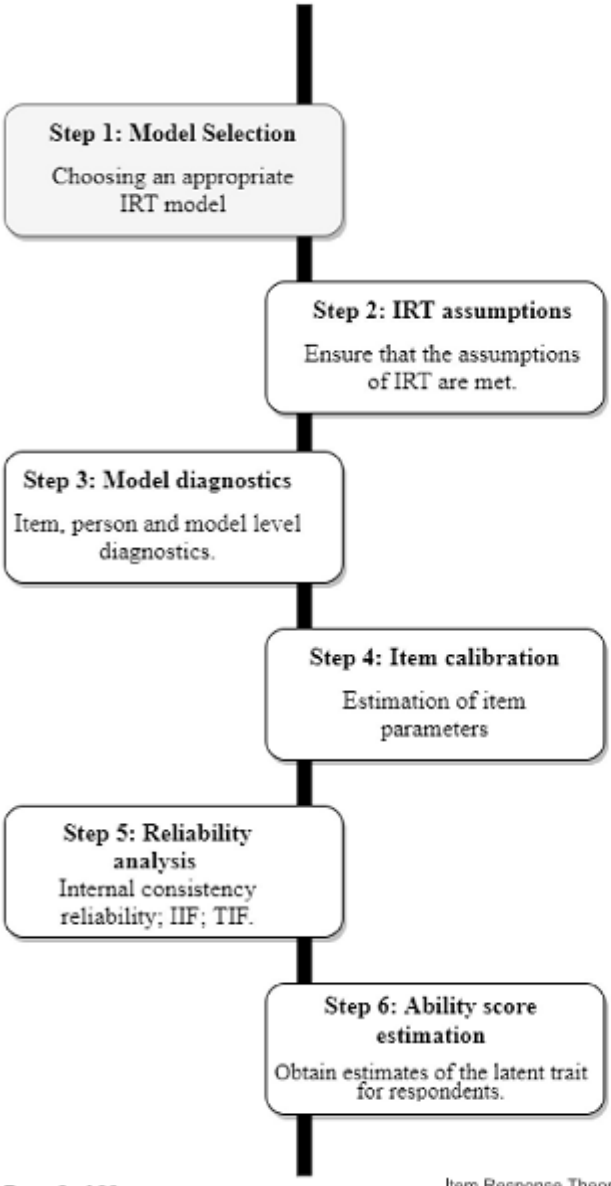
© 2023 SAGE Publications Ltd

characteristics may include the item (1) *difficulty* (i.e., the minimum latent trait level required to correctly answer an item—*b*), (2) *discrimination* (i.e., the ability of an item to distinguish between individuals with low and high θ levels—*a*), and (3) *guessing* (i.e., the probability of correctly answering an item for examinees with very low θ levels—*c*).

If the model is appropriate for the data and the estimates of parameters are accurate, then the instrument will have alluring properties. Based on these parameters, one draws an ICC representing the respondent's probability of correctly answering the item conditional on his θ level. Two final important concepts yet to be mentioned are the item information function (IIF) and the test information function (TIF). These indicate the amount of information for an item/test along the θ continuum, presenting reliability evidence (Desjardins & Bulut, 2018). Specifically, they intend to find the degree to which an item/test reduces the uncertainty in the estimation of a person's θ. Below you can find detailed information on the steps taken in our studies to conduct an IRT analysis. This sequence is represented graphically in Figure 2.

© 2023 SAGE Publications Ltd

Figure 2. Roadmap for an IRT analysis.


Step 1: Model Selection
Choosing an appropriate IRT model
Step 2: IRT assumptions
Ensure that the assumptions of IRT are met.
Step 3: Model diagnostics
Item, person and model level diagnostics.
Step 4: Item calibration
Estimation of item parameters
Step 5: Reliability analysis
Internal consistency reliability; IIF; TIF.
Step 6: Ability score estimation
Obtain estimates of the latent trait for respondents.

© 2023 SAGE Publications Ltd

Step 1. Model Selection

The first step in both studies focused on choosing an appropriate IRT model. As we already observed, several IRT models are available, and these should be selected according to the nature of the data and sample size. In the first study, since we were dealing with dichotomous data and a relatively small sample size (N = 132), we selected the most basic IRT model to conduct our analysis: the RM. This model is relevant for achievement tests with dichotomous scoring due to its determinants of item response (Hohensinn & Kubinger, 2011). Within the RM, the discrimination parameter is set to 1 for all items. As a result of this adjustment, the likelihood of properly answering an item is solely determined by the item difficulty and θ level. For the second study, since we were dealing with polytomous data within a sample of considerable size (N = 1343), we selected a unidimensional polytomous IRT model for ordinal items: the Graded Response Model (GRM) (Samejima, 1969). The GRM is an IRT model that can be used to analyze item responses in several categories. This model details how each item adds to the ability of the measure to distinguish between individuals with varying levels of θ (Adams & Farmer, 2022). Since each item is comprised of k response categories, their respective parameters are estimated for $k - 1$ boundary response functions characterized by two parameters: discrimination and difficulty. These functions will represent the probability of selecting any response option greater than the option of interest (Depaoli et al., 2018).

Step 2. IRT Assumptions

Before applying the models to the data, we ensured that the three main assumptions—of IRT—(1) unidimensionality of the latent trait; (2) local independence; (3) monotonicity between θ levels and true scores—were met. The first requires finding a dominant factor underlying the set of items in the study. Local independence implies the absence of systematic conditional covariance among items, meaning that, given a person parameter, item responses become independent (i.e., no relationship between the items). Finally, monotonicity refers to the need to increase the probability of endorsing an item as the θ level increases (Mair, 2020).

Step 3. Model, Person, and Item Diagnostics

In addition to confirming that assumptions were met, we performed model diagnostics to detect inconsistencies between the IRT models and our data. The diagnostics were conducted at three levels: item, person, and model (Desjardins & Bulut, 2018). Through this procedure, we established a baseline for the chosen model.

Step 4. Item Calibration

By estimating the item parameters, we obtained information on the technical properties of the test items. This allowed us to have a better knowledge of the usefulness of each item in both instruments, as well as a preliminary concept of which items to keep and which to discard.

Step 5. Reliability Analysis

The consistency of results is referred to as reliability. Reliability is critical for making meaningful inferences based on the outcomes of the instrument (Desjardins & Bulut, 2018). Through the IIFs and the TIF, we evaluated both instruments regarding the amount of information and precision across the full range of the trait in the study, which provided reliability. Discrimination values are relevant in terms of the reliability of an instrument because they determine the amount of information an item provides, with more discriminating items providing more information regarding θ (Embretson & Reise, 2013). We also conducted internal consistency analyses to obtain further information about the reliability of the instruments under study.

Step 6. Ability Score Estimation

Following the estimation of item parameters using a specific IRT model, we obtained estimates of θ for each respondent that took the instruments.

Section Summary

- The specification of ICCs, which links the likelihood that a test taker will correctly respond to a question to an underlying ability, is the key component of IRT.
- Our study strategy was based on the psychometric evaluation of various datasets using different IRT models (RM and GRM).
- The following actions are part of an IRT analysis: model selection; assumptions testing; model, person, and item diagnostics; item calibration; reliability analysis; and ability score estimation.

Research Practicalities

Where Should We Conduct Such an Analysis?

One of the main limitations of this type of analysis is the software available. Most of the software employing IRT is proprietary or presents complex interfaces, which discourages researchers interested in performing this type of work. However, there are alternatives, namely in terms of open-source programming languages that, through packages created by a dedicated community, provide free strategies to apply IRT. In our case, we used R software (version 4.0.5), R Studio Desktop (version 1.4.1106), and the following R-packages: "*ltm*" (Navarro, 2015), "*mirt*" (Chalmers, 2012), "sirt" (Robitzsch & Robitzsch, 2021), and "*eRm*" (Patrick et al., 2018).

How Do You Test the Assumptions of an IRT Analysis?

The methods used to address the unidimensionality assumption differed in the two studies. In the first study, unidimensionality was assessed based on two approaches. It is well known that assessing the dimensionality of exam response data is generally problematic, mainly since these data types are rarely strictly unidimensional. Considering this problem, our first approach was based on the possible existence of additional "minor" abilities along with the θ of interest. This refers to the concept of essential unidimensionality. Through the

DETECT index, we evaluated the internal structure of our data. DETECT (e.g., Monahan et al., 2007) is a nonparametric dimensionality assessment test that clusters dichotomously scored items into dimensions and provides an indicator of multidimensionality. The second approach consisted of a consecutive, unidimensional method. More specifically, we used two partition criteria of the sample to evaluate if the item parameters were invariant across subgroup splits through the Andersen's LRT test. In the second study, we took a more straightforward approach to test the factor structure of the measure: since we were dealing with polytomous data, we used an exploratory factor analysis (EFA) followed by a confirmatory factor analysis (CFA) on the items that made up the measurement instrument.

The assumption of local independence is often omitted in practice, most notably because of the difficulty in obtaining adequate results that ensure this assumption. Even though we were aware of this limitation, we conducted a series of tests in both studies to validate this assumption. The first study tested local independence at the global level through the T11-test. If the test result is significant, one can consider that the item responses may be correlated or dependent after controlling for one factor. Unfortunately, we obtained significant results in the T11-test. However, since we were dealing with exam data, one possible reason for this significance may be due to the fact that the items belong to the same strand in a multiple-strand test. In the second study, we evaluated local independence by looking at the residual correlation matrix resulting from the CFA model. In this test, a value of 0.20 or above the average residual correlation is considered for local dependence and therefore fails the assumption.

In both studies, monotonicity was evaluated through graphical visualization of the ICCs. However, in the second study, we decided to deepen the validation of this assumption by fitting a nonparametric IRT model through Mokken scaling (e.g., Lamejier et al., 2020, Van der Ark, 2007). If the scalability coefficients are greater than 0.3, this assumption is assured (e.g., Lamejier et al., 2020; Mokken, 1971).

What About Model Diagnostics?

The way we conducted model diagnostics differed across both studies, much because of the IRT models selected and because there are no generally accepted goodness-of-fit measures. At the fit level of the model, in the first study, we used the Andersen LRT test to evaluate if the item parameters were invariant across subgroup splits. If the result was nonstatistically significant, model fit was assured. In the second study, we

© 2023 SAGE Publications Ltd

compared a common-slope GRM (equal discrimination values) with an unconstrained GRM (Gray-Little et al., 1997). Values of the *p*-value of the log-likelihood ratio test, the corrected Akaike's information criterion (AIC), and the Bayesian information criterion (BIC) values were considered to reject one of the models. At the item level, we used the infit and outfit mean square values in the first study. In the second study, we opted for the Pearson's χ^2 statistics (S-χ^2). Finally, the standardized fit index (Zh) was used in both studies to assess person fit (e.g., Drasgow et al., 1985).

Estimation of Item Parameters: Which Values Should We Pay Attention To?

The available literature presents various rules of thumb regarding the appropriate values for the parameters that characterize the items of a test. We applied the following guidelines throughout our research: items with *b* values greater than 0.75 were considered difficult; between −0.75 and 0.75 were considered medium difficulty, and items with values lower than −0.75 were considered easy (Nunes et al., 2017). Regarding the *a* parameter, we considered estimates below 0.24 and estimates larger than 3.00 problematic (Baker, 1985; Taylor, 2015).

In Need of a Visual Guide: How to Interpret the Graphical Outputs From IRT?

The graphs from an IRT analysis might occasionally be challenging to interpret. However, one must rely heavily on these graphs to interpret the properties of our measurement instrument, which is why an understanding of the plots obtained is necessary—for further readings on this topic, see Partchev (2004). We evaluated our data mainly through the following graphics.

ICCs

In these plots, the person's ability is plotted along the horizontal axis and is denoted with θ. On the other hand, the vertical axis represents the probability of a correct response considering the θ level of the respondent and is denoted with P(θ). As written above, this probability depends on the properties of the item parameters estimated by the IRT model. In our first study, each ICC captured the probability of a correct response through the

© 2023 SAGE Publications Ltd

b parameter and the ability level of the respondent. Figure 3 represents the ICC of an item (M1) analyzed in the first study. Based on this plot, we can identify the position of parameter *b* on the horizontal axis (ability) at the point where the predicted probability of a correct response equals 0.5. Considering the item represented in the figure, we see that it appears to have a difficulty close to 1. In our second study, the model calculated, for each item, a discrimination parameter and four threshold parameters (points of θ where the probability of choosing a response category or higher is 0.5—same as difficulty), considering that there were five response options. In this case, we looked at the probability of endorsing a response—multiple responses ranging from "completely disagree" to "completely agree", as opposed to the correctness of a response like in study one. Each ICC represented the probability of respondents selecting a certain answer on the scale (1–5) at various levels of the θ. Figure 4 represents the ICC of an item (Item 6) analyzed in study two. In this figure, each curve represents the probability of endorsing a response category (P1 = "Strongly disagree," P2 = "Disagree," P3 = "Neutral," P4 = "Agree," and P5 = "Strongly agree"). The ICCs revealed that the response categories covered a wide range of θ. For most response categories, as θ increased, the probability of endorsing a category increased and decreased as responses transitioned to the next higher category. The curve for option 1 (P1) is high at the lowest levels of the latent construct and gradually declines as people become more populist. Option 5 (P5) curve gradually increases as people present more populist attitudes.

Figure 3. Rasch Model item characteristic curve for item M1. Theta: knowledge in surgery on the latent trait continuum.

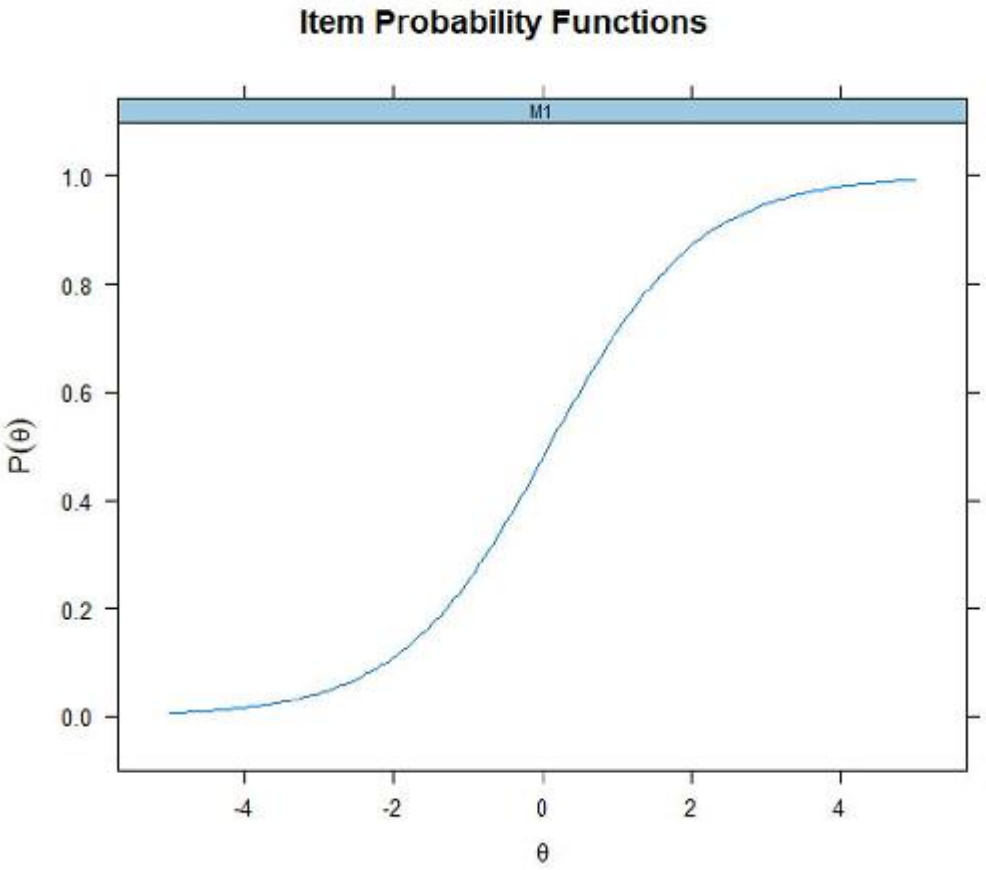
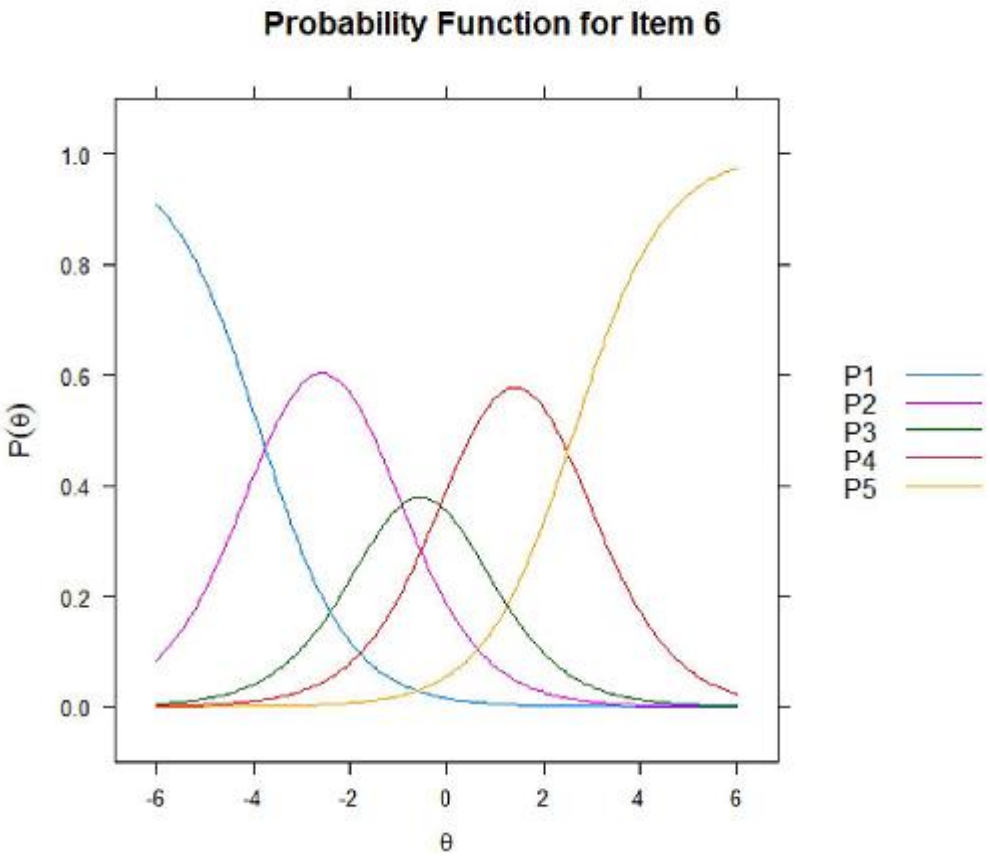


Figure 4. Graded Response Model Item category characteristic curves for item 6. Theta: level of populist attitudes on the latent trait continuum.

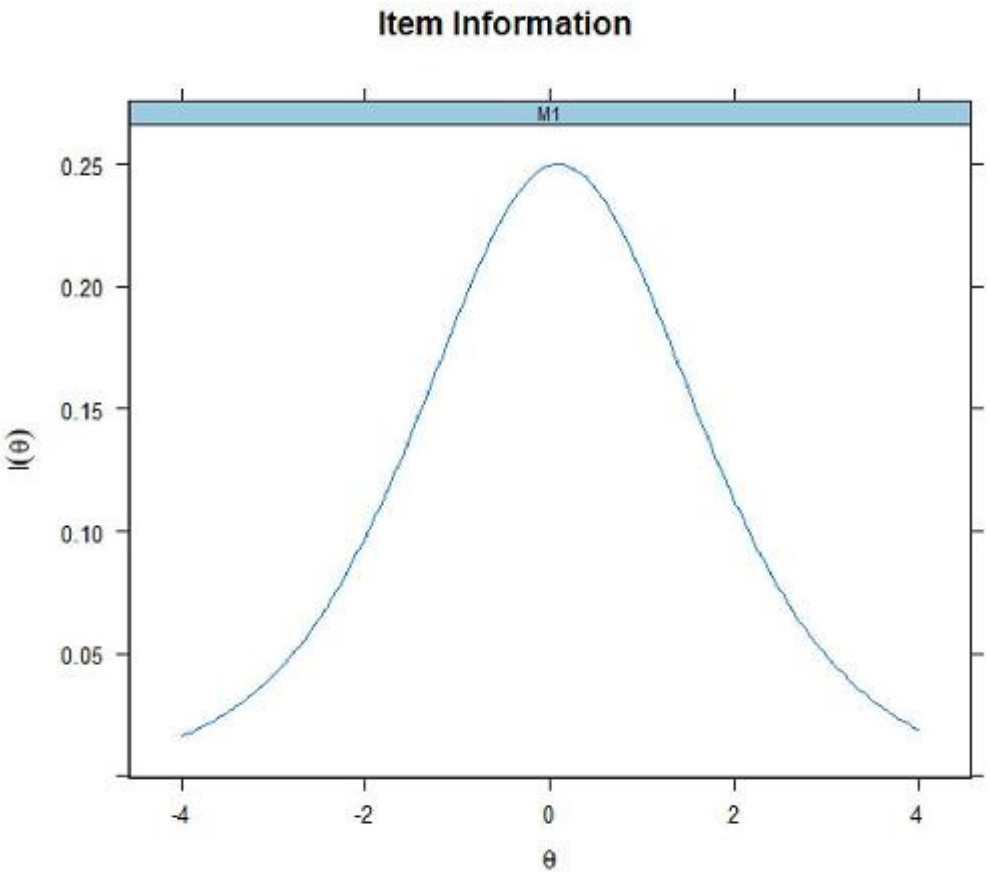


© 2023 SAGE Publications Ltd

IIFs

The IIFs display the accuracy and precision with which each item captures the latent trait at different attribute levels. Similarly to the ICCs, the person's θ is plotted along the horizontal axis. However, the vertical axis represents the amount of information considering the respondent's ability and is denoted with $I(\theta)$. The degree of information an item provides depends on the parameters estimated by the model. In the first study, since we used the RM, the information provided by each item depended only on how well the difficulty of the item corresponded to the respondent's θ. Here, the maximum value of the IIF occurs at the point where the probabilities of a correct/incorrect response are equal to 0.5, with the function decreasing as the ability is smaller/larger than the item difficulty (Partchev, 2004). In the first study, all IIF presented the same shape and maximum, differing only in the position along the ability axis. Figure 5 represents the IIF of an item (M1) analyzed in the first study. The item conveyed relatively high information values in the middle of the θ distribution. In the second study, since we applied the GRM, the IIFs attained their maximum at item difficulty. However, the shapes and values of the maximum were also influenced by the discrimination parameter, leading to dissimilar curves.

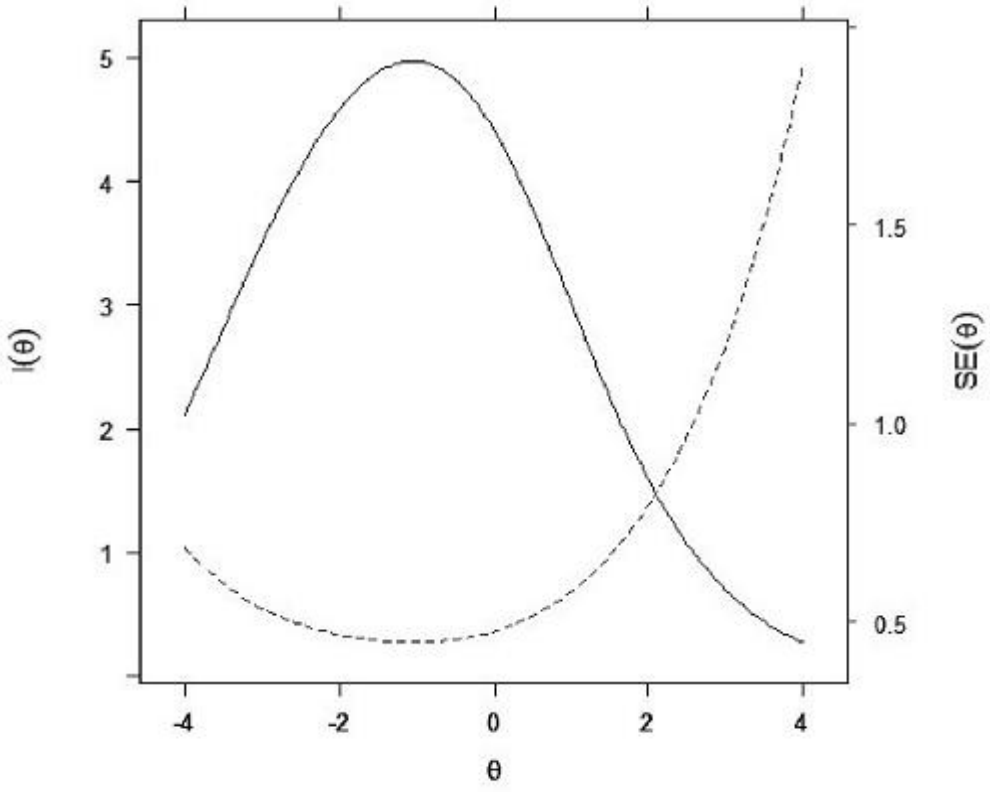
Figure 5. Item information function for item M1 for the Rasch Model.



TIFs

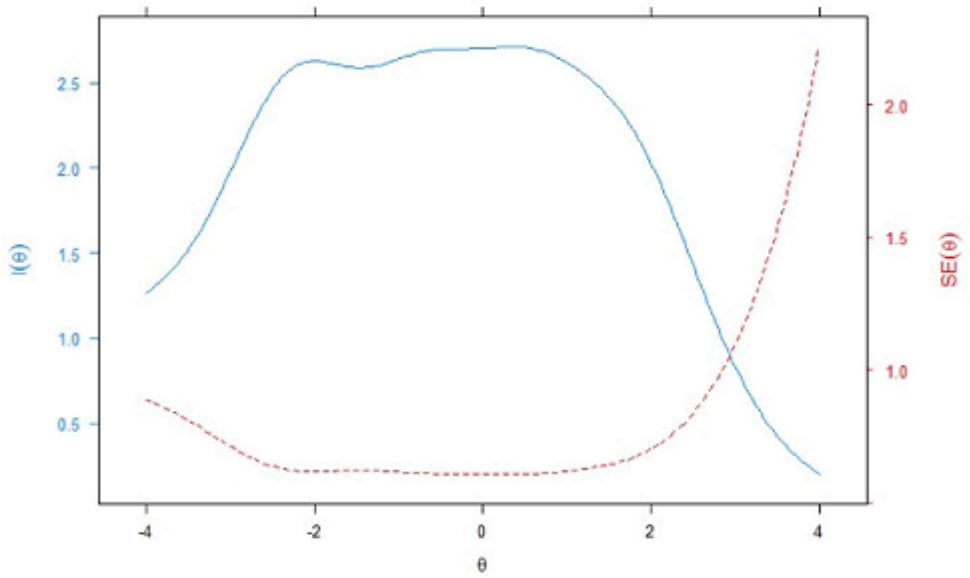
IIFs alone are not informative enough. We need more information to evenly spread over the possible range of the θ under measurement. Since items are assumed to be locally independent, the TIF aggregates all IIFs across the whole test and forecasts the expected test score as a function of θ and item parameters. One can also compute the conditional standard error of measurement (cSEM) with the TIF to find in which trait area our instrument can evaluate a person's location on the trait of interest with good precision (e.g., Mair, 2020). The TIF reveals how accurately the test captures the latent trait at different levels of the attribute and considers the existence of measurement error to compare it with the ability estimate. Figures 6 and 7 present the TIF for the items of the test in studies one and two, respectively. The dashed line represents the cSEM, while the solid line represents the TIF. Like the IIF, the shape of the TIF depends on the item parameters, especially the discrimination parameters. The lowest point of the cSEM matches the highest point of the TIF. This means that we have better accuracy for estimating the amount of θ for respondents near the point where the information on a test or item is at its highest.

Figure 6. Test information function and conditional standard error of measurement plot for the Rasch Model.



Note: The undashed line represents the test information function. The dashed line represents the conditional standard error of measurement function.

Figure 7. Test information function and conditional standard error of measurement plot for the Graded Response Model.



Notes: The undashed line represents the test information function. The dashed line represents the conditional standard error of measurement function.

Ability Score Estimation

The final step in both studies was to acquire estimates of θ for each respondent. IRT accurately estimates the respondent's position on the θ distribution**.** Considering a vector of responses to a set of items for a respondent with a certain θ level, one can calculate the probability of a correct/incorrect response to each item

© 2023 SAGE Publications Ltd

based on the parameters of the IRT model. Through the multiplication of these probability values, one obtains a joint likelihood function for the set of items for each response pattern. In our studies, we used the maximum likelihood (MML) estimator to produce θ estimates. The MML estimator is one of the most widely used methods for the estimation of latent trait scores. It operates by locating the highest point on the likelihood function (local maxima) and returning the value at which it occurs (e.g., Thompson, 2009). Based on this procedure, θ estimates were calculated for each unique observed response string, allowing us to compare test takers and evaluate respondents according to the level of their underlying θ.

Section Summary

- Open-source programming languages such as R provide proper statistical packages to employ an IRT analysis.
- Different approaches and statistical measures can be used to test the IRT assumptions and model diagnostics.
- The available literature offers a variety of guidelines regarding the appropriate values for the parameters that define test items.
- The graphs of an IRT analysis (ICCs, IIFs, and TIFs) are essential tools for understanding the properties of our measurement instruments.
- IRT uses a vector of responses to a set of items from a respondent with a certain level θ to calculate the probability of a correct/incorrect response based on model parameters.

Method in Action

We believe this sequence of steps is useful for any researcher conducting an IRT analysis. This method combines all the fundamentals of an IRT analysis and allows to characterize the psychometric properties of two instruments of different natures. The biggest problems we encountered in these two studies were mainly related to the testing of the unidimensionality and local independence assumptions in the first study. Regarding unidimensionality, when we started the analysis, we expected to find only a single dimension underlying the data. However, we found more than a dozen dimensions which, in essence, were nothing more than small factors aggregated into a single even larger factor underlying the data. The problem of local independence

was essentially related to the usual difficulty in ensuring this assumption. However, the strategies mentioned here allowed us to overcome these problems.

Section Summary

- The method offered combines the core concepts of an IRT analysis and allows describing the psychometric characteristics of instruments of various types.
- Some studies ignore the assumptions of unidimensionality and local independence since they can be challenging to ensure. There are, however, always workarounds and alternatives to put these presumptions to the test.

Practical Lessons Learned

Keep Your Code Clean

Conducting such an analysis can be exhausting, not least because of the number of procedures required to achieve the desired results. Keeping the code well-structured is recommended, and only maintaining what is strictly necessary in our scripts. This way, we avoid potential bugs and endless hours looking for a nagging error.

Importance of Testing Assumptions

Any mathematical model involves a set of premises regarding the data to which it is applied. These premises define the connections between the model's observable and unobservable constructs (Hambleton et al., 1991). Only if the assumptions are secure can we be sure that the results obtained are reliable and that we can extrapolate inferences from our research. It is also important to note that there are other strategies available to assess some of the listed IRT assumptions, even if these were not used in our studies. Besides the DETECT index, the consecutive unidimensional method, EFA, and CFA, there are at least two more statistical approaches for assessing the unidimensionality assumption: multidimensional scaling and comparison of

nomological networks (Bartram et al., 2018; Sireci et al., 2005). Likewise, in addition to the residual correlation matrix and the T11 test, one could also use the Yen's Q3 statistic to determine local independence (Yen, 1984).

Item Performance

Item performance refers to the probability of answering the item correctly/endorsing a particular response category. It is important to know the performance of the items of our measuring instrument to be able to evaluate our participants properly. However, we shouldn't decide whether to keep or remove a particular item from our instrument based simply and only on the values that define it. It is necessary to consider a larger range of factors and different viewpoints. Understanding item construction processes and the guidelines that govern those procedures is always helpful. Likewise, processes for identifying item-level bias phenomena (e.g., Differential item functioning) should also be understood.

Know How to Visually Interpret Your Graphs

More important than knowing how to interpret the numbers in a table is knowing how to "read" the graphs obtained in an IRT analysis. Through the visualization of IRT graphs, we can usually obtain a general reading of the properties of the instrument we are evaluating.

Section Summary

- This kind of analysis requires organization. Keep your code tidy.
- Extrapolating findings from our research is only possible if the assumptions of our analysis are reliable.
- To accurately estimate our respondents, it is crucial to understand how well the items of our measurement instrument performed.
- Understanding how to read the graphs produced by an IRT analysis is of great help.

Conclusion

There are benefits and drawbacks when using IRT in psychometrical analysis. It is nuclear to understand both the benefits and drawbacks of this approach. This cycle of steps enables the application of a reliable IRT analysis that respects the assumptions underlying this theory. IRT analysis provides more reliable results than CTT. Still, it is also more complex and more difficult to apply, considering the limitations listed, much at the level of the software available. It is prudent to weigh the benefits and drawbacks of this approach carefully before conducting this type of analysis. Analyses of this type may seem confusing at first glance due to the number of underlying topics and the set of procedures required. In this study, we believe practical advice has been provided to help junior researchers deal with these issues.

Section Summary

- Understanding both the advantages and disadvantages of an IRT approach is essential for your research.
- The sequence of actions described enables the use of a trustworthy IRT analysis.
- With the aid of some helpful instructions that have been provided here, junior researchers can handle the difficulties of such an analysis.

Classroom Discussion Questions

1. Should researchers still use CTT in light of the growing IRT usage?
2. What kinds of difficulties might you expect when performing this kind of analysis for your research?
3. IRT applications cover the measurement of a variety of latent constructs in a variety of disciplines through a complex framework. Should it, however, be the only method used to judge an instrument's validity and reliability?

Multiple-Choice Quiz Questions

1. The monotonicity assumption should be evaluated using which of the following plots?

a. IIF

Incorrect Answer

Feedback: This is not the correct answer. The correct answer is C.

b. TIF

Incorrect Answer

Feedback: This is not the correct answer. The correct answer is C.

c. ICC

Correct Answer

Feedback: Well done, correct answer

2. 2. Which of these statistics is a useful measure to evaluate person fit?

◯ **a.** Zh statistic

Correct Answer

Feedback: Well done, correct answer

◯ **b.** Pearson's χ2 statistic

Incorrect Answer

Feedback: This is not the correct answer. The correct answer

is A.

c. Yen's Q3 statistic

Incorrect Answer

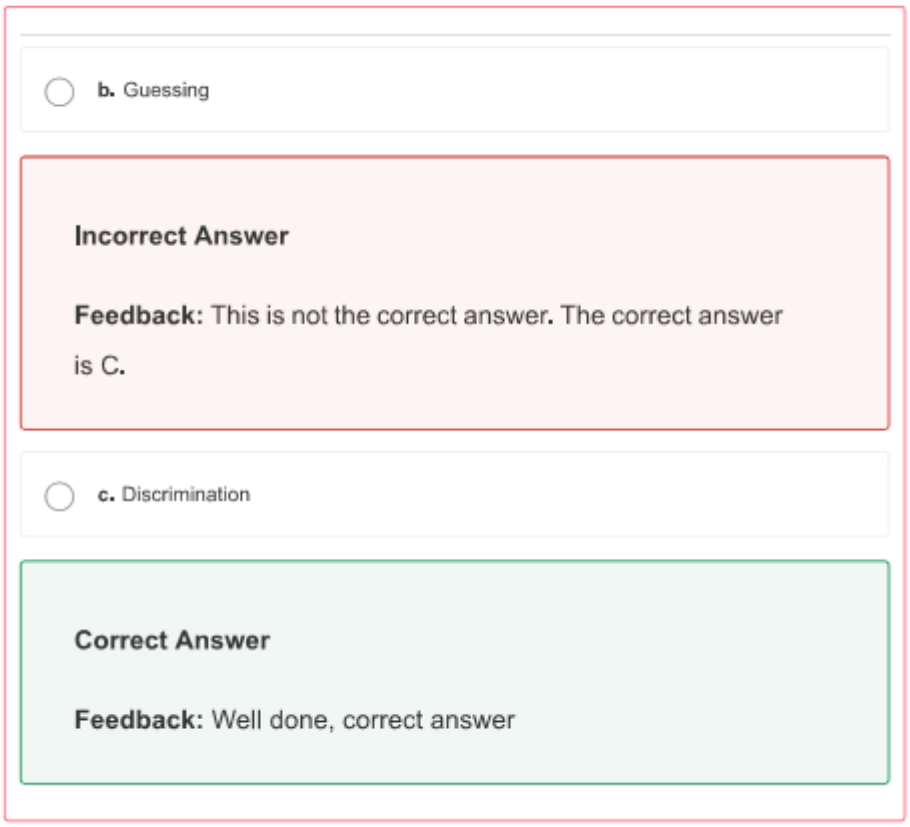
Feedback: This is not the correct answer. The correct answer is A.

3. Which of the following item parameters enables you to distinguish respondents with different levels of θ?

a. Difficulty

Incorrect Answer

Feedback: This is not the correct answer. The correct answer is C.



Further Reading

De Ayala, R. J. (2013). *The theory and practice of item response theory*. Guilford Publications.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press. https://doi.org/10.4324/9781410605269

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Web Resources

Blanchard, A., Epifania, O., & Tanis, C. (2018). A learnr tutorial on item response theory [IRT Tutorial]. https://r.tquant.eu/tquant/KULeuven/IRT/

International Test Comission. (2020). Intestcom.org [Intestcom.org].

Rudner, L. (2001). Item response theory - edres.org [Item Response Theory]. https://edres.org/irt/

References

Adams, D., & Farmer, G. L. (2022). Graded response modeling of the DESSA's self- and social awareness subscales in a special needs sample. *Youth*, 2(2), 98–112. https://doi.org/10.3390/youth2020008

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, 43(4), 561–573. https://doi.org/10.1007/BF02293814

Baker, F. B. (1985). Book review Item response theory: Principles and applications. *Applied Psychological Measurement*, 9(3), 337–339. https://doi.org/10.1177/014662168500900315

Bartram, D, Berberoglu, G, Grégoire, J, Hambleton, R, Muniz, J, & Vijver, F. (2018). ITC Guidelines for Translating and Adapting Tests (Second Edition). *International Journal of Testing*, 18(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166

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

Darrell Bock, R. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, 37(1), 29–51. https://doi.org/10.1007/BF02291411

Depaoli, S, Tiemensma, J, & Felt, J. M. (2018). Assessment of health surveys: Fitting a multidimensional graded response model. *Psychology, Health & Medicine*, 23(sup1), 13–31. https://doi.org/10.1080/13548506.2018.1447136

Desjardins, C. D., & Bulut, O. (2018). *Handbook of educational measurement and psychometrics using R*. CRC Press. https://doi.org/10.1201/b20498

Drasgow, F, Levine, M. V, & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, 38(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Edelen, M. O, & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16 Suppl 1, 5–18. https://doi.org/10.1007/s11136-007-9198-0

Embretson, S. E, & Reise, S. P. (2013). Item Response Theory (Vol. 27). https://doi.org/10.4324/9781410605269

Gray-Little, B., Williams, V. S. L., & Hancock, T. D. (1997). An item response theory analysis of the Rosenberg self-esteem scale. *Personality and Social Psychology Bulletin*, 23(5), 443–451. https://doi.org/10.1177/0146167297235001

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991b). *Fundamentals of item response theory* (Vol. 2). Sage.

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

Lameijer, C. M, van Bruggen, S. G. J, Haan, E. J. A, Van Deurzen, D. F. P, Van der Elst, K, Stouten, V, Kaat, A. J, Roorda, L. D, & Terwee, C. B. (2020). Graded response model fit, measurement invariance and (comparative) precision of the Dutch-Flemish PROMIS® Upper Extremity V2.0 item bank in patients with upper extremity disorders. *BMC Musculoskeletal Disorders*, 21(1), 1–17. https://doi.org/10.1186/s12891-020-3178-8

Mair, P. (2020). Modern Psychometrics With R. *Technometrics*, 62(1), 135–137. https://doi.org/10.1080/00401706.2019.1708675

Masters, G. N. (1982). A rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. https://doi.org/10.1007/BF02296272

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Mouton. https://doi.org/10.1515/9783110813203

Monahan, P. O., Stump, T. E., Finch, H., & Hambleton, R. K. (2007). Bias of exploratory and cross-validated DETECT index under unidimensionality. *Applied Psychological Measurement*, 31(6), 483–503. https://doi.org/

10.1177/0146621606292216

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series*, 1992(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Navarro, D. J. (2015). Learning statistics with R: A tutorial for psychology students and other beginners. University of Adelaide. http://health.adelaide.edu.au/psychology/ccs/teaching/lsr/

Nunes, S., Oliveira, T., & Oliveira, A. (2017). Item response theory – A first approach. In International Conference of Numerical Analysis and Applied Mathematics (ICNAAM 2016) (Vol. 1863, pp. 1–5). https://doi.org/10.1063/1.4992683

Paek, I., & Cole, K. (2020). *Using R for item response theory*. Routledge.

Partchev, I. (2004). A visual guide to item response theory.

Patrick, A., Hatzinger, R., Maier, M. J., Rusch, T., & Mair, M. P. (2018). Package"ERM.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment test. In *Studies in Mathematical Psychology* (Vol. 1). Danmarks Paedagogiske Institut.

Robitzsch, A., & Robitzsch, M. A. (2021). Package"irt.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, 34(S1), 1–97. https://doi.org/10.1007/BF03372160

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). *Statistical methods for identifying flaws in the test adaptation process* (pp. 93–115). Adapting Educational and Psychological Tests for Cross-Cultural Assessment.

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, 75(3), 454–473. https://doi.org/10.1007/s11336-010-9163-7

Taylor, J. M. (2015). Psychometric analysis of the ten-item perceived stress scale. *Psychological Assessment*, 27(1), 90–101. https://doi.org/10.1037/a0038100

Thompson, N. A. (2009). *Ability estimation with item response theory*. Assessment Systems Corporation.

Van der Ark, L. A. (2007). Mokken scale analysis in R. *Journal of Statistical Software*, 20, 1–19.

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-

parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. https://doi.org/10.1177/014662168400800201

https://dx.doi.org/10.4135/9781529627459

CHAPTER VI

PROGRESS IS IMPOSSIBLE WITHOUT CHANGE: IMPLEMENTING AUTOMATIC ITEM GENERATION IN MEDICAL KNOWLEDGE PROGRESS TESTING

This chapter is based on the following publication: **Falcão, F**., Pereira, D.S., Pêgo, J.M. et al. Progress is impossible without change: implementing automatic item generation in medical knowledge progress testing. *Education and Information Technologies* (2023). https://doi.org/10.1007/s10639-023-12014-x

Education and Information Technologies
https://doi.org/10.1007/s10639-023-12014-x

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
© 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

Extended author information available on the last page of the article

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 an 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 judgements 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.

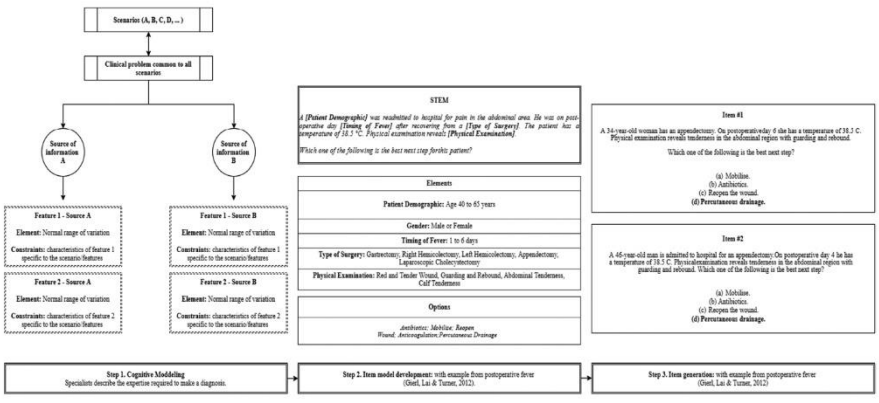


Fig. 1 Three-step process for generating medical MCQs based on AIG

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:

$$\boldsymbol{P}\left(\boldsymbol{Y}_{ij}=1|\theta_j,\boldsymbol{b}_i\right)=\frac{\exp\left(\boldsymbol{D}\left(\theta_j-\boldsymbol{b}_i\right)\right)}{1+\exp\left(\boldsymbol{D}\left(\theta_j-\boldsymbol{b}_i\right)\right)} \quad (1)$$

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, …, J), *bi* is the item difficulty parameter for item I (i=1, …, 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 presumptions 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). Fift 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

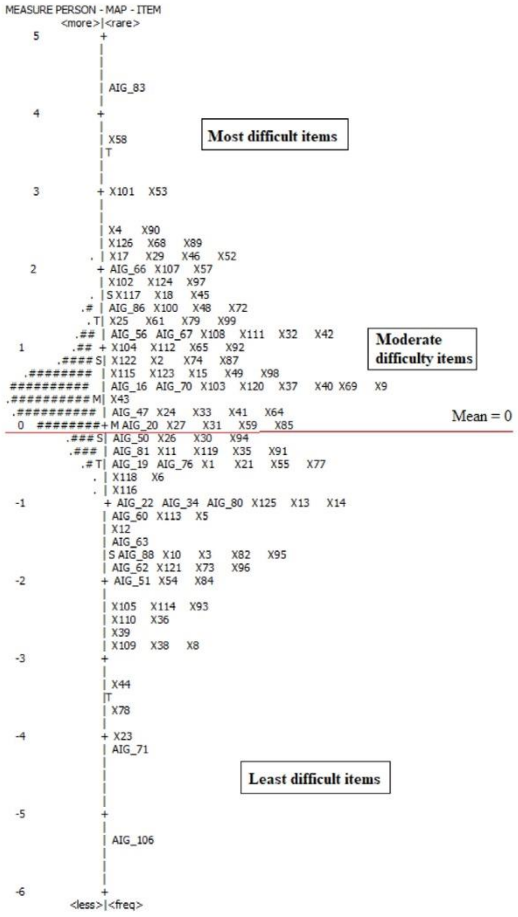


Fig. 2 Map of the difficulty level of the questions from the RM
Note. Area on the left represents the distribution of the subject's θ; Area on the right represents the distribution of items; Items with the highest difficulty level are at the top, while the easiest items are at the bottom; Each # represents 4 students; each '.' represents 1–4 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

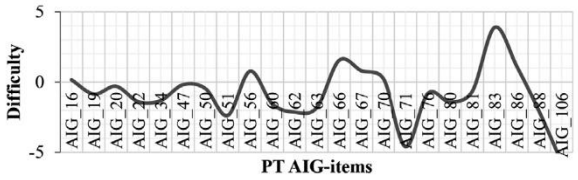


Fig. 3 Bi values for the PT AIG-items

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

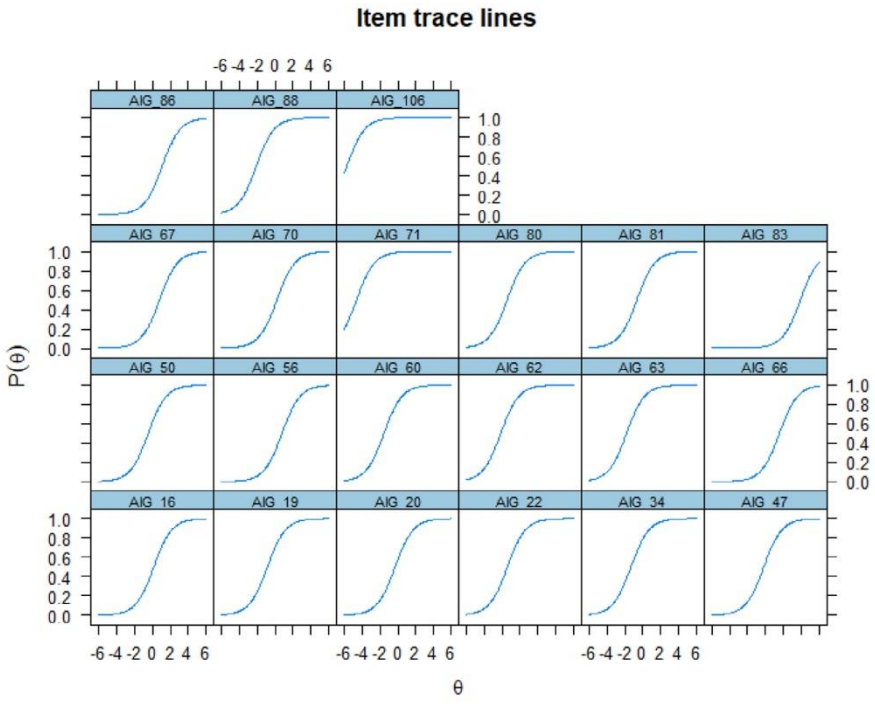


Fig. 4 ICCs for the PT AIG-items

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.

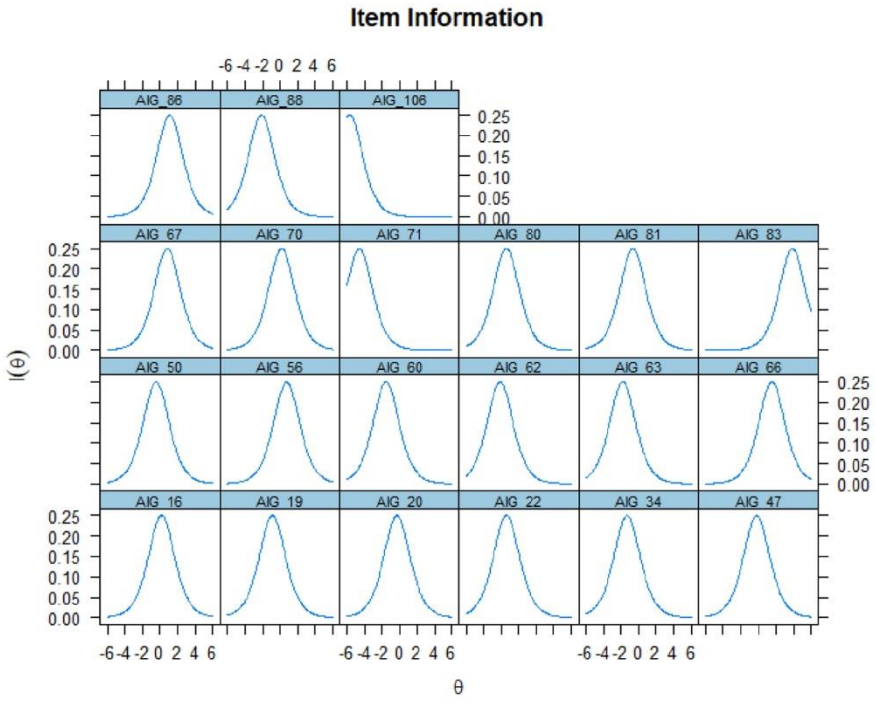


Fig. 5 IIFs for the PT-items

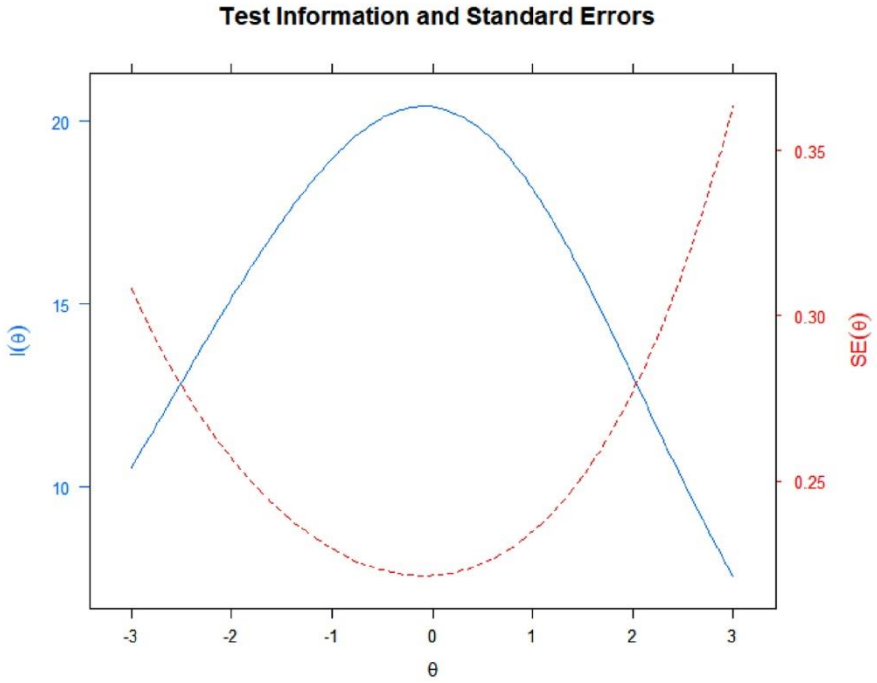


Fig. 6 TIF for the PT

Table 1 MH procedure for measuring and detecting DIF

Item	MH DIF statistic	αMH	ΔMH	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 $\chi 2$) 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 (ΔMH) were used to supplement the chi-square test of statistical significance. DIF in the PT was balanced. Most ΔMH 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 ≤ 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 ≤ 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% 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 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 s ± 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	% of students choosing option					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
Intercept	61	1.02	59.9	1.10	59.9	1.05	59.9	1.05
Student Level								
Number of responses given			15.6**	0.47	15.6**	0.47	15.3*	0.62
Item Level								
Item type			3.29**	0.77	3.29*	0.78	3.29*	0.78
Interaction								
Item type * Number of responses given							0.84	0.92
Residual variance								
Student level	243						223	
Item level	2062						1890	
Explained variance (R^2)								
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 ($B_{switching}$=3.29; B_{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, altough 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 mechani, 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_i, 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 Icognitus4All—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/10.1080/10872981.2018.1560862.

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.

Hubley, 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., Wijnia, L., Köhlen, 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., Adji, 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

CHAPTER VII

GENERAL DISCUSSION

Overview

Innovation in contemporary test delivery and test design poses new challenges to traditional item development methods (Circi et al., 2023; Gierl, Shin, et al., 2022). Concurrently, as AI technology advances and becomes more prevalent in schools, pedagogical issues will only increase (Pereira et al., 2023). AIG may be a suitable approach to address the challenges raised by modern testing due to its integration of models with computer technology and its consistent supply of newly generated items (Embretson & Kingston, 2018; Liu, 2023; Setiawan et al., 2022). While traditional item development methods are time-consuming (because they require experts to perform the taxing chores of manually writing a considerable number of items), AIG delivers a more methodical and cost-effective item development procedure (Fu et al., 2022; Kucharski et al., 2023). Although the current literature seems to support the use of AIG in various educational situations, there is still a lack of information about AIG and its processes, which inhibits the spread of this methodology and highlights the need for more research on the topic (Falcão et al., 2022; Harrison et al., 2017). With this in mind, our goal throughout this Thesis was to provide new insights into the "grey areas" of AIG. Our focus centred primarily on the following points of the AIG procedure in the context of medical education assessment: item quality and psychometric properties of AIG-items; ease of learning; validity; and compatibility with different assessment modalities. In the following paragraphs, we describe the main contributions of this Thesis to the current state of the art and explore how our findings fit into the literature.

Theoretical Approach - Act 1: Technology-based procedures in medical education assessment

As a first note, we intended to perform a theoretical approach to the use of technology in educational assessment and see where AIG could fit into this framework. This strategy resulted in two distinct objectives with a strong theoretical emphasis. The first was to discuss the impact of more technology-based assessment procedures in medical education assessment (Chapter II); The second was to conduct a narrative review of the literature on the feasibility, validity, and item quality of AIG in medical education assessment (Chapter III). The following paragraphs review the findings of these two outcomes, offering explanations and interpretations based on the Thesis objectives and relevant research.

In Chapter II, we built the argument that the use of computer modules already has a significant impact on medical education assessment. We explored how assessment in medical

education is changing, the presence of AI, the trend toward CBT, and how assessment might benefit from using technology in terms of efficiency and accuracy. Our perspective here was based on current literature on the topic, as well as our own personal experiences. Indeed, medical education assessment has evolved in response to societal expectations, the healthcare environment, the information boom, changes in the physician's role, and new instructional methodologies (Guze, 2015; Pei & Wu, 2019). Therefore, the methods employed to evaluate medical students must be rethought to prepare doctors for today's society, where the internet provides widespread digital information (Pei & Wu, 2019). Computer technology should not be considered an obstacle, but a possible solution for creating more realistic/stimulating learning and assessment settings (Pereira et al., 2023).

As mentioned, most medical schools already established CBT as the standard testing format (Huh, 2022). Nevertheless, despite all the progress made, medical education assessment continues to lag behind the scientific breakthroughs made possible by modern technology (Lee et al., 2021). The COVID epidemic has only highlighted this problem, with the transition to online assessment exposing the need for more testing resources, particularly in terms of the availability of test items (Gierl et al., 2021; Hilburg et al., 2020). During this period, medical schools rapidly expanded their online training and pedagogical content offerings, as well as the technological training of their teachers, particularly through online courses (Goh & Sandars, 2020). This increase in large-scale online assessment posed serious challenges to the iterative approaches used to develop, revise, and refine items that are part of more conventional item development methods (Circi et al., 2023; Gierl, Swygert, et al., 2022). This query is congruent with the impetus that led us to briefly unveil AIG as one viable solution to the burden of item development brought on by the advent of large-scale assessment. However, examining the subject from a wide perspective, such as in chapter II, is insufficient to dispel the misinformation surrounding AIG and support its implementation in medical schools and other educational institutions. As a result, we needed to narrow the scope, leading us to our theoretical approach's second act.

Theoretical Approach - Act 2: State of things

The second act of this more theoretical approach was to conduct a narrative review on the feasibility, validity, and item quality of AIG in medical education assessment (Chapter III). Following the line of thought initiated on Chapter II, we began Chapter III by framing the changes in medical education assessment in terms of the urge for more testing items. We then introduced the concept

of AIG, described the three-step procedure used to generate medical MCQs and stressed the importance of conducting a comprehensive synthesis of the available literature to collect relevant data. More specifically, our purpose here was to conduct a broad evaluation of AIG's feasibility, validity, and item quality in medical education assessment. Peer-reviewed, English-language publications published between 2000 and 2021 were searched in 5 electronic databases. 119 records were screened and a total of 13 publications were examined in accordance with the inclusion criteria. However, only 10 of these publications went on to the data synthesis stage. We believe these numbers reflect the paucity of material generated in the last 20 years and underpin the lack of awareness around AIG in medical education assessment.

The synthesized data and the validation framework employed in the paper clearly suggest that AIG is a valid and feasible strategy for producing high-quality test items and to ease the item development burden present in medical education assessment. Furthermore, the data provided demonstrated the role that AIG may play in addressing these difficulties. For example, we determined that the AIG procedure could create 80 (Pugh et al., 2016) to 1248 (Gierl & Lai, 2013) items from a single cognitive model in a reasonably short period (between 1-6 h; approximately 208 items per hour). Regarding ease of learning, we obtained data claiming that both novices and specialists could use AIG relatively well soon after learning the principles of AIG after attending workshops on the topic (Gierl & Lai, 2013; Pugh et al., 2016). Additionally, no significant differences in item quality/psychometric features between items produced using more traditional methods and AIG-items were found. Lastly, the validity framework used in the paper supported the use of AIG-items in medical education assessment as the AIG procedure gathered appraisals of evidence in majority of the inferences contained in the framework. These results were seen as encouraging for the aim of our work and only cemented the argument we had built in chapter II.

As of our publication, no review concentrating on the feasibility, validity, and item quality of AIG in medical education assessment had been conducted. One year following the publication of our study, a mini review focusing on AIG was published (Circi et al., 2023). However, this review was more focused on the knowledge of AIG in the context of educational assessment and not on medical education assessment. In line with our findings, Circi and colleagues (2023) state that using computer modules to generate items enables the efficient manufacture of test items and highlights the need for further work in various fields for the widespread application of AIG. The authors also point out that most AIG-related work to date is experimental and that AIG is not frequently employed in more operational situations because of privacy and confidentiality

regulations that restrict public access to the processes used by testing organizations (Circi et al., 2023).

Apart from our work, this is the only AIG review undertaken. However, one must recognize that our review consists of only ten included citations, a considerably small number of included studies. The very review carried out by Circi and colleagues (2023) is a mini-review. These numbers reveal the scarcity of literature on the topic and the need for more empirical data on the AIG procedure. Therefore, AIG's approaches still require additional testing and research in order to be fully understood (Circi et al., 2023). Since more studies on AIG have been published since the publication of our work, an updated review would be especially useful to obtain the most latent data on the subject.

Empirical Testing

To empirically test the argument built throughout chapters II and III, we conducted two studies in Chapter IV that sought to assess the quality of AIG-items as well as the efficiency, ease of learning (learnability) and validity of the AIG procedure. Based on an operational multimodal approach, we performed a series of comparisons between AIG-items and manually written items to evaluate the AIG procedure and the use of AIG-items in medical education assessment. In the first study, participants with different levels of clinical knowledge and item writing experience created medical MCQs both manually and using AIG. Items produced by the participants were then compared in terms of efficiency and learnability. In the second study, a summative exam composed of manually written and automatically generated items in the content area of surgery was examined through a psychometric lens. In both studies, we found convincing evidence that the AIG procedure is efficient, and that AIG-items were quite similar to the manually written items not only in terms of structure, but also in terms of quality and psychometric properties, meeting the high-standards required for inclusion in high-stakes testing. Furthermore, our results suggested that the time spent constructing cognitive models and the number of items produced using AIG were unaffected by the participants' item writing experience or clinical knowledge. This indicates that the AIG procedure is easy to learn and does not require extensive training.

Most of our findings are supported by past and present literature on AIG (e.g., Gierl et al., 2021; Gierl, Swygert, et al., 2022). However, there are a few considerations to bear in mind. First, it is worth mentioning that asserting AIG as an efficient item development strategy is not necessarily novel. The available literature on the topic demonstrates that AIG can generate many test items,

even in medical education assessment (e.g., Kosh et al., 2019; Royal et al., 2018). Second, regarding the quality of the items, studies prove that the AIG procedure can produce high-quality items with adequate psychometric properties (e.g., Gierl & Lai, 2012; Leslie & Gierl, 2023). Therefore, rather than emphasizing these two topics, we believe that the analyses of learnability and validity were the main strengths of these two studies. There are few papers available that address these topics (e.g., Gierl, Swygert, et al., 2022; Gierl & Lai, 2013; Pugh et al., 2016), and even fewer (if any) that address both topics in the same work. Naturally, medical schools and other educational institutions may be persuaded to use the AIG approach if this procedure is certified valid and an easy-to-learn process through empirical testing. We believe that the work conducted in Chapter IV has the potential to impact the educational context, specifically in medical education assessment, not only in how medical schools can produce testing items but also in how they can manage their item banks due to limited resources. Furthermore, by demonstrating that participants with different levels of item writing experience and clinical knowledge were able to develop cognitive models capable of generating items effectively, we are taking a step forward in the knowledge of AIG and exploring a topic that has rarely, if ever, been explored in the literature. On a side note, it is worth mentioning that Chapter V illustrated our experience with one of the studies conducted in Chapter IV, serving as a reflective guide to some of the analyses conducted to help prospective researchers conduct similar research.

Compatibility with different assessment methods

Following the take-home message of the previous Chapters (AIG as a valid, efficient, and easy to learn assessment procedure capable of generating legions of test items that meet high stakes testing quality standards), it became imperative to understand whether AIG could be implemented in assessment modalities other than summative assessment. Throughout Chapter VI, we set out to study how AIG principles and practices perform in a formative assessment context. Similar to the experiment conducted in Chapter IV, the psychometric properties of AIG-items were compared to those manually written in the context of a medicine progress test (PT). This time, the analytic approach used was more comprehensive (IRT; Differential item functioning; Distractor performance) and included a qualitative component (expert review), and an innovative validity assessment approach (hierarchical linear modelling) focused on the response processes used by medical students to answer both AIG and manually written items. We found that the AIG-items successfully adapted to the PT and were of comparable quality (quantitatively and qualitatively) to those manually written. There were no significant differential item functioning problems, and most

distractors were adequate for assessing student's knowledge. Finally, we found that the response processing mechanisms used to answer both item types were very similar in terms of how much processing is done and how students interpret the questions, which may be considered as further evidence supporting the validity of the AIG procedure. To the best of our knowledge, this is the first time that the validity of AIG-items has been evaluated based on this method.

As previously discussed, the quality of AIG-items in general seems to be well established based on the number of research conducted on the subject. However, the quality of the distractors remains a source of contention (Attali et al., 2022; Gierl & Lai, 2018). Although the difference was not statistically significant, it is worth noting that manually written items performed somewhat better in terms of distractor performance. Our results are congruent with the literature on the topic (e.g., Gierl & Lai, 2013; Lai, Gierl, Touchie, et al., 2016), which asserts that AIG-items may be less credible than those manually written in terms of distractors. A possible reason for this could be that the three-step procedure has been explicitly described for item and correct option creation, but only implicitly for the incorrect alternatives, which requires new data-driven systematic processes to develop both correct and incorrect options (Lai, Gierl, Touchie, et al., 2016; Shin et al., 2019). Systematic distractor generation approaches refer to extensions of the modelling process used to construct test items that guarantees each item has options suited to the cognitive model (Lai, Gierl, Touchie, et al., 2016; Leslie & Gierl, 2023). Thus, the features associated with the correct option can also produce distractors (Gierl & Lai, 2018). However, few research has been conducted on the use of enhanced intelligence in the development of distractors so far (Shin et al., 2019).

Furthermore, other aspects of the AIG procedure within formative assessment merit further consideration, such as the strategies for developing the rationales for solving the problems. Since feedback can be developed as part of the generative process (because the logic used to gather all of the correct and incorrect possibilities may also be used to identify the solutions and provide the explanation for that decision), it is important to discuss the methods available for the generation of the rationales (Gierl & Lai, 2018). Throughout our work in Chapter VI, we were not explicit about the processes used to generate feedback on the PT under study. In addition, there was no evaluation of the feedback used. These data would be of particular relevance to the objective of this section of this Thesis. The explanation for this omission is that the approach employed in Chapter VI is still in the experimental stage, leading us to analyse the items used in the PT rather than the feedback. This detail represents a significant restriction of our effort. It is important to

note, however, that generating rationales is complex because the approaches available should consider different sources of feedback (Gierl et al., 2021). Indeed, studies focus on the methods that can generate rationales (Gierl et al., 2021; Gierl & Lai, 2018). However, few researchers (Choi et al., 2018; Gierl & Lai, 2018; Xinxin, 2019) have tested these methods and evaluated the rationales generated. For an in-depth explanation of the methods used to generate rationales, we recommend the work of Gierl and colleagues (2021).

Another topic that deserves further consideration is the novel approach employed to assess the validity of the AIG procedure. To the best of our knowledge, no study has evaluated the validity of AIG-items in terms of response processes. Our approach is novel from a modern psychometric standpoint, as we tested the validity of AIG-item response processes using nested data. Since we have used data that is linked at the time students take the exam to obtain evidence concerning the fit between the construct and the nature of the response engaged in by the students (Padilla & Benítez, 2014), our work can be very useful in strengthening the argument for the validity of the AIG procedure.

Feedback is vital for clinical teaching by facilitating continuous learning and improvement (Kornegay et al., 2017). However, educators still approach providing student feedback with caution (Chowdhury & Kalu, 2004). The results acquired throughout this chapter indicate that the AIG procedure is compatible with formative assessment. By generating a considerable number of test items and allowing the possibility of creating rationales based on computer algorithms, medical schools will not only be making more forms of testing available but will also be ensuring that students receive adequate and secure feedback.

In summary, we believe that this Thesis addressed many shadow regions of AIG. Throughout the many chapters of this work, we have acquired evidence that promotes further research and encourages more implementations of AIG by combining theoretical and empirical approaches to evaluate its potential. Throughout this Thesis, concrete examples have been presented on how the AIG procedure can be used to generate thousands of quality test items in the field of medical education assessment through different statistical and psychometric analyses that prove the efficiency, quality, validity, and ease of learning of this procedure.

Future Perspectives and Practical Implications

AIG can be a powerful assessment tool for producing exam content more efficiently than traditional item writing methods (Babcock & Brunnert, 2023). It's simple learning approach, as well as its mass manufacturing, intelligent item calibration and management procedures, flairs to transform the future of medical education assessment, resulting in fundamental changes in the way we conduct testing, why we design tests, who we test, and, lastly, what do we test (Gierl et al., 2021). By employing the AIG procedure, item banks can be routinely updated and include enough testing items to be used in safe large-scale evaluations. Additionally, experts will no longer need to write down information individually, but instead construct highly structured and uniform representations of the areas under evaluation, leaving computer algorithms to handle the rest of the task (Kucharski et al., 2023). Because most medical schools have already deployed CBT systems as their primary assessment format, and due to the constant need for more testing items, we believe that the interest in AIG is projected to expand in the near future.

This Thesis sheds new perspectives on AIG and the potential role that this procedure could play in medical education assessment. However, there are several questions surrounding the topic that we would like to investigate further in future research. In general, our findings support a position occasionally expressed in the literature, namely that AIG seems to have found a footing in medical education assessment (Götz et al., 2023). Indeed, the research on AIG in medical education assessment has grown in recent years, gathering a significant number of papers compared to applications in other areas of knowledge. However, this does not mean that the AIG procedure is exclusive to medical education assessment. There are papers available that have focused on applications of AIG beyond medical education assessment, such as mathematics (Embretson & Kingston, 2018), pharmacy (Leslie & Gierl, 2023) and even dentistry (Lai, Gierl, Byrne, et al., 2016). The findings are encouraging and seem to refute the notion that AIG's application resides solely within medical education assessment. Nonetheless, more research is required to determine whether AIG may be expanded to other areas of knowledge and domains of specialisation.

The validity of the AIG procedure also needs further exploration. Our interest in this point is understandable, considering that validity is the most essential consideration in the development of tests (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). Although we have offered evidence of the validity of the AIG procedure throughout the chapters of this Thesis, additional sources of evidence can be explored to further establish this approach's validity. Evidence based on test content and

evidence based on relationships to other variables external to the test could be relevant sources of the validity evidence that may be further explored. Future literature on the topic could contribute to generalising validity around the AIG procedure, attesting to it as a legitimate instrument for producing test items. This could help to debunk any assumption claiming that creating test items based on algorithms could somehow undermine educational assessment.

As mentioned, one of the most challenging topics to address under the AIG framework continues to be the generation of distractors. To date, there is no clear methodology for developing distractors. To make matters worse, standard proposals for distractor generation using traditional methods cannot be employed within the AIG-procedure (Gierl et al., 2021). Given this issue, we firmly believe and reinforce the message that new methods for developing distractors are required. While these new methods are not yet available, the existing methods for generating distractors should be further examined and compared (in terms of performance, efficiency, advantages, and limitations) to see which can produce quality distractors and ensure test security.

As CBT permits educational tests to yield both formative and summative inferences, we believe that the feedback received by students in formative assessments composed by AIG-items should also be further analysed. As previously stated, the available literature already provides some suggestions on how AIG might be used to provide feedback to students (Lai, Gierl, Touchie, et al., 2016). According to our knowledge, at least three methods for rationale generation have already been introduced to support formative feedback systems. However, one must consider the different types of feedback that can be provided to examinees before implementing one of these methods. Furthermore, when combined with the rationales, the development time of the items must also be considered since the development of answers and the rationales will add time to the three-stage AIG procedure due to the large number of possible outcomes.

We also expect that it will only be a matter of time before tests use AIG-items generated on the fly. This feature will be particularly attractive for Computerized Adaptive Testing (CAT). CATs are established evaluation methods that function in accordance with the characteristics of the respondents, measuring individual proficiency on the fly while adapting to the ability level (θ) of the student (Stochl et al., 2016). They promise accurate knowledge assessment through interactive test administrations where tailored sets of items are algorithmically selected depending on respondents' earlier responses, offering unbiased results regardless of the student's score (Seo & Choi, 2018; Weiss, 2004). However, building and maintaining sets of high-quality test items with

consistent psychometric properties estimates is one of the most daunting challenges CAT systems must face (Arendasy & Sommer, 2012; Luecht, 2012). This laborious procedure is avoided when items are generated algorithmically, which is why AIG may be particularly important for the future of CAT (Harrison et al., 2017). More specifically, AIG could offer CAT three primary advantages: first, a potentially low-cost solution to the otherwise expensive option of developing items by hand; Second, the enhancement of its item-writing procedures through the use of a cognitive basis for each item; Third and finally, it can reduce the need for pilot testing every item if they comply with the established standards for quality control on the permitted variation in statistical item attributes, which will build tests able to make a more precise estimate of the examinee's θ (Lai et al., 2009; Luecht, 2012).

The potential for AIG to supply large sets of items on the fly and with the necessary qualities to generate optimal scores under CAT is considerable (Luecht, 2012). Rather than building and calibrating each item separately, the algorithmic approach used by AIG is employed for test construction and in the estimation of the psychometric properties of the items beforehand (Harrison & Müllensiefen, 2018). There have been few attempts to combine CAT with AIG (Susanti et al., 2020). Thus far, strategies center on item modelling techniques (Bejar et al., 2003) and item difficulty estimators (Susanti et al., 2020). As of the date of submission of this Thesis, as far as medical education assessment is concerned, there is no standard approach for implementing AIG in the context of CAT with concrete results. Congruently, we are unaware of any study focused on the use of AIG with CAT in medical education assessment. This may be due to an unfamiliarity of medical schools regarding AIG or CAT itself, which reveals the need to explore both topics in depth.

In addition to the changes in how tests are administered, we must also consider the linguistic context in which tests are used nowadays. For the bulk of the twentieth century, exams were solely written and administered at the local or regional level in the native language of the culture in which they were set. However, considering globalization and the increasing relevance of computers in training/assessment situations, educational tests must now be designed in different languages to suit students from different countries (Gierl et al., 2021). The conventional procedures used to develop and translate test items are time-consuming, costly, and laborious. This is a parallel issue to the item development burden currently experienced in other sectors of the educational setting. More modern AIG item modeling approaches could also be employed to generate items in multiple languages, as different languages may be added as an additional layer in the item model

to enable multilingual AIG (Gierl et al., 2021; Gierl et al., 2016). However, it is important to note that little research on the topic has been conducted to date (Gierl et al., 2016).

It is also worth noting that AIG does not replace (and does not wish to) test development specialists (Gierl et al., 2016). On the contrary, it seeks to assist these professionals in concentrating their efforts on the task of identifying, organizing, and adapting the content required to generate more test items (Gierl, Shin, et al., 2022; Gierl et al., 2016). AIG should not be viewed as a threat, but as an effective and contemporary tool that can considerably assist specialists with their duties, freeing them of time-consuming and costly tasks. It is also crucial to note that, while AIG can tackle the problems described, it should not be regarded as a panacea for all assessment problems, and the same holds true for all other AI tools out there that are surfacing in the educational setting (Swiecki et al., 2022).

Last but not least, it is critical to determine how content specialists (and students) feel about using AIG. Although the production of test questions based on computer algorithms appears enticing, it is important to understand the perspectives of individuals dealing directly with the procedure's output. If medical schools (and other educational institutions) decide to implement AIG systems to produce test items, students and specialists should be informed about how the AIG procedure is conducted and, most importantly, assured that the items, rationales and respective options created are safe and reliable for their assessment.

Conclusion

The worldwide spread of the internet and the increasing demand for new and large-scale assessment opportunities have pushed traditional forms of assessment toward modern computer-based evaluations. AIG is a next-gen item construction method in applied psychometrics that has sparked interest due to its promise of efficiently providing high-quality test items. Throughout this Thesis, we have found evidence that sustains AIG as an efficient, valid, and easy-to-learn assessment procedure capable of automatically generating quality test items and alleviating the item development burden that affects medical education assessment, which in itself is a promise for better assessment practices and significant cost savings for medical education institutions.

References

1. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
2. Arendasy, M., & 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
3. Attali, Y., Runge, A., LaFlair, G. T., Yancey, K., Goodwin, S., Park, Y., & von Davier, A. A. (2022). The interactive reading task: Transformer-based automatic item generation. *Frontiers in Artificial Intelligence, 5*, 903077. https://doi.org/10.3389/frai.2022.903077
4. Babcock, B., & Brunnert, K. (2023). A Simple Sorting Selection Method for Picking Results from an Automatic Item Generator. *Journal of Applied Testing Technology*. Retrieved from: http://jattjournal.net/index.php/atp/article/view/172787
5. Bejar, I., Lawless, R., Morley, M., Wagner, M., Bennett, R., & Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. The Journal of Technology, Learning and Assessment, 2(3). Retrieved from: https://ejournals.bc.edu/index.php/jtla/article/view/1663
6. 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
7. Chowdhury, 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
8. Circi, R., Hicks, J., & Sikali, E. (2023). Automatic item generation: foundations and machine learning-based approaches for assessments. *Frontiers in Education*. 8:858273. https://doi.org/10.3389/feduc.2023.858273
9. Embretson, S., & Kingston, N. (2018). Automatic Item Generation: A More Efficient Process for Developing Mathematics Achievement Items? *Journal of Educational Measurement, 55*(1), 112–131. https://doi.org/10.1111/jedm.12166
10. Falcão, F., Costa, P., & Pêgo, J. (2022). Feasibility assurance: a review of automatic item generation in medical assessment. *Advances in Health Sciences Education*, 1–21. https://doi.org/10.1007/s10459-022-10092-z

11. Fu, Y., Choe, E. , Lim, H., & Choi, J. (2022). An Evaluation of Automatic Item Generation: A Case Study of Weak Theory Approach. *Educational Measurement: Issues and Practice*. https://doi.org/10.1111/emip.12529
12. 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
13. Gierl, M., & 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
14. Gierl, M., Lai, H., Fung, K., & Zheng, B. (2015). Using technology-enhanced processes to generate test items in multiple languages. In F. Drasgow (Ed.), *Technology and Testing: Improving Educational and Psychological Measurement* (pp. 109–126). Routledge. https://doi.org/10.4324/9781315871493-5
15. Gierl, M., & 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
16. 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
17. Gierl, M., Lai, H., & Tanygin, V. (2021). Advanced Methods in Automatic Item Generation. Routledge. https://doi.org/10.4324/9781003025634
18. Gierl, M., Shin, J., Firoozi, T., & Lai, H. (2022). Using Content Coding and Automatic Item Generation to Improve Test Security. *Frontiers in Education*, 1–14. https://doi.org/10.3389/feduc.2022.853578/full
19. Gierl, M., Swygert, K., Matovinovic, D., Kulesher, A., & Lai, H. (2022). Three Sources of Validation Evidence Needed to Evaluate the Quality of Generated Test Items for Medical Licensure. *Teaching and Learning in Medicine*, 1–11. https://doi.org/10.1080/10401334.2022.2119569
20. Goh, P.-S., & Sandars, J. (2020). A vision of the use of technology in medical education after the COVID-19 pandemic. *MedEdPublish, 9*(49), 49. https://doi.org/10.15694/mep.2020.000049.1

21. Götz, F. M., Maertens, R., Loomba, S., & van der Linden, S. (2023). Let the algorithm speak: How to use neural networks for automatic item generation in psychological scale development. *Psychological Methods*. https://doi.org/10.1037/met0000540
22. Guze, P. A. (2015). Using technology to meet the challenges of medical education. *Transactions of the American Clinical and Climatological Association*, 126, 260.
23. Harrison, P., 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
24. Harrison, P., & Müllensiefen, D. (2018). Development and validation of the computerised adaptive beat alignment test (CA-BAT). *Scientific Reports, 8*(1), 1–19. https://doi.org/10.1038/s41598-018-30318-8
25. Hilburg, R., Patel, N., Ambruso, S., Biewald, M., & Farouk, S. (2020). Medical education during the coronavirus disease-2019 pandemic: learning from a distance. *Advances in Chronic Kidney Disease, 27*(5), 412–417. https://doi.org/10.1053/j.ackd.2020.05.017
26. Huh, S. (2022). Application of computer-based testing in the Korean Medical Licensing Examination, the emergence of the metaverse in medical education, journal metrics and statistics, and appreciation to reviewers and volunteers. *Journal of Educational Evaluation for Health Professions, 19*. https://doi.org/10.3352/jeehp.2022.19.2
27. Kornegay, J. G., Kraut, A., Manthey, D., Omron, R., Caretta-Weyer, H., Kuhn, G., Martin, S., & Yarris, L. (2017). Feedback in medical education: a critical appraisal. *AEM Education and Training, 1*(2), 98–109. https://doi.org/10.1002/aet2.10024
28. Kosh, A., Simpson, M., 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
29. Kucharski, S., Damnik, G., Stahr, F., & Braun, I. (2023). Revision of the AIG Software Toolkit: A Contribute to More User Friendliness and Algorithmic Efficiency. *Proceedings of the 15th International Conference on Computer Supported Education (CSEDU 2023)*, 410–417. https://doi.org/10.5220/0011982000003470
30. Lai, H., Alves, C., & Gierl, M. (2009). Using Automatic Item Generation to Address Item Demands for CAT. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*, 1–16. Retrieved from: www.psych.umn.edu/psylabs/CATCentral

31. Lai, H., Gierl, M., Byrne, B., Spielman, A., & Waldschmidt, D. (2016). 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.0022-0337.2016.80.3.tb06090.x
32. Lai, H, Gierl, M., Touchie, C., Pugh, D., Boulais, A-P., & De Champlain, A. (2016). Using Automatic Item Generation to Improve the Quality of MCQ Distractors. *Teaching and Learning in Medicine, 28*(2), 166–173. https://doi.org/10.1080/10401334.2016.1146608
33. Lee, J., Wu, A. S., Li, D., & Kulasegaram, K. M. (2021). Artificial intelligence in undergraduate medical education: a scoping review. *Academic Medicine, 96*(11S), S62–S70. https://doi.org/10.1097/ACM.0000000000004291
34. Leslie, T., & Gierl, M. J. (2023). Using Automatic Item Generation to Create Multiple-Choice Questions for Pharmacy Assessment. *American Journal of Pharmaceutical Education*, 100081. https://doi.org/10.1016/j.ajpe.2023.100081
35. Liu, M. (2023). Applying Natural Language Processing to Assessment. In T. Kantrowitz, D. Reynolds, & J. Scott (Eds.), *Talent Assessment: Embracing Innovation and Mitigating Risk in the Digital Age* (p. 66). Oxford University Press. https://doi.org/10.1093/oso/9780197611050.003.0005
36. Luecht, R. (2012). Automatic Item Generation for Computerized Adaptive Testing. In M. Gierl & T. Haladyna (Eds.), *Automatic Item Generation: Theory and Practice* (pp. 196–216). Routledge.
37. Padilla, J., & Benítez, I. (2014). Validity evidence based on response processes. *Psicothema, 26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259
38. Pei, L., & Wu, H. (2019). Does online learning work better than offline learning in undergraduate medical education? A systematic review and meta-analysis. *Medical Education Online, 24*(1), 1666538. https://doi.org/10.1080/10872981.2019.1666538
39. Pereira, D., Falcão, F., Nunes, A., Santos, N., Costa, P., & Pêgo, J. (2023). Designing and building OSCEBot ® for virtual OSCE - Performance evaluation. *Medical Education Online, 28*(1), 2228550. https://doi.org/10.1080/10872981.2023.2228550
40. 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

41. Royal, K. D., Hedgpeth, M.-W., Jeon, T., & Colford, C. M. (2018). Automated item generation: The future of medical education assessment? *European Medical Journal Innov, 2*(1), 88–93. https://doi.org/10.33590/emjinnov/10313113.
42. Seo, D. G., & Choi, J. (2018). Post-hoc simulation study of computerized adaptive testing for the Korean Medical Licensing Examination. *Journal of Educational Evaluation for Health Professions, 15*, 14. https://doi.org/10.3352/jeehp.2018.15.14
43. Setiawan, H., Hidayah, I., & Kusumawardani, S. (2022). Automatic Item Generation with Reading Passages: A Systematic Literature Review. *8th International Conference on Education and Technology (ICET)*, 250–255. https://doi.org/10.1109/ICET56879.2022.9990738
44. Shin, J., Guo, Q., & Gierl, M. (2019). Multiple-choice item distractor development using topic modeling approaches. *Frontiers in Psychology, 10*, 825. https://doi.org/10.3389/fpsyg.2019.00825
45. Stochl, J., Böhnke, J., Pickett, K., & Croudace, T. (2016). Computerized adaptive testing of population psychological distress: simulation-based evaluation of GHQ-30. *Social Psychiatry and Psychiatric Epidemiology, 51*(6), 895–906. https://doi.org/10.1007/s00127-015-1157-4
46. Susanti, Y., Tokunaga, T., & Nishikawa, H. (2020). Integrating automatic question generation with computerised adaptive test. *Research and Practice in Technology Enhanced Learning, 15*(1). https://doi.org/10.1186/s41039-020-00132-w
47. Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075
48. Weiss, D. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development, 37*(2), 70–84. https://doi.org/10.1080/07481756.2004.11909751
49. Xinxin, Z. (2019). Using automatic item generation to create content for computerized formative assessment [Doctoral Dissertation, University of Alberta]. University of Alberta. https://doi.org/10.7939/r3-vbt3-6598