Using automatic item generation to create multiple-choice test items

Mark J Gierl,1,2 Hollis Lai2 & Simon R Turner1,2

CONTEXT Many tests of medical knowledge, from the undergraduate level to the level of certification and licensure, contain multiple-choice items. Although these are efficient in measuring examinees’ knowledge and skills across diverse content areas, multiple-choice items are time-consuming and expensive to create. Changes in student assessment brought about by new forms of computer-based testing have created the demand for large numbers of multiple-choice items. Our current approaches to item development cannot meet this demand.

METHODS We present a methodology for developing multiple-choice items based on automatic item generation (AIG) concepts and procedures. We describe a three-stage approach to AIG and we illustrate this approach by generating multiple-choice items for a medical licensure test in the content area of surgery.

RESULTS To generate multiple-choice items, our method requires a three-stage process. Firstly, a cognitive model is created by content specialists. Secondly, item models are developed using the content from the cognitive model. Thirdly, items are generated from the item models using computer software. Using this methodology, we generated 1248 multiple-choice items from one item model.

CONCLUSIONS Automatic item generation is a process that involves using models to generate items using computer technology. With our method, content specialists identify and structure the content for the test items, and computer technology systematically combines the content to generate new test items. By combining these outcomes, items can be generated automatically.

Medical Education 2012; 46: 757–765
doi:10.1111/j.1365-2923.2012.04289.x

Discuss ideas arising from this article at www.mededuc.com ‘discuss’

1Department of Surgery, University of Alberta, Edmonton, Alberta, Canada
2Centre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, Alberta, Canada

Correspondence: Dr Mark J Gierl, Professor of Educational Psychology, Canada Research Chair in Educational Measurement, Centre for Research in Applied Measurement and Evaluation, 6-110 Education North, University of Alberta, Edmonton, Alberta T6G 2G3, Canada. Tel: 601 780 492 2996; Fax: 601 780 492 1318; E-mail: mark.gierl@ualberta.ca
INTRODUCTION

Multiple-choice items provide the foundation for many assessments used in medical education. From small-scale classroom assessments to large-scale licensure examinations, the multiple-choice item format is used extensively to measure examinees’ knowledge and skills across diverse medical content areas and domains of specialisation. Clear guidelines exist for creating multiple-choice items.1–3 Content specialists draw on these guidelines to develop items because they represent well-established standards in practice (e.g. they include a central idea in the stem; they have one correct answer). Content specialists must also draw on their experience, expertise and judgement to create each new multiple-choice item.1 However, despite the availability of item-writing guidelines and the willingness of content specialists, creating multiple-choice items for medical tests is still a challenging task. One aspect of the challenge stems from the need to develop items across diverse and often specialised content areas. Another aspect is rooted in the complexity of the development task itself, in which cognitive problem-solving skill and content knowledge must be expressed by the content specialist in the multiple-choice item format, and then replicated over and over again to produce new items.

Addressing the challenges posed by multiple-choice item development is feasible for small-scale applications, like classroom assessments, in which only a small number of items are required (e.g. a 50-item, end-of-unit test). However, this task becomes more prohibitive in the context of a large-scale examination in which hundreds or even thousands of items are needed. For instance, many medical licensure examinations are now administered using computerised adaptive testing systems. A computerised adaptive test (CAT) is a paperless examination that implements a prescribed method of selecting and administering items, scoring examinees’ responses, and estimating examinees’ ability. The adaptive process of selecting new items based on examinees’ responses to previously administered items is continued until a stopping rule is satisfied. The benefits of the CAT are well documented.4,5 For instance, a CAT permits testing on demand, thereby allowing examinations to be administered continually throughout the year. A CAT can be administered via the Internet and therefore at multiple sites simultaneously. A CAT also decreases overall testing time by administering items appropriate to each examinee’s ability level, thereby shortening the examination without losing measurement precision compared with a paper-and-pencil test. However, this form of testing has one important cost: CATs require large banks containing thousands of multiple-choice items. The use of large banks permits continuous testing while minimising item exposure so that test security can be maintained. Breithaupt et al.6 estimated, for example, that the number of items required for a high-stakes licensure examination using a relatively small 40-item CAT with two administrations per year was, at minimum, 2000.

To create the items needed for a medical licensure examination, extensive multiple-choice item development is required. Item development is a time-consuming process because each item is individually written, reviewed and revised by a content specialist. It is also an expensive process. Rudner7 claimed that the cost of developing a single item for a high-stakes licensure examination ranged from US$1500 to US$2000. Given this estimate, it is easy to see how the costs of item development quickly escalate. If we combine the Breithaupt et al.6 item size estimate with the Rudner7 cost per item estimate, we find that the bank for a 40-item CAT licensure examination requires an expenditure of US$3 000 000–4 000 000. In short, medical testing agencies that adopt new and innovative computer-based testing methods, of which CAT is just one example, are faced with the formidable task of creating thousands of new and expensive multiple-choice items. One way to develop these items is to hire more content specialists. Although this approach is permissible, it adds to the cost of an already expensive item development process. An alternative approach to developing multiple-choice items for medical examinations is automatic item generation (AIG).8–11

Automatic item generation is a process of using models to generate items using computer technology. Unlike the current approach to item development, in which content specialists create each item individually, AIG promotes a generative process where an item model, which is a prototypical representation of a multiple-choice test item or question, is used.12–14 With this model, the item is decomposed into elements that are varied during the generation process to produce new items. Because the elements in the item model are the only variables manipulated, content specialists have the role of defining those elements that yield feasible items, and computer technology has the role of systematically combining the elements to produce new items. Much like the current approach to multiple-choice item development, AIG items must adhere to the highest standards of quality through the use of rigorous
guidelines and item development practices. Items developed using AIG are also created to measure examinees’ knowledge and problem-solving skills in specific and diverse medical content areas. However, AIG has the added benefit of scalability, whereby large numbers of multiple-choice items can be produced very quickly compared with current item development practices. The purpose of this study is to present a methodology for AIG unique to medical education testing that can be used to generate large numbers of multiple-choice items. Although our illustration is in the context of a medical licensure test for the content area of surgery, it should be noted that our method can be applied to any medical testing situation in which large numbers of items are required or desired.

METHODS AND RESULTS

To generate multiple-choice items for medical tests, a three-stage process is required. In the first stage, a cognitive model structure is created. In the second stage, the item models are developed using the content from the cognitive model. In the third stage, items are generated from the item models using computer software. We will describe and illustrate each stage.

Stage 1: creating a cognitive model structure

The content included in a medical examination is identified by specialists who balance the demands of the medical profession with the knowledge and skills students should acquire during their medical school training. A wealth of literature now exists on how medical knowledge can be conceptualised. Norman et al. summarised these conceptual frameworks as causal (basic mechanisms of medicine), analytic (relationships of specific symptoms and features with specific conditions) and experiential (prior case experiences). The application of this knowledge is an important area of study. Decision trees, clinical problem analysis, and clinical reasoning education are examples of methods that can be used to study how knowledge is applied in practice to make medical diagnoses.

Just as frameworks are needed to study the structure and application of medical knowledge, frameworks are required to generate medical test items. In the current study, we present a framework for structuring the knowledge, skills and content required to make a medical diagnosis. The knowledge, skills and content are used, in turn, to generate items. We call our framework a cognitive model structure for AIG because it highlights and helps organise the knowledge, skills and content required to make a medical diagnosis. The model also organises this cognitive- and content-specific information into a coherent whole, thereby presenting a succinct yet structured organisation of the content relationships and sources of information used in formulating medical diagnoses (Fig. 1).

Two content specialists, both of whom were experienced medical examination item writers and practising surgeons, were asked to describe the knowledge and clinical reasoning skills required to solve items on a medical licensure examination in surgery. The knowledge and skills were identified in an inductive manner using a verbal reporting method, meaning that the content specialists were given an existing multiple-choice item and asked to identify and describe the key information that would be used to solve the item. To produce this information, they solved the item together by verbally articulating the knowledge, skills and content they would draw upon to generate the correct solution. The content specialists began by identifying the problem (i.e. post-operative fever) specific to the existing test item. They also identified different scenarios related to that problem (i.e. urinary tract infection, atelectasis, wound infection, pneumonia, deep vein thrombosis, deep space infection). These outcomes are presented in the top panel in Fig. 1.

Using a set of related scenarios that hold the underlying problem in common, the content specialists then identified sources of information required to diagnose the scenarios. Again, this information was identified by verbally describing the type of information needed to solve the item. Sources of information were deemed by the content specialists to be either generic (i.e. patient demographics, physical examination) or case-specific (i.e. timing of fever, type of surgery). These outcomes are presented in the middle panel in Fig. 1.

Finally, the content specialists identified features within each information source as they verbalised their solution to the item (e.g. guarding and rebound, fever, calf tenderness, abdominal examination, and examination of the wound are features of physical examination). Each feature, in turn, contains two nested components. The first nested component for a feature is the element. Elements contain content specific to each feature that can be manipulated for item generation. For the Guarding and Rebound feature of the Physical Examination source of information, Yes or No Guarding and Rebound serve as the element. The
second nested component for a feature is the constraint. Each element is constrained by the scenarios specific to this problem. For instance, Deep Space Infection (DSI) is associated with Guarding and Rebound (i.e. DSI: Present, for the Guarding and Rebound feature). These outcomes are presented in the bottom panel in Fig. 1.

In its entirety, Fig. 1 serves as a cognitive model structure for AIG. It provides a structure articulated by content specialists using a verbal reporting method for identifying the contextual relationships between domain-specific content and how associated information is assembled to make a medical diagnosis for complications related to postoperative fever. The modelling approach in our example can be generalised to other medical diagnoses for the purpose of generating items. A generic cognitive model structure for AIG is presented in Fig. 2.

**Stage 2: creating item models**

In the next stage of our method, an item model is created so that the content in the cognitive model structure can be cast as a multiple-choice item. Recall, an item model is a prototypical representation of a test item that guides the generation process. The item model includes all of the information necessary to generate a set of unique items. A sample item model is presented in Fig. 3. The item model contains the stem, which specifies the context, content, item and/or question the examinee is required to answer. The stem also highlights the sources of information that will be manipulated for item generation, as well as the location of those sources in the model itself. For example, in Fig. 3, the sources of information include Patient Demographic, Timing of Fever, Type of Surgery and Physical Examination. These four sources serve as the variables in the item model.
Next, the elements in the item model are specified. These are the variables that will be manipulated to generate items in the Postoperative Fever example. The elements are extracted directly from the cognitive model structure described in Stage 1. In our example, these elements are the sources of information that direct which features (i.e., elements, subject to the constraints) will be inserted into the item model to create new test items. For example, the elements Very Likely, Less Likely and Unlikely for the features Gastrectomy, Left Hemicolectomy, Laparoscopic Cholecystectomy, Right Hemicolecetomy and Appendectomy will be inserted in the Type of Surgery source of information to create new items.

The item model also contains the options. The options are the components of the model that specify the correct alternative and one or more incorrect alternatives or distracters. For the Postoperative Fever example, the possible treatment options are Antibiotics, Mobilise, Reopen Wound, Anticoagulation, and Percutaneous Drainage.
Stage 3: item generation

Finally, to generate items from the Stage 1 and 2 models, software is used to assemble all permissible combinations of elements, subject to their constraints, to create new items. One piece of software designed for this purpose is called Item GeneratOR or IGOR. IGOR is a JAVA-based program that will run on any desktop computer. IGOR is research software available from the first author, by request. IGOR is an ‘iterator’, meaning that it contains algorithms for combining elements, subject to the logical constraints in the item model, to generate items. Given item model input, IGOR first generates a set of all possible items. Then, IGOR checks the constraints for each item and removes all illogical element combinations. The outcome is a set of generated items and their corresponding answer keys. The generated items are saved in HTML or in a Word document format in a database. Figure 4 provides a summary of the outcomes of Stages 1–3 that are required for item generation.

Using the AIG process presented in Fig. 4, IGOR generated 1248 different multiple-choice items for diagnosing complications with postoperative fever. Table 1 contains a random sample of four generated items, in which the correct answer is marked with an asterisk. The 1248 items generated in this example required 6 hours to produce across our three-stage method: Stage 1 (3 hours); Stage 2 (2 hours), and Stage 3 (1 hour).

DISCUSSION

The number of assessment points in the typical experience of a medical student is large. Often, these assessments include multiple-choice items. This item format is used extensively in medical education to evaluate examinees’ knowledge and skills across diverse content areas in an efficient and economical manner. However, multiple-choice items are challenging to develop. They require content specialists to implicitly structure the knowledge, skills and content required to solve a medical problem and to express it in the form of a test item. They must also use their interpretation of professional guidelines, as well as their experience, expertise and judgement, to maintain the quality of that item. Because interpretation, experience and judgement play such important roles, multiple-choice item development has sometimes been characterised as an ‘art’.

The purpose of the current study is to offer a new methodology for developing content for multiple-choice examinations using AIG concepts and procedures. Automatic item generation is the process of
creating models to generate items using computer technology. We described and illustrated a method for AIG that yields large numbers of items for medical licensure testing using the example Postoperative Fever. We presented a three-stage process, in which the content needed to generate items is first identified and structured, the item models are then created and, finally, items are generated using computer software. From a single item model, 1248 unique items were generated for our Postoperative Fever example.

Automatic item generation could, therefore, be characterised as representing a shift away from the ‘art’ of item development, in which multiple-choice items are created solely from expertise, experience and judgement, toward a new ‘science’ of item development. This new science is characterised by combining the knowledge and skill of content specialists with the combinatoric and algorithmic power of the computer. However, it is important to note that the characterisation of AIG as an item development ‘science’ does not diminish the role of content specialists. Instead, it helps to focus their role on the important task of identifying, organising and evaluating the content needed to develop the stem and the options; the role of the content specialist in AIG is critical to the creative task in a manner that we typically associate with the art of test development in that the content specialist identifies the knowledge and skills required to solve medical problems, casts the knowledge, skills and content into a cognitive model, designs meaningful item models, and organises the content required to produce the stem as well as the plausible options. This creative task relies on the judgement, expertise and experience of content specialists. The role of computer technology in AIG is critical to the generative task that we often associate with the science of modern computing in that it systematically merges large numbers of elements in each item model. By combining the outcomes of the content-based creative task and the technology-based gener-

| Table 1 | A set of multiple-choice items generated for measuring diagnoses of complications related to postoperative fever |
|-----------------------------------------------|
| **1** A 34-year-old woman has an appendectomy. On postoperative day 6 she has a temperature of 38.5 °C. Physical examination reveals tenderness in the abdominal region with guarding and rebound. Which one of the following is the best next step? |
| (a) Mobilise |
| (b) Antibiotics |
| (c) Reopen the wound |
| (d) Percutaneous drainage* |
| **2** A 46-year-old man is admitted to hospital for an appendectomy. On postoperative day 4 he has a temperature of 38.5 °C. Physical examination reveals tenderness in the abdominal region with guarding and rebound. Which one of the following is the best next step? |
| (a) Mobilise |
| (b) Anticoagulation |
| (c) Reopen the wound |
| (d) Percutaneous drainage* |
| **3** A 54-year-old woman has a laparoscopic cholecystectomy. On postoperative day 3 she has a temperature of 38.5 °C. Physical examination reveals a red and tender wound and calf tenderness. Which one of the following is the best next step? |
| (a) Mobilise |
| (b) Antibiotics |
| (c) Anticoagulation* |
| (d) Reopen the wound |
| **4** A 62-year-old man is admitted to hospital for a laparoscopic cholecystectomy. On postoperative day 1 he has a temperature of 38.5 °C. Physical examination reveals no other findings. Which one of the following is the best next step? |
| (a) Mobilise* |
| (b) Antibiotics |
| (c) Reopen the wound |
| (d) Percutaneous drainage |

* Correct option
ative task, a new composite science of AIG becomes possible.

Limitations and directions for future research

Evidence to support the reliability and validity of test items can be garnered in many ways. For example, the outcomes from an item analysis (e.g. item difficulty, item discrimination, inter-item correlations) provide evidence about the psychometric characteristics of items that can support their reliable and valid use. In the current study, the development process itself was the focal point for producing reliable and valid test items. That is, content specialists combine their expertise, experience and judgement with test development guidelines and standards of practice to create items that yield a reliable and valid probe of students’ medical knowledge and skills. Our AIG methodology uses a process similar to the more traditional approach, except that the knowledge, skills and content used by specialists are now specified overtly in the form of a cognitive model. This model can, and should, be scrutinised carefully. In our study, two content specialists verbalised their strategies for solving a medical test item. The knowledge, skills and content they identified were then organised, structured and used to create a cognitive model for AIG (Fig. 1). This model highlights and coordinates the cognitive- and content-specific information used in formulating medical diagnoses.

However, the cognitive model must still be evaluated to ensure that it provides an accurate account of the diagnostic process required to identify a particular problem (i.e. postoperative fever) in a specific content area (i.e. surgery). The reliability and validity of the generated items are therefore limited to the cognitive model produced by our two content specialists because our study focused only on the AIG methodology. No attempt was made to validate their model or to determine its generalisability. However, before these items are used on an operational test, a validation step, in which empirical evidence is collected to ensure that the model is accurate and generalisable, is essential. Two methods that could be used to collect additional evidence on the veracity and generalisability of the cognitive model include protocol and verbal analysis drawing on a larger sample of content specialists who would solve a more diverse sample of test items. General theories of cognition and medical expertise, as they pertain to the knowledge structures and cognitive processes used in medical diagnoses, could also be used to evaluate our cognitive model, thereby providing even stronger evidence to support the reliability and validity of the generated items.

To summarise, AIG serves as a new, technology-enhanced method for item development that may help to address one of the most pressing and challenging assessment issues facing many medical educators: the rapid and efficient production of large numbers of high-quality, content-specific, medical multiple-choice items.

Contributors: MJG and HL made substantial contributions to the study conception and design, and to the acquisition, analysis and interpretation of data, and drafted the article. MJG, HL and SRT made substantial contributions to the acquisition, analysis and interpretation of data, and to the critical revision of the article. All authors approved the final manuscript for publication.

Acknowledgements: the authors would like to thank the Medical Council of Canada (MCC) for its support of this research. However, the authors are wholly responsible for the methods, procedures and interpretations expressed in this paper, which do not necessarily reflect the views of the MCC.

Funding: funding for this research was provided to the first author by the Social Sciences and Humanities Research Council of Canada, Ottawa, Ontario, Canada (430-2011-0011).

Conflicts of interest: none.

Ethical approval: this study was approved by the Research Ethics Board of the University of Alberta.

REFERENCES


Automatic generation of test items


Received 8 December 2011; editorial comments to authors 7 February 2012; accepted for publication 11 March 2012