Applied Psychological Measurement
2018, Vol. 42(1) 42–57
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0146621617726788
journals.sagepub.com/home/apm
SAGE

Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing

Mark J. Gierl[1] and Hollis Lai[1]

Abstract

Computerized testing provides many benefits to support formative assessment. However, the advent of computerized formative testing has also raised formidable new challenges, particularly in the area of item development. Large numbers of diverse, high-quality test items are required because items are continuously administered to students. Hence, hundreds of items are needed to develop the banks necessary for computerized formative testing. One promising approach that may be used to address this test development challenge is automatic item generation. Automatic item generation is a relatively new but rapidly evolving research area where cognitive and psychometric modeling practices are used to produce items with the aid of computer technology. The purpose of this study is to describe a new method for generating *both* the items and the rationales required to solve the items to produce the required feedback for computerized formative testing. The method for rationale generation is demonstrated and evaluated in the medical education domain.

Keywords

item development, test development, automatic item generation, technology and assessment

For much of the 20th century, educational measurement was synonymous with large-scale testing intended, almost exclusively, to satisfy demands for accountability. These tests were taken by all students within a specific context, they were administered under standardized conditions, and they produced results that guided decisions often described as ''high stakes.'' In other words, educational testing was focused on summative assessment. Summative assessments are administered at the end of a unit or sequence of instruction to provide a succinct measure of educational attainment at one point in time for the purpose of supporting decision makers.

But in the 1990s, a noteworthy shift began to occur in educational measurement. Rather than focusing solely on summative assessment, researchers and practitioners started to develop and evaluate testing procedures that yielded explicit evidence to help teachers monitor their instruction and to help students improve how and what they learn. That is, educational testing began

[1] University of Alberta, Edmonton, Canada

Corresponding Author:
Mark J. Gierl, Faculty of Education, University of Alberta, 6-110 Education North, Edmonton, Alberta, Canada T6G 2G5.
Email: mark.gierl@ualberta.ca

to focus on formative assessment (see, for example, Black & Wiliam, 1998; Sadler, 1989). Formative assessment is a process used by teachers and students during instruction to produce feedback required to adjust teaching and improve learning so students can better achieve the intended outcomes of instruction (Council of Chief State School Officers, 2008; Wylie & Lyon, 2012). Feedback has maximum value when it yields specific information in a timely manner that can direct instructional decisions designed to help each student acquire knowledge and skills more effectively. Outcomes from empirical research demonstrate that formative feedback can enhance teacher effectiveness and produce student achievement gains ranging from one half to one full course grade, reflecting effect sizes of 0.4 to 0.7 or higher (Kluger & DeNisi, 1996; see also Black & Wiliam, 1998, 2010; Hattie & Timperley, 2007; Hwang & Chang, 2011; Popham, 2011; Shute, 2008).

Many different assessment strategies can yield feedback to guide the types of adjustments that improve teaching and learning. In the current study, the authors describe and illustrate how computerized formative testing (also described as e-assessment for learning; see Ras, Whitelock, & Katz, 2016) can serve as an assessment strategy to help students learn. While the outcomes from computerized formative testing can also be used to adjust teaching, this study's focus will be on formative assessment for students. *Computerized formative testing* is a type of technology-enhanced assessment that provides students with evidence so they can monitor and, if need be, modify their learning activities.

Computerize Formative Testing: A Case Study

To provide a context for this research, a short case study is presented. A third-year medical student is progressing through her 8-week surgery rotation. She will be assessed at many points in her program. For instance, medical students are evaluated throughout each rotation, at the end of each rotation, and in a final high-stakes qualifying examination required to enter a postgraduate residency program. She is in the fourth week of her surgery rotation. The student is implementing different learning strategies to improve her surgical knowledge base. She is currently focusing on how to diagnose serious problems that result from chest trauma. She has a smartphone that permits her to access a computerized testing system. The system has a bank of 1,000 test items that measure specific learning outcomes in surgery. The content in this bank is based on the learning outcomes of the clerkship, the required clinical encounters within the program, and the objectives of the licensing body of the medical profession. The student studies the content in her surgery textbook, and she draws on her experiences during the surgery rotation. To evaluate how well she understands the surgical content she is studying, the student uses a computerized testing system to conduct self-assessments by solving different types of chest trauma test items. She can access this testing system using her smartphone at any time. When she solves an item correctly, the solution is immediately presented as the keyed option. When she solves the item incorrectly, the solution is immediately presented as the keyed option along with the rationale required to solve the item. Using this strategy, the student can identify what knowledge she possesses (correct solutions), what knowledge she lacks (incorrect responses), why she lacks specific knowledge for a given item (rationales), and, possibly, where her knowledge is weak (for example, if the rationale consistently contains the blood pressure variable, then the student may have a misconception about the relationship between chest trauma and blood pressure). Based on the feedback from the rationales, the student intends to adjust her learning strategies to focus on specific topics within the area of chest trauma or more general topics such as blood pressure and surgery. Once the strategy has been implemented, the student can use the computerized testing system to, again, evaluate her surgical knowledge. This iterative, self-directed approach continues until the student is confident with her knowledge of the surgical learning outcomes.

Requirements for Computerized Formative Testing

Computerized formative testing, as presented in the case study, has five requirements. First, it must permit students to evaluate themselves. Second, it must be available to students on demand. Third, it must provide students with immediate feedback. Fourth, it must permit students to access large numbers of diverse but content-specific test items that are developed to measure specific learning objectives across different instructional topics and units. Fifth, it must provide students with specific feedback by presenting the solution and rationale required to solve each test item. The purpose of this feedback is to allow students to adjust their learning strategies to better achieve the intended outcomes of instruction. Recall that feedback has maximum value when it yields specific information in a timely manner that can direct students' learning decisions and strategies.

The first three requirements of computerized formative testing—self assessment, on-demand, and immediate feedback—are related to the availability of technology through the use of a computerized testing system. On-line test administration made possible through the rapid expansion of Internet access has created the foundation necessary for the widespread use of computerized testing. Computerized testing systems are now readily available from companies who offer a broad assortment of products and services ranging from proprietary computer-based testing systems (e.g., Pearson-Vue) to open-source computer-based testing software (e.g., TAO testing from OAT). Tests from a computerized system can be accessed with different types of devices ranging from mobile technologies such as smartphones to standalone desktop computers typically found in testing centers. Regardless of the device that is used, these systems are all capable of providing students with on-demand, computerized testing that yield immediate feedback. Because of these important benefits, the broad adoption of computerized testing at both the K-12 and postsecondary levels of education is well underway.

Purpose of the Current Study

The last two requirements of computerized formative testing—large numbers of content-specific test items and the rationales required to solve these items—are related to the availability of content for use with the testing system. These requirements are far more challenging to address. Large numbers of items are required to create the banks necessary for continuous test administration. A bank, which is a repository of items, must be initially developed and then constantly replenished to ensure that students receive a continuous supply of unique, content-specific, items anchored to detailed learning outcomes. The traditional approach used to create the content for these banks relies on a method where a subject-matter expert (SME) creates each test item individually (Gierl & Lai, 2013; Haladyna & Rodriguez, 2013). Traditional item writing is an iterative process where SMEs use their experiences and expertise to produce new test items. Once the items are created, they are then edited, reviewed, and revised until they meet the required standards of quality (Lane, Raymond, & Haladyna, 2016; Perie & Huff, 2016). Unfortunately, this traditional item development process is time-consuming and expensive. Hence, the first challenge is to create large item banks in a manner that is efficient and economical. One possible way to address this challenge is with *automatic item generation* (AIG; Gierl & Haladyna, 2013; Gierl & Lai, 2016a; Irvine & Kyllonen, 2002). AIG is the process of using models to generate items with the aid of computer technology. AIG is a relatively new but rapidly evolving research area where the outcomes from cognitive theories, computer technologies, and psychometric practices are used to create large numbers of test items.

The second challenge is to create student feedback within the bank of test items. The rationales for each test item must be available so students can receive feedback on how to solve each testing task. Rationales provide explanations of how to solve a test item based on the features or characteristics of the problem-solving task. The use of rationales for producing formative feedback has maximum value when it yields specific information in a timely manner. Outcomes from empirical research demonstrate that formative feedback can enhance teacher effectiveness and produce significant student achievement gains (Black & Wiliam, 1998, 2010; Hattie & Timperley, 2007; Kluger & DeNisi, 1996; Popham, 2011; Shute, 2008). Using a traditional item writing approach, rationales are created by SMEs specific to each item. But because AIG produces large numbers of generated items, a correspondingly large number of rationales are also needed. The purpose of this study is to expand the existing AIG framework to address the content development challenge of creating the rationale for each generated item. To begin, the authors describe how items can be generated. They use an example of complex problem solving drawn from the domain of surgical education, as was first introduced in the case study. Then, they introduce and demonstrate a new method for creating rationales that can be incorporated into the item generation process. Finally, they evaluate the quality of generated rationales by presenting the results from an expert content review.

Automatic Item Generation: An Overview

AIG is the process of using models to generate items with the use of computer technology. Gierl and Lai (2013, 2016a) described a three-step approach for AIG. In Step 1, the content for item generation is identified by SMEs. SMEs are content specialists who have extensive experience with the students, with the learning environment and, often, with the test development process. They organize and structure the content required for item generation using a cognitive model. A *cognitive model for AIG* is a representation that highlights the knowledge, skills, and abilities required to solve a problem in a specific domain (Gierl, Lai, & Turner, 2012). In Step 2, an item model is developed to specify where the content from the cognitive model must be placed to generate new items. An *item model* is a template that specifies the features in an item that can be manipulated (Bejar et al., 2003; LaDuca, Staples, Templeton, & Holzman, 1986). Item models specify which parts of the assessment task can be manipulated for item generation. In Step 3, computer-based algorithms place the cognitive model content specified in Step 1 into the item model developed in Step 2. Gierl, Zhou, and Alves (2008) demonstrated the use of the IGOR software system to conduct the content assembly task. IGOR is a JAVA-based program designed to assemble test items using the items models from all combinations of content specified in the cognitive model.

Step 1: Identifying Content for Item Generation Using Cognitive Modeling

Cognitive models are used to generate test items. A cognitive model for AIG highlights the knowledge, skills, and abilities required to solve a problem in a specific domain (Gierl et al., 2012). This model also organizes the cognitive- and content-specific information into a structured representation of how the SME expects that students will think about and solve questions on tests. It contains three types of data: the problem and associated scenarios, sources of information, and features. These three types of data are specified as separate panels in the cognitive model (see Figure 1). The top panel identifies the problem and its associated scenarios. The middle panel specifies the relevant sources of information. Sources of information can be

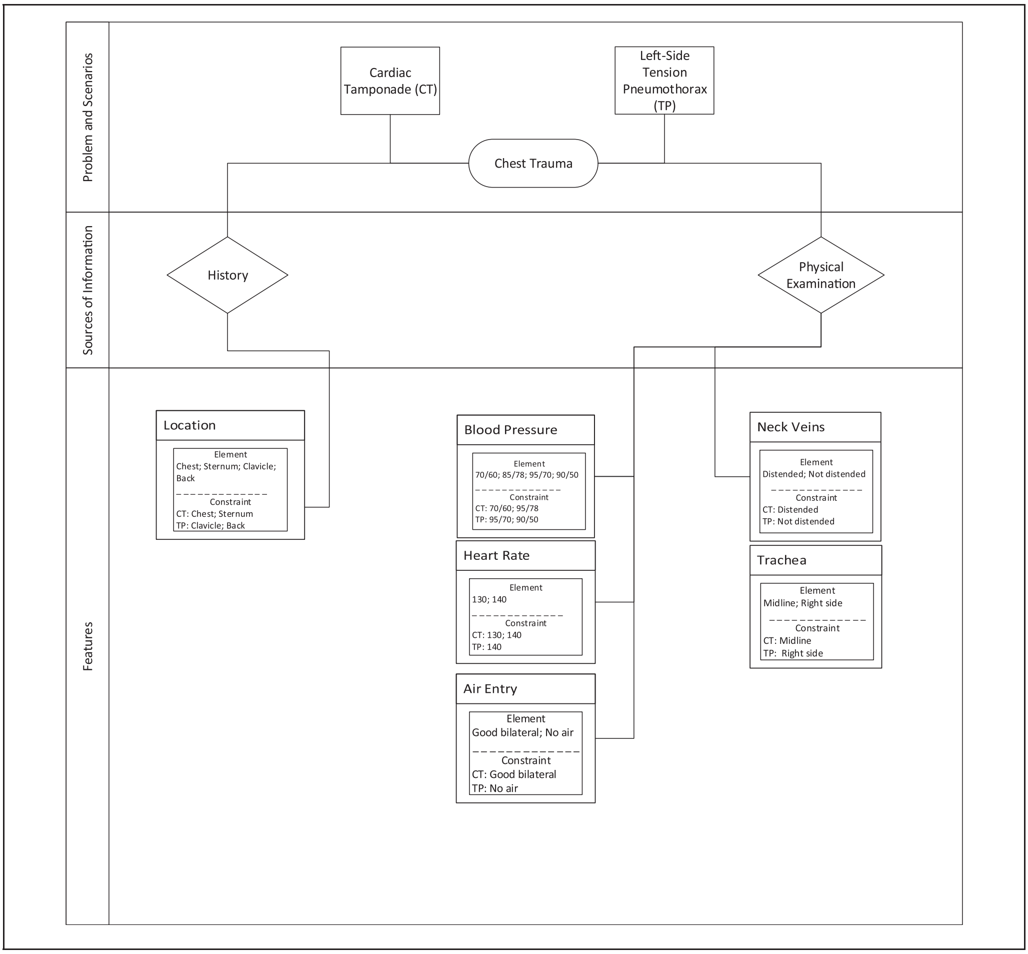

Figure 1. A cognitive model for chest trauma.

specific to a particular problem or generic thereby applying to many types of problems. The bottom panel highlights the salient features, which includes the elements and constraints. *Elements* contain content specific to each feature that can be manipulated for item generation. *Constraints* serve as restrictions that must be applied during the content assembly task in Step 3 (see below) to ensure that meaningful items are generated. Item generation can therefore be considered a form of constraint logic programming (Frühwirth & Abdennadher, 2003).

To identify the content in the cognitive model, SMEs are required to describe the knowledge, content, and reasoning skills required to create a problem-solving task. The knowledge and skills specified in a cognitive model are identified using an inductive process by asking the SMEs to review a parent item and then to identify and describe information that could be used to create new items. *Parent items* are high-quality items which produce strong item analysis results drawn from existing operational tests. Parent items highlight the underlying structure of the cognitive model thereby providing a point-of-reference for the development that occurs in Step 1. By referencing the parent items, SMEs are responsible for identifying, organizing, and evaluating the content required to create the cognitive model. This model development process relies on human judgment acquired through AIG training. Cognitive model development is also a subjective practice because a model is an expression of the SMEs' understanding of knowledge and skill within a specific content area (Haladyna & Rodriguez, 2013; Lane et al., 2016; Schmeiser & Welch, 2006).

One approach to item generation focuses on the relations among key features in the cognitive model (Gierl & Lai, 2016b). The *key features cognitive model* is used when the features or characteristics of a task are systematically combined to produce meaningful outcomes across the item feature set. The systematic combination of permissible features is defined by the constraints specified in the feature panel (i.e., the bottom panel) of the cognitive model. The key features cognitive model is most suitable for measuring the students' ability to assemble and apply key features within a domain as well as to solve problems using these key features.

Figure 1 contains a cognitive model for AIG that describes how to diagnose a serious problem after chest trauma has occurred. The key features cognitive model shown in Figure 1 outlines the knowledge and skills that produce test items required to make medical diagnostic inferences to identify the problem outlined in this stem of the item model. Two medical SME created the model in Figure 1. The SMEs were third-year surgical residents. In addition to advanced medical training, both SMEs acquired research and assessment training by completing the Royal College Clinical Investigators Program. The top panel identifies the problem and its associated scenarios. The SMEs began by identifying the medical problem (i.e., chest trauma). Two types of diagnoses were associated with this problem, cardiac tamponade (CT) and left-side tension pneumothorax (TP). The middle panel specifies the relevant sources of information. In this example, two sources were identified (history and physical examination). The bottom panel highlights the salient features, which includes the elements and constraints. In this study's example, six features (i.e., location, blood pressure, heart rate, air entry, neck veins, trachea) were identified. Each feature specifies two nested components. The first nested component for a feature is the element. For example, the location feature contains four elements: chest, sternum, clavicle, and back. The second nested component for a feature is the constraint. For instance, the location feature is constrained to chest and sternum when the diagnosis is cardiac tamponade (i.e., top left side of the features box contains the coding CT: Chest; Sternum). Again, additional features could easily be added to create a more complex model.

Step 2: Structuring Content for Automatic Item Generation

In Step 2, an item model is developed to specify where the content from the cognitive model must be placed to generate new items. An item model is a template that specifies the features in an item that can be manipulated to generate new items (Bejar et al., 2003; LaDuca et al., 1986). In the current study, the authors focus on the multiple-choice, selected-response item type because of its prevalence and popularity in medical testing. Downing and Haladyna (2009) claimed, for instance, that selected-response items, typified by the multiple-choice question, serve as the most useful written testing format for evaluating cognitive knowledge in the health sciences. A multiple-choice item includes the stem and the options. The stem contains the content or question the student is required to answer. The options include a set of alternative answers with one correct option and one or more incorrect options.

When developing multiple-choice items, the correct option can readily be created for each stem. The incorrect options or distracters, on the contrary, are often much more challenging to create. Distracters are typically designed from a list of plausible but incorrect alternatives linked to misconceptions or errors in thinking, reasoning, and problem solving (Rodriguez & Haladyna, 2013, p. 106). Using a traditional item writing approach, distracters must also be created by SMEs specific to each item. But because AIG produces large numbers of generated items, a correspondingly large number of distracters are needed. Lai et al. (2016) recently described a method for generating the incorrect options using a method called *systematic distracter generation*. This method involves specifying the key features related to errors and misconceptions to generate plausible but incorrect options. The process of modeling incorrect options is complex because as the stem is changing in a generative system, a large number of constraints are required to ensure the incorrect options contain erroneous but plausible information that is matched to the content in each of the generated stems. Moreover, three or four incorrect options must be produced for each generated item when four- and five-option multiple-choice items, respectively, are generated.

To systematically generate distracters, the key features related to the correct option are also used to create distracters. For example, when the SMEs were asked to identify the key features required to diagnose cardiac tamponade, they listed blood pressure is 70/60, heart rate is 130, air entry is good bilaterally with muffled heart sounds, and trachea is midline (see Figure 1). This set of key features is associated with the correct option cardiac tamponade. But this list of key features can also be associated with incorrect options. Sternal fracture, for example, is also associated with a blood pressure of 70/60, heart rate of 130, good air entry bilaterally with muffled heart sounds, and midline trachea. As a result, sternal fracture serves as one plausible but erroneous distracter because it shares some but not all of the key features required to diagnose cardiac tamponade. Using the logic for systematic distracter generation, a list of plausible but erroneous answers that served as incorrect options was identified and used as the distracters for the generated items in the current study.

Table 1 is the item model that contains the content specified in the chest trauma cognitive model. Table 1 lists all of the content required for the stem in the item model (top—stem), for the elements of each feature within the two source of information (middle—elements), and for the options in each generated item, including the correct answer and the distracters (bottom—options). By combining the logic specified in Figure 1 with the content presented in Table 1, we can illustrate how the key features of the item are manipulated to produce the generated items. For this example, the elements and the constraints in Figure 1 define the key features that produce the correct response. If we follow the diagnosis of cardiac tamponade in Figure 1, we can see that this diagnosis is correct when location is chest or sternum (i.e., Location Feature CT: Chest, Sternum), when blood pressure is 70/60 or 95/78 (i.e., Blood Pressure Feature CT: 70/60, 95/78), when heart rate is 130 or 140 (i.e., Heart Rate Feature CT: 130, 140), when air entry is good bilateral (i.e., Air Entry Feature CT: Good Bilateral), when neck veins are distended (i.e., Neck Veins Feature CT: Distended), and when trachea is midline (i.e., Trachea Feature CT: Midline). In other words, the student must be able to recognize that any combination of the elements in these six key features will produce the correct diagnosis of cardiac tamponade. The set of key features that lead to the diagnosis of left-side tension pneumothorax can be identified using the same approach (i.e., the element and constraint combinations for TP in Figure 1). Students must be able to differentiate the key features across these two diagnoses to select the correct answer.

Table 1 also contains a list of plausible but incorrect options. In this study's example, six incorrect options are included in the item model (myocardial infarction, pulmonary contusion, splenic injury, atelectasis, subclavian artery laceration, sternal fracture). Each of these incorrect options was identified by the SMEs for inclusion in the item model because they share some but not all of the key features required to identify the correct diagnosis of cardiac tamponade or left-side tension pneumothorax.

Table 1. An Item Model for Generating Chest Trauma Test Items.

Stem
A 68-kg, 50-year-old security guard was **[Location]** during an altercation with another man. He is immediately brought to the hospital where his vitals are as follows: **[Blood Pressure]** and **[Heart Rate]**. He has **[Air Entry]**. His neck veins are **[Neck Veins]** and his trachea is **[Trachea]**. What is the most likely diagnosis?
Elements
Location: 1. stabbed in the left lower chest 2. kicked multiple times in the sternum 3. stabbed below the clavicle on the left side 4. stabbed multiple times in the left back Blood Pressure: 1. blood pressure is 70/60 2. blood pressure is 85/78 3. blood pressure is 95/70 4. blood pressure is 90/50 Heart Rate: 1. heart rate is 130 2. heart rate is 140 Air Entry: 1. good air entry bilaterally, with muffled heart sounds 2. no air entry on left side, with normal heart sounds Neck Veins: 1. distended 2. not distended Trachea: 1. midline 2. deviated to the right side
Options
1. Cardiac tamponade* (Correct Option) 2. Left-side tension pneumothorax* (Correct Option) 3. Myocardial infarction 4. Pulmonary contusion 5. Splenic injury 6. Atelectasis 7. Subclavian artery laceration 8. Sternal fracture

Step 3: Generating Items Using Computer Algorithms

In Step 3, computer-based algorithms place the cognitive model content specified in Step 1 into the item model developed in Step 2, subject to the constrained specified in Figure 1. Content assembly is conducted with a computer algorithm because it is a complex constraint logic programming task, particularly when large problems with many sources of information that contain a lot of different features are specified in the cognitive model. Gierl et al. (2008) created the IGOR software as a generalized item generation system to conduct the content assembly task. IGOR is a JAVA-based program designed to assemble test items using the items models from all combinations of elements specified in the cognitive model. That is, once the content is specified in Step 1 and the template is created in Step 2, IGOR systematically places the content into the template, subject to the constraints, to produce new items as part of Step 3. IGOR generated 48 chest trauma items using the content from the cognitive model in Figure 1 with the item model content in Table 1. This three-step process has also been used to generate items in other content areas including mathematics (Gierl, Lai, Hogan, & Matovinovic, 2015), science (Gierl & Lai, 2016a), dentistry (Lai, Gierl, Byrne, Spielman, & Waldschmidt, 2016), and nonverbal reasoning (Gierl, MacMahon-Ball, Vele, & Lai, 2015).

Creating Rationales as Part of the Generation Process

Feedback can also be created as part of the generative process because the logic used to assemble all of the correct and incorrect options based on the key features can also be used to identify the solution and to generate the rationale for the solution. That is, AIG can be used as a method to produce the items as well as the corresponding rationale for each item. The new method for rationale generation introduced in this study is made possible by expanding the item model in Step 2. In this section, the authors begin by identifying important characteristics of formative feedback. Then, they describe and illustrate a method for generating rationales that possesses these important formative feedback characteristics. Finally, they evaluate the quality of the generated rationales using the results from an expert content review.

Characteristics of Formative Feedback

In 2008, Shute published a comprehensive study in the journal *Review of Educational Research* that analyzed more than 100 articles related to formative feedback (see Shute, 2008). The purpose of her review was to identify the features, functions, interactions, and links between formative feedback and learning but also to consolidate these findings into a set of guidelines for using formative feedback. Shute (2008) defined formative feedback as information communicated to learners so they can modify their behavior to improve learning (p. 154). She identified nine prescriptions in her guidelines for how to use formative feedback (Shute, 2008, Table 2) as well as nine prescriptions for how not to use formative feedback (Shute, 2008, Table 3). Of the 18 prescriptions, the authors identified five that are directly relevant to the development of rationales for a computerized formative test. Generated rationales intended to link formative feedback to learning should have the following characteristics. The feedback is focused on a specific task. Elaborated (i.e., what, how, why) feedback is used when rationales are presented. The rationale is presented as a small, manageable piece of text that is specific, clear, and simple. These five characteristics were used to guide the development of the rationale generation method by specifying how to structure and present student feedback. These five guidelines were also used to evaluate the quality of the generated rationales.

Methodology for Generating Rationales, With Examples

Solutions and rationales can be added to the generation process in the item modeling step by focusing on the key features in the task. An example of a rationale item model is presented in Table 2. This example contains three different types of rationales based on the content in the chest trauma item model presented in Table 1. The content for the rationale methodology was created by the same two medical SME who created the item models presented earlier in this study. The SMEs were asked to describe the knowledge and clinical-reasoning skills required to solve parent items, as described in Step 1 of the AIG method. In addition to solving the parent items, the SMEs were also asked *to specify the key features* required to produce the rationales for the parent items. To produce this information, they solved the parent items together

Table 2. An Item Model for Generating Three Different Types of Chest Trauma Rationales.

Rationales
Rationale 1 (Key Features List): **[Correct Option]** is the most correct option for this patient who was **[Location]** because his blood pressure is **[Blood Pressure]**, his heart rate is **[Heart Rate]**, he has **[Air Entry]**, his neck veins are **[Neck Veins]**, and his trachea is **[Trachea]**. These are the key features of **[Correct Option]**.
Rationale 2 (Key Features Set): **[Correct Option]** is the most correct option as the patient had a blood pressure of **[Blood Pressure]** with a high heart rate after being **[Location]**. The key features in this diagnosis include **[Key Features]**.
Rationale 3 (Key Features Set with Distracter): **[Correct Option]** is the most correct option as the patient had a blood pressure of **[Blood Pressure]** with a high heart rate after being **[Location]**. The key features in this diagnosis are **[Key Features]**. The incorrect options are not plausible because **[Distracter Rationale]**.
Elements
Correct Option: 1. Cardiac tamponade 2. Left-side tension pneumothorax
Location: 1. stabbed in the left lower chest 2. kicked multiple times in the sternum 3. stabbed below the clavicle on the left side 4. stabbed multiple times in the left back
Blood Pressure: 1. blood pressure is 70/60 2. blood pressure is 85/78 3. blood pressure is 95/70 4. blood pressure is 90/50
Heart Rate: 1. heart rate is 130 2. heart rate is 140
Air Entry: 1. good air entry bilaterally, with muffled heart sounds 2. no air entry on left side, with normal heart sounds
Neck Veins: 1. distended 2. not distended
Trachea: 1. midline 2. deviated to the right side
Key Features: 1. distended neck veins, trachea that is midline, good air entry, and limited heart sounds. 2. neck veins that are not distended, trachea that is deviated to the right side, and reduced air entry on the left side.
Distracter Rationale: 1. they all could have been presented without distended neck veins. 2. they would only present with a deviated trachea and they would only present when the patient was kicked in the sternum. 3. they would only present with a deviated trachea and they would only present when the patient had faint heart sounds.

by verbally articulating the knowledge, skills, content, and features they would draw on to produce the correct solution and the rationale for the solution. Using the approach, three different types of rationales were created based on the key features in the item model.

Table 3. The Solutions and Rationales for a Generated Item.

1. A 68-kg, 50-year-old security guard was stabbed in the left lower chest during an altercation with another man. He is immediately brought to the hospital where his vitals are as follows: blood pressure 70/60 and heart rate 130. He has good air entry with muffled heart sounds. His neck veins are distended and his trachea is midline. What is the most likely diagnosis?

A. Atelectasis
B. Splenic injury
C. Sternal fracture
D. Cardiac tamponade*
*correct option

Rationale 1 (Key Features List): Cardiac tamponade is the most correct option for this patient who was stabbed in the left lower chest because his blood pressure is 70/60, his heart rate is 130, he has good air entry, his neck veins are distended, and his trachea is midline. These are the key features of cardiac tamponade.

Rationale 2 (Key Features Set): Cardiac tamponade is the most correct option as the patient had a blood pressure of 70/60 with a high heart rate after being stabbed in the left lower chest. The key features in this diagnosis include distended neck veins, trachea that is midline, good air entry, and limited heart sounds.

Rationale 3 (Key Features Set with Distracter Rationale): Cardiac tamponade is the most correct option as the patient had a blood pressure of 70/60 with a high heart rate after being stabbed in the left lower chest. The key features in this diagnosis are distended neck veins, trachea that is midline, good air entry, and limited heart sounds. The incorrect options are not plausible because they all could have been presented without distended neck veins.

The first rationale is based on a list of all key features that were used in the stem of the item model in Table 1. As a result, it is called a *key features list* rationale. This rationale includes all of the key features—correct option, location, blood pressure, heart rate, air entry, neck veins, trachea—that were used in the stem of the generated items. The text for this rationale is structured as a succinct list for the student that is focused on all of the key features in the problem. This is the simplest type of generated rationale.

The second rationale is based on a set of the key features in the stem of the item model. It is called a *key features set* rationale. The text for this rationale may contain some of the elements from the key features list. In this study's example, the key features include correct option, blood pressure, and location. But the text also contains at least one new element that includes a new set of the key features. Hence, a new variable called ''key features'' must be created for the item model. The text for the new key features element contains content from the neck veins, trachea, and air entry features. This is a more complex type of rationale because it contains existing key features from the model used to generate the items. But it also contains new elements that combine some of the existing key features from the model used to generate the items. The benefit of this type of rationale over a more simplistic list structure is that it permits feedback designed to draw attention to specific combinations of key features. This combination, in turn, may yield a more complete explanation of how to solve the problem. In this study's example, the key features of blood pressure and location are presented in the first sentence of the rationale. The intended effect of this rationale structure is that the student is immediately drawn to the blood pressure and location key features. In fact, any key feature in the cognitive model can be used to highlight or emphasize a specific element required in the rationale explanation. Then, in the second sentence of the rationale, the remaining key features of neck veins, trachea, and air entry are combined into a single element and presented to the student as part of the rationale explanation. Again, any combination of the key features can be included or created to support the explanation in the rationale. For example, the key features can be organized by the sources of information in cognitive model. In this case, the history element (i.e., location) would be presented first in the rationale, followed by the physical examination elements (blood pressure, heart rate, air entry, neck veins, and trachea). Because many different combinations of key features are permissible, many rationales emphasizing different aspects of the problem-solving task can be generated for each item.

The third rationale is based on a set of the key features in the stem of the item model as well as a rationale for the distracters. Hence, it is called *key features set with distracter* rationale. The text for this rationale may contain elements from the key features list (as with the first rationale), it may contain elements from the key features set (as with the second rationale), but it always contains elements that provide an explanation for why the distracters yield the incorrect solution. This is the most complex type of rationale because it provides an explanation for both the correct option and for the incorrect options in each generated item. Producing distracter rationales requires an approach that draws on logic that is comparable to the approach that was used for systematic distracter generation. That is, the common elements among the distracters that produce an incorrect option are presented to the student as the explanation for why the distracters would be erroneous responses. To permit these associations among the elements, each distracter rationale is coded so that it will appear whenever a set of distracters with the same features are used as the incorrect options for the generated items. For instance, the incorrect options of splenic injury, atelectasis, subclavian artery laceration, and sternal fracture are all associated with nondistended neck veins. Therefore, when these options are used as incorrect options for a generated item, the distracter rationale can state ''they all could have been presented without distended neck veins.'' The third type of rationale provides students with the most comprehensive feedback because one explanation is included for the correct option and another explanation is included for the incorrect options. An example of rationale Types 1, 2, and 3 is provided in Table 3. In a computerized formative testing system, the student would only receive one type of rationale in the form of feedback for each item.

Rationale generation has many benefits. For example, it can be integrated into the three-step AIG process described by Gierl and Lai (2013, 2016a). With the addition of a rationale item model in Step 2, the developer can generate both the items and the rationales for the items. The key features that lead to the correct and incorrect options can be explicitly identified and modeled and then used as feedback. Item generation relies on explicit dependencies that exist between the stem, the correct option, and the incorrect options though specific key features. These dependencies can be used to produce the correct response. Rationale generation is also an adaptive method because the content for rationales is selected conditionally using relationships among the key features specified in the stem, the correct option, and the incorrect options. Consequently, the rationales created using this method will be plausible and appropriate for every generated case within an item model. Rationale generation also has drawbacks. Rationale modeling requires that specific information be collected for the all of the correct and incorrect options. This is a time-consuming process because data must be collected from SMEs, these data must be coded to identify the shared features, and then computer programs must be written to assemble the content in the form of a rationale using constraint logic programming. The rationale may include a list of key features or a set of key features. The rationale can focus on the correct option or it can include both the correct option and the distracters. Because of the large number of different outcomes that are possible, rationale generation will add extra time to the three-step AIG process.

Evaluating the Quality of Generated Rationales

The authors also evaluated the quality of the generated rationales. Rationale quality was evaluated using judgments from SMEs where the guidelines from Shute (2008), as described earlier

Table 4. Medical Subject-Matter Expert Rationale Ratings.

	Rationale type		
Characteristic	Key features list	Key features set	Key features set with distracters
Focus feedback on task	3.67	3.78	3.11
Provide elaborated feedback	4.00	3.89	3.56
Present in manageable units	3.44	3.40	3.00
Be specific and clear	3.92	4.00	3.83
Feedback simple and focused	3.76	3.83	3.75

in this study, were used as the basis for scrutinizing the rationales. A four-member panel of medical SMEs was convened to evaluate the three types of rationales presented in this study. The panelists represented four different medical specializations (Surgery, Radiology, Psychiatry, Pediatrics) with an extensive range of professional (14-20 years) experience. The panelists were blind to the development process used to create the items and rationales. They were asked to evaluate the quality of the rationales generated from three different item models using the characteristics from Shute (2008). That is, three different item models were used to generate items. For each generated item, one of the three types of rationales was also generated. The nine generated items and rationales were randomly presented within a single test form to the SMEs. Each SME, working independently, rated each rationale on a 5-point scale ranging from *strongly disagree* to strongly agree, where a higher score indicates more agreement that the item satisfies the feedback characteristic.

The mean score from the evaluation are presented in Table 4. For the first characteristic, focus feedback on the task, the overall mean score was 3.52. The mean rating for each rationale type ranged from a low of 3.11 for key features set with distracters to a high of 3.78 for key features set. For the second characteristic, provide elaborated feedback, the overall mean score was 3.81. The mean rating for each rationale type ranged from a low of 3.56 for key features set with distracters to a high of 4.00 for key features list. For the third characteristic, present feedback in manageable units, the overall mean score was 3.28. The mean rating for each rationale type ranged from a low of 3.00 for key features set with distracters to a high of 3.44 for key features list. For the fourth characteristic, be specific and clear, the overall mean score was 3.92. The mean rating for each rationale type ranged from a low of 3.83 for key features set with distracters to a high of 4.00 for key features set. For the fifth characteristic, the rationale is simple and focused, the overall mean score was 3.78. The mean rating for each rationale type ranged from a low of 3.75 for key features set with distracters to a high of 3.83 for key features set.

Two outcomes are apparent from the results in the item quality review. First, the average ratings for each characteristic and for each rationale type were above 3 indicating that, overall, the SMEs agreed that the three types of rationales satisfied, to some extent, all five characteristics of feedback. Second, the key features set with distracters was consistently rated the lowest for each of the five feedback characteristics relative to the other two rationale types. The key features list had the highest agreement rating for providing elaborated feedback and presenting feedback in manageable units while the key features set had the highest agreement rating for focus feedback on the task, be specific and clear, and provide simple and focused feedback. These results suggest that different types of rationales may be required to meet different feedback characteristics. The results also reveal that the feature set with distracters may be the most challenging type of rationale to create.

Summary and Discussion

Computerized testing offers many important benefits to support and promote key principles in formative assessment, such as on-demand testing with immediate feedback. But the advent of computerized formative testing has also raised daunting challenges in the area of item development. Large numbers of test items are required to implement computerized formative testing because they are continuously administered to students. Automatic item generation is one promising approach that may be used to address this test development challenge. But even if large numbers of items are available, a second challenge is to create the feedback—in the form of rationales—for these newly generated items. The purpose of this study is to expand the AIG framework to address the content development challenge of creating the rationale for each generated item required for computerized formative testing. A new method for rationale generation was demonstrated and evaluated using items from the surgical content area of chest trauma.

Three different types of rationales were created, presented, and evaluated in this study. The first type, called a key feature list rationale, is based on a list of all key features that were used in the stem of the item model. This is the simplest type of feedback. The second type, called key feature set rationale, is based on a set of the key features in the stem of the item model. The text for this rationale may contain some of the variables from the key features list but is also contains at least one new variable that include a set of the key features designed to provide examines with a more complete explanation of how to solve the problem. The third type, key features set with distracter rationale, is based on a set of the key features in the stem of the item model as well as a rationale for the distracters. This is the most complex type of feedback because it provides an explanation for both the correct option and for the incorrect options in each generated item. The results from a validation study on rationale quality indicated that, overall, the three types of rationales satisfied all five feedback characteristics. However, the first two rationales—key feature list and key feature set—had the highest agreement ratings across the five feedback characteristics while the key feature set with distracters consistently had the lowest ratings. These results suggest that different types of rationales may be required to meet different feedback characteristics.

Directions for Future Research

Schmeiser and Welch (2006), in their seminal chapter on test development in the fourth edition of the handbook *Educational Measurement*, claimed that traditional item writing is a standardized process that required iterative refinements. It is often standardized through the use of item writing guidelines (Haladyna & Rodriguez, 2013). These guidelines provide SMEs with information that helps to structure their task and to control for the potentially diverse outcomes that can be produced when different SMEs write items. Guidelines provide a summary of best practices, common mistakes, and general expectations that help ensure that the SMEs have a shared understanding of their tasks and their responsibilities (Lane et al., 2016). Because item writing is complex, large numbers of guidelines must be considered to produce high-quality items across diverse content areas. For example, Haladyna and Rodriguez (2013) presented 65 guidelines for developing selected-response items.

To the authors' knowledge, no guidelines exist for creating rationales within either a traditional item development or AIG framework. In much the same way that SMEs require structure for item writing, SMEs will also require structure for creating rationales to introduce some degree of standardization to this task. It was noted that any combination of the key features can be included in a rationale item model which implies that many different rationales can be generated for each item. Research is therefore needed to develop guidelines for rationale generation. This type of research would focus on identifying and aligning rationale purposes with rationale item models. It would focus on the use of experimental research studies designed to evaluate the effectiveness of different rationale types for improving student learning (cf. Sadler, 1989). It could also be used to outline the time and human resources required to produce different types of rationales.

A second line of research should focus on expanding the item model structure to include different item types. Downing and Haladyna (2009) claimed that selected-response items serve as the most useful written testing format for evaluating cognitive knowledge in the health sciences. But alternative item types can also be developed to measure other cognitive skills, such as reasoning and complex problem solving. The clinical decision-making (CDM) item, for instance, is a commonly used item in the health sciences to evaluate a student's ability to gather data (e.g., obtain medical history), interpret data (e.g., diagnosis a problem), and/or treat a problem (e.g., prescribe medication) within the context of a complex clinical scenario (Page & Bordage, 1995). With a CDM item, the student is first presented with the case and then asked to make a series of key decisions related to the case by answering two to four different items using either the extended multiple-choice (meaning an item that has two or more correct options) or short-answer (student must produce their own answers for each item) format. Alternative item formats and item types can be modeled with the goal of generating new content. The feasibility of producing rationales for these new AIG item models should also be studied.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported with funds from the Social Sciences and Humanities Research Council of Canada, Grant Number 435-2016-0016.

References

Bejar, I. I., Lawless, R. 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), 1-32.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, *5*, 7-74.

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *92*, 81-90.

Council of Chief State School Officers. (2008). *Formative assessment: Examples of practice (FAST)* (A work product initiated and led by Caroline Wylie, ETS, for the Formative Assessment for Students and Teachers [FAST] Collaborative). Washington, DC: Author.

Downing, S., & Haladyna, T. M. (2009). Validity and its threats. In S. M. Downing & R. Yudhowsky (Eds.), *Assessment in the health professions* (pp. 21-55). New York, NY: Routledge.

Frühwirth, T., & Abdennadher, S. (2003). *Essentials of constraint programming*. New York, NY: Springer.

Gierl, M. J., & Haladyna, T. (2013). *Automatic item generation: Theory and practice*. New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*, 36-50.

Gierl, M. J., & Lai, H. (2016a). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 410-429). New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2016b). The role of cognitive models in automatic item generation. In A. Rupp & J. Leighton (Eds.), *Handbook of cognition and assessment: Frameworks, methodologies, and applications* (pp. 124-145). New York, NY: John Wiley.

Gierl, M. J. and Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice*, *35*, 6-20. doi:10.1111/emip.12129

Gierl, M. J., Lai, H., Hogan, J., & Matovinovic, D. (2015). A method for generating test items that are aligned to the Common Core State Standards. *Journal of Applied Testing Technology*, *16*, 1-18.

Gierl, M. J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education*, *46*, 757-765.

Gierl, M. J., MacMahon-Ball, M., Vele, V., & Lai, H. (2015). Method for generating nonverbal reasoning items using n-layer modeling. In E. Ras & D. Joosten-ten Brinke (Eds.), *Proceedings from the 2015 International Computer Assisted Assessment Conference, Communications in Computer and Information Science* (pp. 1-10). New York, NY: Springer. doi:10.1007/978-3-319-27704-2_2

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment*, *72*(2), 1-52.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81-112.

Hwang, G. J., & Chang, H. F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, *56*, 1023-1031.

Irvine, S. H., & Kyllonen, P. C. (2002). *Item generation for test development*. Hillsdale, NJ: Lawrence Erlbaum.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modeling procedures for constructing content-equivalent multiple-choice questions. *Medical Education*, *20*, 53-56.

Lai, H., Gierl, M. J., Byrne, B. E., Spielman, A., & Waldschmidt, D. (2016). Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of Dental Education*, *80*, 339-347.

Lai, H., Gierl, M. J., Touchie, C., Pugh, D., Boulais, A., & De Champlain, A. (2016). Using automatic item generation to improve the quality of MCQ distractors. *Teaching and Learning in Medicine*, *28*, 166-173.

Lane, S., Raymond, M., & Haladyna, R. (2016). Test development process. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 3-18). New York, NY: Routledge.

Page, G., & Bordage, G. (1995). The Medical Council of Canada's key features project: A more valid written examination of clinical decision-making skills. *Academic Medicine*, *70*, 104-110.

Perie, M., & Huff, K. (2016). Determining content and cognitive demands for achievement tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 119-143). New York, NY: Routledge.

Popham, J. (2011). *Transformative assessment in action: An inside look at applying the process*. Alexandria, VA: Association for Supervision & Curriculum Development.

Ras, E., Whitelock, D., & Kalz, M. (2016). The promise and potential of e-assessment for learning. In P. Reimann, S. Bull, M. Kickmeier-Rust, R. Vatrapu, & B. Wasson (Eds.), *Measuring and visualising learning in the information-rich classroom* (pp. 21-40). London, England: Routledge.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119-144.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: National Council on Measurement in Education and American Council on Education.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153-189.

Wylie, C., & Lyon, C. (2012). Formative assessment—Supporting students' learning. *ETS R&D Connections*, *19*, 1-12.