Educational Measurement
ISSUES AND PRACTICE

NCME national council on measurement in education

Educational Measurement: Issues and Practice
Winter 2016, Vol. 35, No. 4, pp. 6–20

A Process for Reviewing and Evaluating Generated Test Items

Mark J. Gierl and Hollis Lai, *University of Alberta*

Testing organization needs large numbers of high-quality items due to the proliferation of alternative test administration methods and modern test designs. But the current demand for items far exceeds the supply. Test items, as they are currently written, evoke a process that is both time-consuming and expensive because each item is written, edited, and reviewed by a subject-matter expert. One promising approach that may address this challenge is with automatic item generation. Automatic item generation combines cognitive and psychometric modeling practices to guide the production of items that are generated with the aid of computer technology. The purpose of this study is to describe and illustrate a process that can be used to review and evaluate the quality of the generated item by focusing on the content and logic specified within the item generation procedure. We illustrate our process using an item development example from mathematics drawn from the Common Core State Standards and from surgical education drawn from the health sciences domain.

Keywords: automatic item generation, item development, technology and assessment

Testing organization requires large numbers of diverse, high-quality items to support innovative new approaches to test delivery and test design. But the demand for items now exceeds the supply. Traditional item development requires a process this is time-consuming and expensive because each item is first written by a subject-matter expert (SME), and then reviewed, edited, and revised by groups of SMEs to ensure that the items yield reliable and valid information. Hence, item writing serves as a critical bottleneck in our current approach to item development and banking. One promising method that may be used to address this problem is automatic item generation (AIG). AIG is a relatively new but rapidly evolving research area where cognitive theory and psychometric practice guide the production of items that are generated with the aid of computer technology. AIG can be used to produce hundreds of items from a single item model. Despite the important benefit of producing large numbers of generated items, content reviews are still required to ensure that these items meet appropriate standards of quality. But it is not feasible to review thousands of new items. In this article, we describe and illustrate a comprehensive but feasible review process that can be used to evaluate generated test items. To begin, we provide some context to help account for our growing item demands.

Mark J. Gierl, Professor of Educational Psychology, University of Alberta, 6-110 Education North, Edmonton, AB, Canada, T6G 2G5; mark.gierl@ualberta.ca. Hollis Lai, Assistant Professor of Dentistry, Director of Assessment, Undergraduate Medical Education, University of Alberta, Edmonton, AB, Canada, T6G 2G5; hollis.lai@ualberta.ca.

Factors That Account for Growing Item Demands

Educational testing is undergoing dramatic changes. These changes can be characterized, in part, by how we deliver and how we design educational assessments. Test delivery marks an important paradigm shift for educational assessment. Simply put, large-scale testing is no longer feasible or desirable when delivered in a paper-based format because printing, scoring, and reporting require tremendous time, effort, and expense. As the demand for more frequent testing escalates, the cost of administering paper-based tests will only continue to increase. One solution that can help curtail some of the test delivery cost while providing important benefits for examinees is to transfer to a computer-based testing (CBT) system. CBT permits testing on demand, thereby allowing examinees to write their exams on a more frequent and flexible schedule. CBT provides examinees with instant feedback. CBT supports the development of multimedia item types that allows test developers to measure more complex performances. Because of these important benefits, the broad adaption and wide-spread transition to CBT is now under way. Dr. Richard Patz, past president of the National Council on Measurement in Education and Chief Measurement Officer at ACT Inc., recently stated:

> If there was a tipping point in the transition from number 2 pencils to digital devices, we seem to have reached it in the realm of large-scale educational assessment. After years of anticipation, hard work, blazed trails, and more than a few difficult lessons learned, the broad adoption of technology in our schools has at last enabled the widespread migration to online testing. (National Council on Measurement in Education, 2015, p. 1)

A second important change, made possible through the use of CBT, is how we design tests. Many examples of innovative test designs are now emerging. For example, CBTs can be used

to identify examinee's cognitive-problem solving strengths and weaknesses (e.g., Hoijtink, Béland, & Vermeulen, 2014; van der Kleij, Eggen, Timmers, & Veldkamp, 2012; van der Kleij, Vermeulen, Schildkamp, & Eggen, 2015). CBTs can be paired with computer "agents" to engage examinees in conversational dialogues (Graesser, Li, & Forsyth, 2014). CBTs can be designed to score examinees' written and spoken responses (Shermis, Burstein, Brew, Higgins, & Zechner, 2016). CBTs can include innovative item types and interactive item response formats that prompt the examinee to respond to a variety of stimuli, and then to provide examinees with feedback on their responses using dynamic multimedia score reports (Sireci & Zenisky, 2016; Zenisky & Hambleton, 2016). One important benefit of coupling CBT with advanced test design is that educators can evaluate examinees more frequently. But this desire to assess examinees more frequently combined with the need to use tests for different purposes also exposes one critical problem in educational testing that must be resolved—our current approach to item development.

CBTs and innovative test designs must be supported by large banks of items because frequent testing means, initially, that large numbers of items are required for delivery, and then, large numbers of items are required for the maintenance of these banks. A bank is a repository of test items. These banks must be initially created with new items and then frequently replenished with new items to ensure that examinees receive a continuous supply of unique, content-specific, items while limiting exposure to maintain test security. For traditional paper-based tests, a relatively small number of items were sufficient because tests were delivered in fixed length forms for a small number of administrations per year. By way of contrast, modern CBTs require large numbers of items because exams are delivered in variable length forms or with multiple forms at many times during the year, possibly using a test design where examinees are provided with detailed feedback on their performance. Breithaupt, Ariel, and Hare (2010) claimed, for instance, that a high-stakes 40-item computer-adaptive test with two administrations per year would require a bank containing, at minimum, 2,000 items. Increased examination frequency, test form diversity, and examinee formative feedback within a computer-based context can only be implemented when large numbers of items are available.

The Challenge Facing Subject Matter Experts

Unfortunately, the large, content-specific, item banks required for modern educational testing are not readily available. Moreover, the means by which large numbers of new items can quickly be developed to satisfy these banking requirements are unclear. The current approach used to create the content for these banks relies on a method where a SME creates each test item individually. Item writing is an iterative process where SMEs use their experiences and expertise to produce new test items. Then, when the items are created, they are edited, reviewed, and revised until they meet the required standards of quality (Lane, Raymond, Haladyna, & Downing, 2016). The SME is responsible for identifying, organizing, and evaluating the content required for this complex and creative development process. This approach relies on human judgment acquired through extensive training and practical experiences. Item writing is also a subjective practice because an item is an expression of the SME's understanding of knowledge and skill within a specific content area. This expression, distinctive to each SME, renders each item unique. As a result, item writing has often been described as an "art" because it relies on the knowledge, experience, and insight of the SME.

Schmeiser and Welch (2006), in their seminal chapter on test development in the 4th edition of the handbook *Educational Measurement*, claimed that traditional item writing is a standardized process that required iterative refinements (see also Lane et al., 2016). Because this process must yield fair and equitable assessment tasks, it is often standardized through the use of item writing guidelines (Haladyna & Rodriguez, 2013). These guidelines, which are often paired with item writing training programs, provide SMEs with information to structure their task. Standardization helps control for the potentially diverse outcomes that can be produced when different SMEs perform the same item writing task. 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. Iterative refinement supports the practice of item writing through the use of a structured and systematic item review. That is, once an item has been written, it is then reviewed to evaluate whether it has met important outcomes described in the guidelines and training materials. Typically, reviews are conducted by committees of SMEs who were not involved in the initial item writing activities. Reviews can focus on a range of standards and objectives related to, for instance, item content (e.g., does the item match the test specifications?), fairness (e.g., does the item elicit construct-irrelevant variance due to subgroup differences?), cognitive complexity (e.g., is the linguistic complexity of the item aligned to grade-level expectations?), and presentation (e.g., is the item grammatically correct?) (Perie & Huff, 2016; Schmeiser & Welch, 2006). The review yields detailed feedback on different standards of item quality that, in turn, can be used to revise and improve the original item. In sum, traditional item writing is the process where SMEs use their judgments, expertise, and experiences to create assessment tasks. Item writing is often anchored to a review process where another group of SMEs evaluates the items using well established standards of quality. These writing and reviewing exercises are standardized and coordinated so that the process can be used to produce items that meet high standards of quality.

Unfortunately, this traditional approach to item development has two noteworthy limitations. First, item writing is time-consuming and expensive because it relies on the item as the unit of analysis (Drasgow, Luecht, & Bennett, 2006). That is, each item in the process is unique, and therefore each item must be individually written, edited, reviewed, and revised. Many different components of item quality can be identified. For example, item quality can be determined by the content alignment of the item to a curriculum, by the difficulty level of the item, by the accuracy of the content in each item, by the use of a clear stem in each item, by the use of an appropriate key, by the plausibility of the distractors of each item, and by the accuracy of the punctuation, capitalization, item structure, typeface, and formatting. Because each element in an item is unique, each component of item quality must be reviewed and, if necessary, revised. This view of an item where every element is unique, both within and across items, was highlighted by Drasgow et al. (2006) when they stated:

> The demand for large numbers of items is challenging to satisfy because the traditional approach to test development uses the item as the fundamental unit of currency. That is, each item is individually hand-crafted—written, reviewed, revised, edited, entered into a computer, and calibrated—as if no other like it had ever been created before.

Because writing and reviewing is conducted by highly trained SMEs, the traditional approach is also expensive. Rudner (2010, p. 157) estimated that the cost of developing *one operational item* for a large-scale test using the traditional item writing approach ranged from US $1,500 to $2,500.

Second, the traditional approach to item writing is challenging to scale in an efficient and economical way. The scalability of the traditional approach is again linked to the item as the unit of analysis. When one item is required, one item is written and reviewed by the SME. When 100 items are required, 100 items must be written and reviewed by the SMEs. Hence, a large numbers of SMEs who can write and review items are needed to scale the development process. Using a traditional item writing approach can most certainly result in an increase in item production when large numbers of SMEs are available. But, as noted earlier, item writing and reviewing is a time-consuming process due to the human effort needed to create, review, edit, and revise large numbers of new items. As a result, building banks using this approach is costly and may even prove to be unsustainable.

These two limitations highlight the importance of an efficient and cost-effective approach to item development. These limitations are also challenging to address using the traditional approach (Haladyna & Rodriguez, 2013; Quellmalz & Pellegrino, 2009; Webb, Gibson, & Forkosh-Baruch, 2013). Consequently, an alternative approach to item development may be needed to meet the demands required to support our contemporary test delivery and design initiatives. One approach that may help address the growing need to produce large numbers of new test items in an efficient and economical manner is through the use of *AIG* (Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002). AIG is the process of using models to generate items using computer technology. Next, we describe a method that could be used to produce large numbers of new test items in both an efficient and cost-effective manner.

Automatic Item Generation

Gierl and Lai (2013, 2016) described a three-step approach to AIG. In Step 1, the content for item generation is identified. SME must identify the content that will be used to produce new items. This content is identified using design principles and guidelines that highlight the knowledge, skills, and abilities required to solve problems and perform tasks in a specific domain. This content must also be organized and structured in a logical manner that can promote item generation. One framework that can be used to organize and structure the content is called a *cognitive model for AIG* (Gierl, Lai, & Turner, 2012). A cognitive model for AIG highlights the knowledge, skills, and abilities required to solve a problem in a specific domain. This model also organizes the cognitive- and content-specific information into a coherent whole, thereby presenting a succinct yet structured representation of how the SMEs, expect that examinees will think about and solve problems.

In Step 2, an item model is developed to specify where the content from the cognitive model must be placed to generate new items. Item models (LaDuca, Staples, Templeton, & Holzman, 1986) specify which parts of the assessment task can be manipulated for item generation. For a selected-response item type, it includes the stem, the options, and the auxiliary information. For a constructed-response item type, only the stem and auxiliary information is required. The stem contains the content or question the examinee is required to answer. The options include a set of alternative answers with one correct option and one or more incorrect options. Auxiliary information includes any supplementary material, such as graphs, tables, figures, or multimedia exhibits that augment the content presented in the stem and/or options.

In Step 3, computer-based algorithms place the cognitive model content specified in Step 1 into the item model developed in Step 2. Assembly is conducted with a computer algorithm because it is often a complex combinatorial task. Different types of software have been written to generate test items (e.g., Gütl, Lankmayr, Weinhofer, & Höfler, 2011; Higgins, 2007; Higgins, Futagi, & Deane, 2005; Singley & Bennett, 2002). Gierl, Zhou, and Alves (2008) proposed the use of the IGOR (which is the acronym for **I**tem **G**enerat**OR**) software system to conduct the 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.

Item development using AIG has two important benefits that help overcome the limitations of the traditional approach. First, AIG is an efficient and cost-effective approach to item creation because it relies on the item model, rather than the item, as the unit of analysis. AIG treats the item model as the fundamental unit of currency where a single model is used to generate many items compared with the traditional item writing approach where the item is treated as the unit of analysis. Because the item model can be reused to yield many test items compared with developing each item individually, quality is easier to monitor and maintain. The item model serves as a template for which the SME only manipulates specific, well defined, elements. The remaining elements are not altered during the AIG process. The view of an item model as a template with both fixed and variable elements contrasts with the more conventional view of a single item where every element is unique, both within and across items.

Second, AIG is a process that is easily scaled, unlike traditional item writing, because the item model is the unit of analysis. AIG permits the SME to create a single item model that, in turn, yields many test items. Hence, AIG is scalable because one item model can generate many test items. When one item is required, one item model is written and reviewed by the SMEs. But when 100 items are required, one item model may be sufficient. As a result, the number of required items is no longer tied to the number of SMEs who can write and review items. Rather, item production is linked to the number of available item models, where a small number of SMEs can create the models that produce large numbers of new items. Gierl et al. (2012) described the development of one medical item model in the area of general surgery that generated 1,248 items. More recently, Gierl, Lai, Hogan, and Matovinovic (2015) reported on the development of 18 different mathematics item models that were created to measure 18 outcomes described in the Common Core State Standards, in Mathematics. The 18 item models generated 109,300 new items, meaning that each model produced, on average, 6,072 items. Because of this unit of analysis shift, the cost per item will decrease because SMEs are producing item models

that yield multiple items rather than producing single unique items. As a result, building banks using SMEs who implement AIG could be an efficient and sustainable approach to item development.

A Two-Part Process for Reviewing and Evaluating Generated Test Items

Despite these important benefits, AIG also raises new challenges in the area of item development, particularly as it relates to the content review. With thousands of newly generated items, the traditional item review is infeasible because SMEs can no longer evaluate each item. If one item model yields 1,248 surgery items or 6,072 math items, then it is not possible to have medical and mathematics SMEs, respectively, review each generated item. One possible approach for evaluating quality is to review a sample of the generated items. That is, use the content and the logic specified in the cognitive model along the task structure formulated in the item model to generate new items. Then, ask the SMEs to review a random sample of the generated items in order to evaluate quality. This approach, where items are generated and then samples are reviewed, is continued iteratively until the samples are judged to be error-free. The error-free sample indicates that the AIG modeling process yields high-quality items. While feasible, this approach is still limited. One limitation is that large numbers of iterations across the item reviews—which is both time-consuming and labor-intensive—may be required to achieve error-free results. A second limitation is that, despite the time and effort required to conduct this type of iterative review, there is no guarantee that all of the generated items are error-free. A sample of generated items could still be produced that contain errors. As a result, it may not be possible for the SME to "sample item review" the AIG models to perfection.

Another possible approach to evaluate the quality of the generated items is to shift the review task from the generated *items* to the generating *models* where the focus becomes reviewing, evaluating, and revising the cognitive and item models described in Steps 1 and 2 of the AIG process. SMEs evaluate the content and the logic specified in the cognitive model along with the task structure formulated in the item model. As with a traditional item review, the model review can be an iterative process conducted by the SMEs. But the purpose is not to review a large number of generated items in order to evaluate quality but rather to review a small number of models that provide the core instructions for the generation task. If the instructions for generation in the models are sound, then outcomes from the generative process (i.e., the individual items) should also be sound. The focus of the content review therefore is on evaluating and refining the AIG models.

Models Required to Generate Test Items

Our process contains two types of reviews. The first review is focused on evaluating the quality of the AIG models. It is conducted by the SMEs who developed these models. AIG contains two types of models. The first is the cognitive model. It links cognitive principles to item generation practices for the purpose of producing new test items. A cognitive model for AIG is intended to highlight the knowledge, skills, competencies, and content required to create test items. Figure 1 contains the components in the cognitive model for AIG. 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 defined for the cognitive model in Figure 1 are identified in an inductive manner by asking the SMEs to review a parent item, and then to identify and describe information that could be used to create new items. A cognitive model guides item generation by specifying the cognitive- and content-related components of problem-solving tasks and by identifying how these components can be assembled to produce new tasks.

Three types of information required to solve the parent item are described. They include the main concept, sources of information, and features. These three types of information are presented as separate panels in the cognitive model for AIG. The top panel identifies the main concept or problem and its associated scenarios. The SME begins by identifying the main concept or problem specific to the parent item. The middle panel specifies the relevant sources of information. Sources of information can be specific to a particular problem or generic thereby applying to many problems. The bottom panel highlights the salient features, which include the elements and constraints. The first component for a feature is the element. Elements contain content specific to each feature that can be manipulated for item generation. The second component for a feature is the constraint. Constraints serve as restrictions that must be applied during the assembly task to ensure that content in the elements is combined in a meaningful way so that useful items can be generated.

The second model used in AIG is the item model. Item models (Bejar, 1996, 2002; Bejar et al., 2003; LaDuca et al., 1986) have been described using different terms and concepts, including schemas (Singley & Bennett, 2002), blueprints (Embretson, 2002), templates (Mislevy & Riconscente, 2006), forms (Hively, Patterson, & Page, 1968), frames (Minsky, 1974), and shells (Haladyna & Shindoll, 1989). Regardless of the term that is used, its purpose remains the same. An item model contains the components in an assessment task that can be manipulated for item generation. These components include the stem, the options, and the auxiliary information. The goal of item generation using the item model is to produce new assessment tasks by manipulating a relatively small number of elements in the model. A general item model structure is presented in Table 1. The first panel is the stem of the generated item. The stem contains different elements. The second panel contains the elements, where each element is associated with a list of content that will be used to create the generated items. The third panel identifies the correct and incorrect options for the generated items. The fourth panel specifies the correct option. Often, the starting point for creating an item model is to begin with a parent item. The parent item highlights the underlying structure that has the potential to be used as an item model thereby providing a point of reference for creating items. Then, an item model is created from the parent by identifying elements that can be manipulated to produce new items.

AIG Models With a Logical Structure

Gierl and Lai (in press) described two different types of cognitive models for AIG. The first type is called a logical structures cognitive model. For this model, item generation is guided by the logical structure of the content specified in the cognitive model. It is used when a specific concept or a set of closely

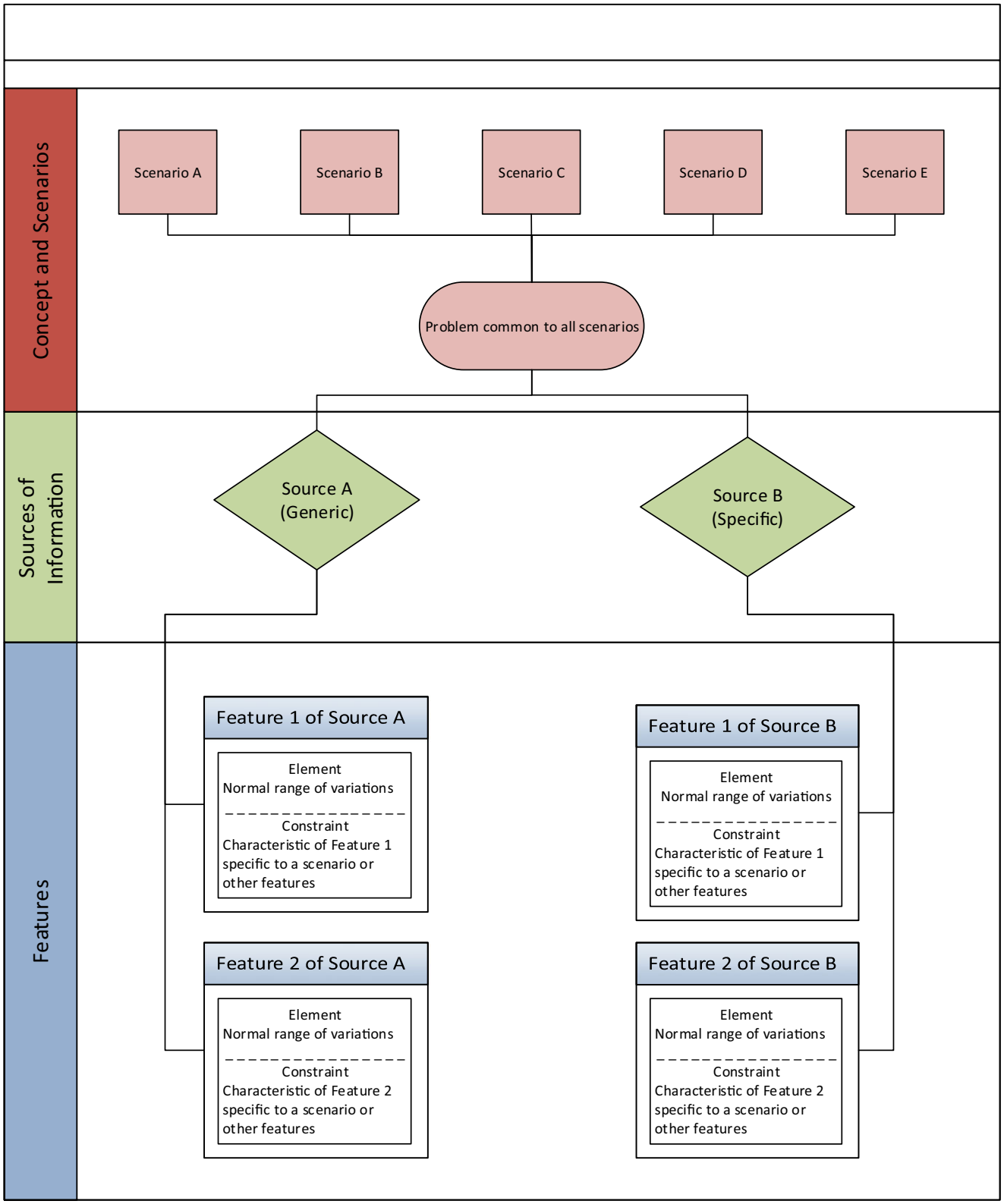

FIGURE 1. A general cognitive model structure for AIG.

related concepts is operationalized as part of the generative process. The logical structures cognitive model is used to measure the examinees' ability to use a concept across a variety of different content representations. The concept is often used to implement a main idea, formula, algorithm, and/or logical outcome. The defining characteristic of this cognitive model is that the content for the item can vary, but the concept remains fixed across the generated items.

The logical structures cognitive model is suitable for many topics found in mathematics and science. To illustrate this

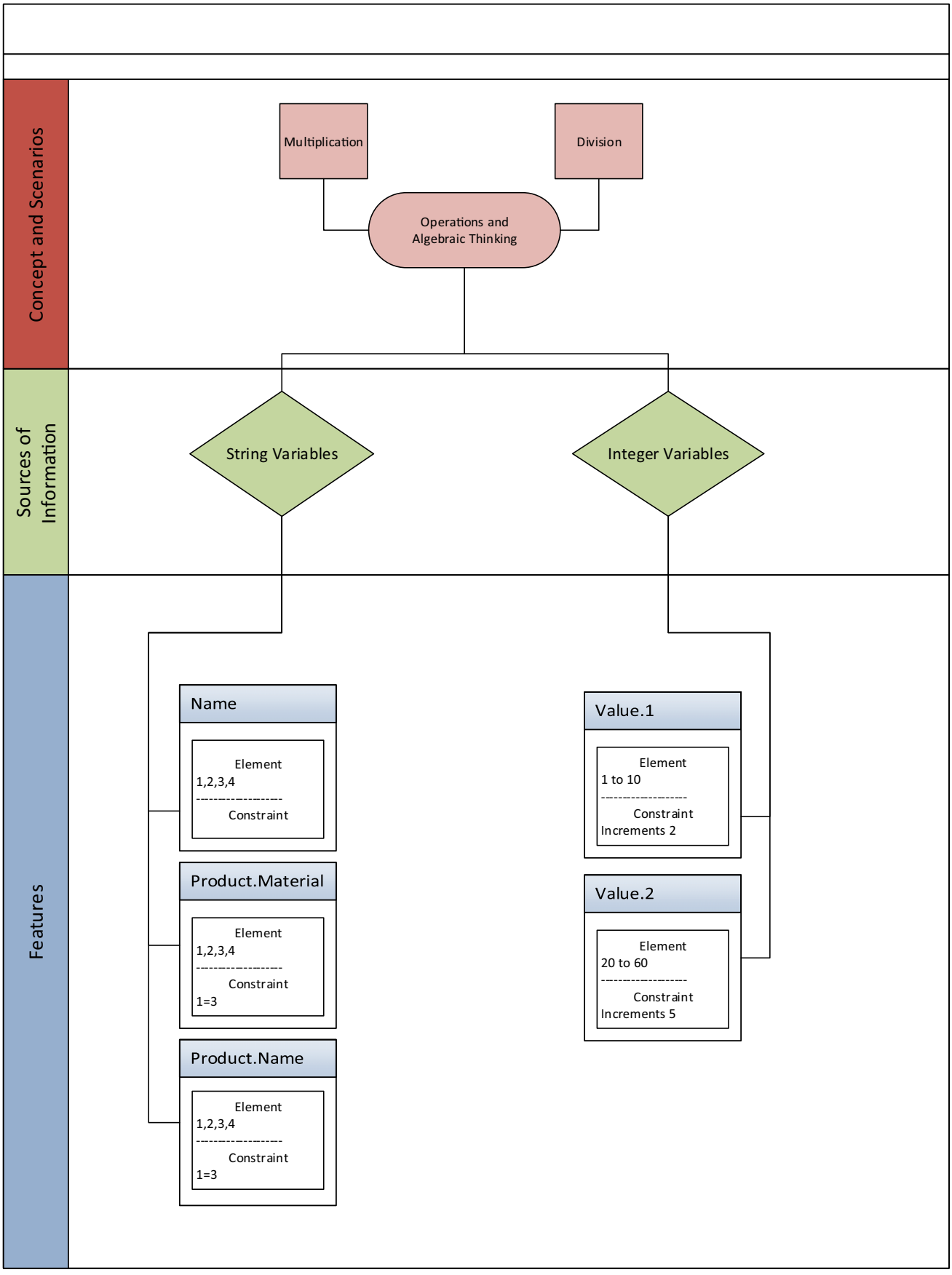

FIGURE 2. A cognitive model for generating algebraic assessment tasks.

Table 1. A General Item Model Structure for AIG

Stem	Text [Element 1] [Element 2] [Element 3] etc.
Elements	**Element 1**: **1**: Variable 1 **2**: Variable 2 **3**: Variable 3 **Etc.** **Element 2**: **1**: Variable 1 **2**: Variable 2 **3**: Variable 3 **Etc.** **Element 3**: **1**: Variable 1 **2**: Variable 2 **3**: Variable 3 **Etc.**
Options	1. Option 1 2. Option 2 3. Option 3 4. Option 4 5. Etc.
Key	Correct Option

Table 2. An Item Model for Generating Algebraic Assessment Tasks

Stem	**[Name] has a total of [Value 1][Product Material][Product Name]. Each [Product Name] will require [Value 2] different [Product Material]. What is the maximum number of [Product Name][Name] can make?**
Elements	**Name**: **1**: Tammy **2**: Lisa **3**: Henry **4**: Michael **Value 1**: 1–10, increments of 2 **Product Material**: **1**: crayons **2**: stamps **3**: felt pens **4**: seeds **5**: ornaments **Product Name**: **1**: paper murals **2**: stamp collections **3**: posters **4**: gardens **5**: trees **Value 2**: 20–60, increments of 5
Options	1. [Total. No. of **Product Material**] / [**Product Material** per **Product Name**] 2. [Total. No. of **Product Material**] + [**Product Material** per **Product Name**] 3. [Total. No. of **Product Material**] - [**Product Material** per **Product Name**] 4. [Total. No. of **Product Material**]
Key	[Total. No. of **Product Material**] / [**Product Material** per **Product Name**]

model, an example of a mathematics concept in operations and algebraic thinking is presented. This concept is drawn from the Operations and Algebraic Thinking domain in the Common Core State Standards in Mathematics. Figure 2 contains the logic specified in the cognitive model required to coordinate the content specified in the item model. Table 2 contains the actual content and the item model used to generate the mathematics items.[1] The top panel in Figure 2 contains the main concept. The main concept measured by this model requires examinees to solve algebraic problems using multiplication and division. The middle panel specifies the sources of information required in the model. In our example, the model has two sources of information. The first source is specified as three string variables related to the context of the problem (i.e., Name, Product Material, and Product Name). The second source is two integer variables related to the numeric operations (i.e., Value 1, Value 2). The bottom panel contains the features. Each feature has elements and constraints. The elements contain the specific information required to generate each item. These specific elements are numbered in the cognitive model (see Figure 2). The numbers correspond to lists of content presented in the item model (see Table 2). For example, the elements 1–4 for the Product Name feature in the cognitive model correspond to 1: to make paper murals, 2: to make stamp collections, 3: to make posters, 4: to make gardens, and 5: to decorate trees in the item model. The constraints are restrictions or rules required for combining the elements to ensure that the content combinations yield feasible outcomes during item generation. For the Product Name feature, as an example, the rule is that the paper murals and posters elements can be used as interchangeable content during item generation. The logic specified in the logical structures cognitive model in Figure 2 and the content in the item model in Table 2 are used to specify how content can be manipulated by describing the rules required to produce the appropriate combinations when the computer-based algorithms generate new items.

In addition to the practical purpose of guiding the item generation algorithms, cognitive models have a second purpose that is more abstract. A cognitive model organizes the cognitive- and content-specific information in a structured representation of a task. This information can also be used to make inferences about how examinees are likely to think about and solve problems. This structured representation can be expressed as the *AIG validation table* (see Table 3; see also Pugh, DeChamplain, Gierl, Lai, & Touchie, 2016). The table is a unique summary for item developers that is only available with AIG. It can be used during a content review to evaluate the structure of the knowledge and skills required to produce the correct and incorrect options for the generated items. It also provides a way to describe and evaluate the target construct for the assessment (Reckase, 2016). For these reasons, we introduce the validation table as a critical new component in the AIG review process. For the logical structures cognitive model, item generation is guided by the logical structure in the cognitive and the content in the item models in Figure 2 and Table 2, respectively. Recall, the logical structures cognitive model is most suitable for measuring the examinees' ability to use a concept across a variety of differ-

Table 3. Validation Table for the Cognitive and Item Models Used to Generate Algebra Assessment Tasks

Stem	Option 1 (Key)	Option 2	Option 3	Option 4
[**Name**] has a total of [**Value 1**][**Product Material**] to make [**Product Name**]. Each [**Product Name**] will require [**Value 2**] different [**Product Material**]. What is the maximum number of [**Product Name**][**Name**] can make?	[Total. No. of **Product Material**] / [**Product Material** per **Product Name**]	[Total. No. of **Product Material**] + [**Product Material** per **Product Name**]	[Total. No. of **Product Material**] - [**Product Material** per **Product Name**]	[Total. No. of **Product Material**]

ent content representations. The concept required to produce the correct option for the mathematics example is [Total. No. of Product Material]/[Product Material per Product Name]. Examinees must be able to implement this algebraic operation across different content representations specified in the Product Material and Product Name variables. Three different incorrect options are also presented. Each incorrect option contains an error or misconception related to the correct option. The first incorrect option uses the correct content but an incorrect operation (i.e., addition rather than division). The second incorrect option used the correct content with an incorrect operation (i.e., subtraction rather than division). The third incorrect option only used the correct content in the numerator and excludes an algebraic operation altogether.

During the first review, the SME is responsible not only for evaluating the content and the logic specified in Table 2 and Figure 2, but also for implementing the content and logic as it is presented in the AIG validation table in Table 3. If the models are correctly specified, then the generated items will reflect the correct combination of the content and logic outlined in AIG models. For the logical structures cognitive model, the SME must verify that the main idea, formula, algorithm, and/or logical outcome are correctly implemented and presented across different types of content. In our example, both correct and incorrect formulas are used to measure algebraic thinking using diverse content in the generated algebra test items. A sample of the generated items is presented in Table 4.

Table 4. A Sample of Generated Algebra Items

12. Lisa has a total of 1,224 crayons to make paper murals. Each paper mural will require 40 different crayons. What is the maximum number of paper murals Lisa can make?

A. 24
B. 30*
C. 40
D. 54

13. Tammy has a total of 1,432 stamps to prepare a stamp collection. Each stamp collection will require 35 different stamps. What is the maximum number of stamp collections Tammy can make?

A. 32
B. 35
C. 40*
D. 72

14. Henry has a total of 1,149 seeds to plan a garden. Each garden will require 60 different seeds. What is the maximum number of gardens Henry can make?

A. 9
B. 19*
C. 28
D. 60

15. Michael has a total of 572 felt pens to make posters. Each poster will require 35 different felt pens. What is the maximum number of posters Michael can make?

A. 12
B. 16*
C. 28
D. 35

*Correct option, where options are listed from the lowest to highest value.

AIG Models With Key Features

The second type of cognitive model for AIG is called a key features model (Gierl & Lai, in press). For this model, item generation is guided by the relationships among the key features specified in the cognitive model. It is used when the attributes or features of a task are systematically combined to produce meaningful outcomes across the item feature set. The use of constraint programming ensures that the relationships among the features yield meaningful items. The key features cognitive model is most suitable for measuring the examinees' ability to assemble and apply key features within a domain as well as to solve problems using these key features. The defining characteristic of this modeling approach is that the content for the item can vary and the key concept also varies across the generated items due to the meaningful combination of features.

The key features cognitive model is suitable for many topics found in the health and medical sciences. To illustrate this model, a medical example from surgery is presented to

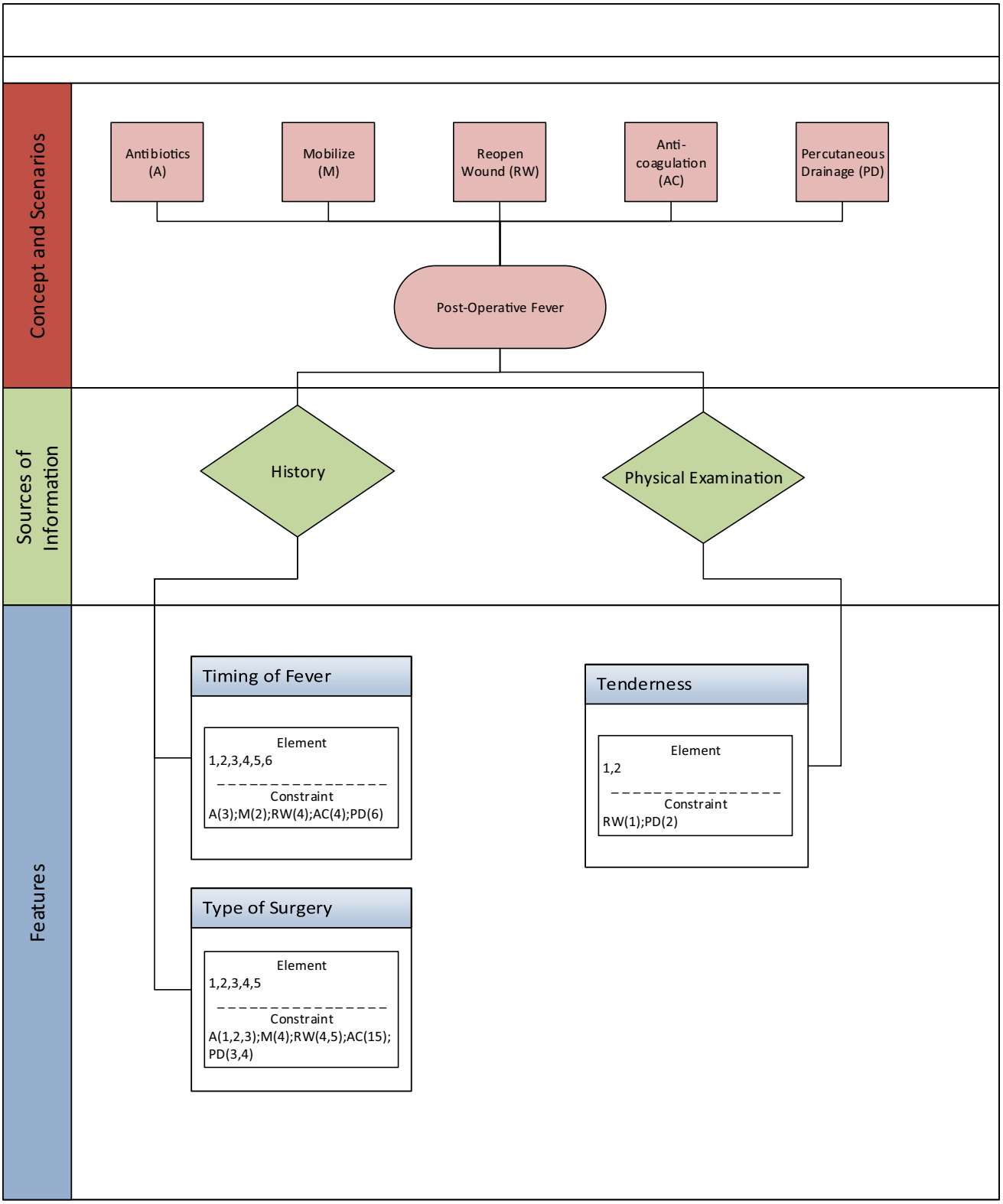

FIGURE 3. A cognitive model for generating post-operative fever assessment tasks.

describe how physicians could manage a patient who is experiencing postoperative fever. Figure 3 contains the logic specified in the cognitive model required to coordinate the content presented in the item model. Table 5 contains the content and the item model used to generate postoperative fever items.[2] The top panel in Figure 3 contains the main concept measured by this model. Examinees are required to respond appropriately when it has been discovered that their patients have developed a fever after a surgery. Five scenarios related to this concept include the use of antibiotics,

Table 5. An Item Model for Generating Post-Operative Fever Assessment Tasks

Stem	**A 40-year-old man has a [Type of Surgery]. On postoperative day [Time of Fever] he has a temperature of 38.5 °C. [Tenderness]. Which one of the following is the best next step?**
Elements	**Timing of Fever**: **1**: 1 **2**: 2 **3**: 3 **4**: 4 **5**: 5 **6**: 6 **Type of Surgery**: **1**: gastrectomy **2**: right hemicolectomy **3**: left hemicolectomy **4**: appendectomy **5**: laparoscopic cholecystectomy **Tenderness**: **1**: Physical examination reveals a red and tender wound at the opening. **2**: Physical examination reveals tenderness in the abdominal region with guarding and rebound. **3**: EMPTY
Options	**1**: Antibiotics (A) **2**: Mobilize (M) **3**: Reopen wound (RW) **4**: Anticoagulation (AC) **5**: Percutaneous drainage (PD)
Key	Variable

Table 6. Validation Table for the Cognitive and Item Models Used to Generate Post-Operative Fever Assessment Tasks

	Correct Options				
Variables	**Antibiotics (A)**	**Mobilize (M)**	**Reopen Wound (RW)**	**Anticoagulation (AC)**	**Percutaneous Drainage (PD)**
Timing of Fever	• Day 3	• Day 2	• Day 4	• Day 4	• Day 6
Type of Surgery	• Gastrectomy • Right Hemicolectomy • Left Hemicolectomy	• Appendectomy	• Appendectomy • Laparoscopic Cholecystectomy	• Gastrectomy • Laparoscopic Cholecystectomy	• Left Hemicolectomy • Appendectomy
Tenderness			• Red and tender wound at the opening		• Reveal tenderness in the abdominal region with guarding and rebound

mobilization, reopen wound, anticoagulation, and percutaneous drainage. The middle panel specifies the sources of information required to manage the post-operative fever. In our example, two sources of information are included: history and physical examination. The bottom panel contains the features. Each feature has elements and constraints. The features include timing of fever, type of surgery, and tenderness. Elements contain content specific to each feature that can be manipulated for item generation. As an example, the type of surgery feature contains five elements: gastrectomy, right hemicolectomy, left hemicolectomy, appendectomy, and laparoscopic cholecystectomy. Each element is constrained by the scenarios specific to this problem. For instance, antibiotics are used to manage post-operative fever when the types of surgery are gastrectomy, right hemicolectomy, or left hemicolectomy (i.e., A [1,2,3] at the bottom left corner in Figure 3). The logic in the cognitive model in Figure 3 and content in the item model in Table 5 are used together to specify how content can be manipulated by describing the rules required to produce the appropriate combinations when the computer-based algorithms generate new items.

As with the logical structures cognitive model, the logic specified in Figure 3 can be combined with the content in Table 5 to produce the AIG validation table (see Table 6). The purpose of the key features model is to measure the examinees' ability to identify, assemble, and apply key features within a domain as well as to solve problems using these key features. Hence, the AIG validation table describes the structure of the knowledge and skills required to produce the

correct and incorrect options for the generated items. Note, however, that the representation of knowledge in the AIG validation table differs markedly between the logical structures and key features cognitive model. For the logical structures cognitive model, the same formula is used to produce the correct and incorrect options. The only difference among the generated items is the content for each generated algebra word problem. For the key features cognitive model, a different combination of key features is required to produce the correct and incorrect options. Hence, each generated item has a different combination of key features. For example, antibiotics is the correct option when timing of fever is day 3, type of surgery is either gastrectomy, right hemicolectomy, or left hemicolectomy, and tenderness is blank. Conversely, mobilization is the correct option when timing of fever is day 2, type of surgery is appendectomy, and tenderness is blank. An examinee who selects mobilization as the correct option when the features include timing of fever day 3, type of surgery as gastrectomy, and tenderness as blank is confusing the management option of mobilization with the use of antibiotics because of a misunderstanding associated with time of fever (day 2 vs. day 3) and/or type of surgery (appendectomy vs. gastrectomy).

During the first review, the SME is responsible for evaluating the content and the logic specified in Table 5 and Figure 3 as well as implementing the content and logic presented in the AIG validation table in Table 6. If the models are correctly specified, then the generated items will reflect the correct combination of the content and logic. For the key features model, the SME must verify that each specific combination of key features will produce the correct option. The incorrect options serve as erroneous but plausible combinations of the key features in the model. In our example, different combinations of correct and incorrect features are used to measure management solutions for patients who develop a fever after their surgery. A sample of the generated items is presented in Table 7.

Table 7. A Sample of Generated Post-Operative Fever Items

A 40-year-old man has an appendectomy. On post-operative day 4, he has a temperature of 38.5 °C. Physical examination reveals a red and tender wound at the opening. Which one of the following is the best next step?
1. Mobilize
2. Antibiotics
3. Anticoagulation
4. Reopen wound *
5.Percutaneous drainage

A 40-year-old man has a laparoscopic cholecystectomy. On post-operative day 4, he has a temperature of 38.5 °C. Which one of the following is the best next step?
1. Mobilize
2. Antibiotics
3. Anticoagulation*
4. Reopen wound
5. Percutaneous drainage

A 40-year-old man has a left hemicolectomy. On post-operative day 6, 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?
1. Mobilize
2. Antibiotics
3. Reopen wound
4. Anticoagulation
5. Percutaneous drainage *

A 40-year-old man has a right hemicolectomy. On post-operative day 3, he has a temperature of 38.5 °C. Which one of the following is the best next step?
1. Mobilize
2. Antibiotics*
3. Reopen wound
4. Anticoagulation
5. Percutaneous drainage

*Correct option.

Rating Rubric for the External Review

In keeping with the claim made by Schmeiser and Welch (2006) that item writing is a standardized process that requires iterative refinements, a second review is included in our process. This review is also focused on evaluating the quality of the AIG models. However, it is conducted independently from the first review by SMEs who did *not* develop the AIG models and it requires the use of a standardized rating rubric for providing specific feedback. Hence, the purpose of the second review is to independently evaluate model quality and provide feedback for model improvement using a standardized rating scale. To demonstrate how this review could be conducted, a sample rubric is presented in Table 8. This rubric is provided as an example. Other rubrics with other measures of model quality could also be used. The rubric in our example contains three components: content, logic, presentation. These three components are evaluated in each model to ensure that the generated items meet these standards of quality. The rubric contains a 4-point rating scale where each of the three components (content, logic, and presentation) is evaluated on the scale. The SME's task is to apply the 4-point scale to each of the three components in order to evaluate the quality of the results in the validation table. The 4-point scale ranges from a low of 1 (*reject*) to a high of 4 (*accept*). The midrange values are (2-Reject, major revisions) and 3 (Accept, minor revisions). A 4-point scale was selected because a succinct but meaningful indicator of quality was required. Based on our discussions with SMEs, psychometricians, and managers, items in the development cycle are often classified into two categories—acceptable or unacceptable. With our rubric, models judged to be acceptable (rating of 4) meet the content, logical, and presentation requirements needed to generate operational test items. Models that require minor revisions (rating of 3) can be quickly modified to meet these content, logical, and presentation standards. Hence, models judged to be either acceptable or requiring minor revisions are seen as acceptable because they require little or no additional work. Conversely, models that require major revisions (rating of 2) warrant a considerable amount of time and extra work. Models that are rejected (rating of 1) have no value for an operational testing program because they fail to meet the content, logical, and presentation requirements. The rubric therefore provides a standardized way to evaluate the AIG models using the judgments from SMEs who are independent of the initial development process. Because the 4-point scale is applied to three components of the model, it also allows the SME to provide the original AIG model developer with feedback on specific areas in the model that require improvement.

Table 8. Rubric for Evaluating the Content and Logic Presented in the AIG Validation Table

Score	Criterion
4—Accept	**Content**: The content is appropriate for measuring the outcome in the model. **Logic**: The content combinations produce an accurate representation of the knowledge and skills required in the domain. **Presentation**: The item model produces tasks that are grammatically and linguistically accurate.
3—Accept, Minor Revisions	**Content**: The majority of the content is appropriate for measuring the outcome in the model. **Logic**: Most of the content combinations produce an accurate representation of the knowledge and skills required in the domain. **Presentation**: The item model, for the most part, produces tasks that are grammatically and linguistically accurate—only minor changes are required.
2—Reject, Major Revisions	**Content**: Much of the content is not appropriate for measuring the outcome in the model. **Logic**: Many of the content combinations do not produce an accurate representation of the knowledge and skills required in the domain. **Presentation**: The item model produces tasks that contain many grammatical and linguistic errors—major changes are required.
1—Reject	**Content**: The content is not appropriate for measuring the outcome in the model. **Logic**: The content combinations do not produce an accurate representation of the knowledge and skills required in the domain. **Presentation**: The item model produces tasks that are grammatically and linguistically incorrect.

Conclusions

Testing organization requires large numbers of diverse, high-quality items because computer-based tests are now delivered with variable forms at many times during the year, sometimes using designs where examinees receive detailed feedback. AIG is a method that may be used to produce these items. But even with item generation, reviews are still required to ensure that the appropriate standards of item quality are met. The purpose of our article was to describe and illustrate a comprehensive process for evaluating generated items by focusing on the content and logic provided in the cognitive and item models. The first review was internal to the AIG model development. The content and the logic specified in the cognitive and item models was used to create a validation table that permits the SME to review, evaluate, and revise the models without the need to review thousands of generated items. As a result, the SME can discern the quality of the generated items by reviewing a small amount of content and a small number of instructions in the AIG models. The second review was external to the AIG model development. It also focused on evaluating the quality of the AIG models using the validation table. However, the purpose of this second review was to conduct an independent evaluation of model quality and to provide feedback for improvement using a standardized rating scale. This two-part review is intended to support an iterative refinement approach to item development and to ensure the generated items meet high standards of quality.

Benefits of AIG Model Development and Validation

Item development will always be a complex process that requires content expertise. These demands will not change, but the development process can be modified, so it captures expertise more efficiently. AIG is an example of one contemporary item development method that was designed to efficiently capture content expertise. It has at least four important benefits for test developers. First, AIG permits the SME to create a single item model that will produce many test items. That is, AIG treats the item model as the unit of analysis where a single model is used to generate many items compared with a traditional item writing approach where the item is treated as the unit of analysis. The shift in the unit of analysis means that the SME manipulates only a small number of elements in each item model rather than manipulating every element in each new item (Gierl & Lai, 2012). This model-based approach, when combined with the review process presented in our article, can yield large numbers of high-quality items in an efficient and cost-effective manner.

Second, AIG is a scalable process because one item model can generate many test items. With a traditional item writing approach, the test item is the unit of analysis where each item is created individually. Because of this unit of analysis shift, the cost per item will decrease because SMEs are producing models that yield multiple items rather than producing single unique items. As a result, the efficiency of content development will increase because SMEs are creating models that produce many items. The models can be archived and used again, which will also have long-term efficiency and economic benefits (Gierl & Lai, 2012).

Third, AIG is a flexible approach to content production. Knowledge has become fluid and dynamic, particularly in the health sciences (Norman, Eva, Brooks, & Hamstra, 2006; Patel, Arocha, & Zhang, 2010). The creation of new drugs, the development of new clinical interventions, and the identification of new standards for best practice mean that test content in the health sciences is constantly changing. These changes are difficult, if not impossible, to accommodate in a flexible manner when the item is the unit of analysis because each element in the item is fixed. For example, if medical researchers conclude that antibiotics is no longer an effective method to manage fever after a gastrectomy (see Table 4), then all items directly or indirectly related to antibiotic treatment for gastrectomy must be identified and modified or deleted from the bank. But when the model is the unit of analysis, knowledge can be easily and readily updated to accommodate changes. In our example, gastrectomy is removed from Tables 4 and 5 and Figure 3 to accommodate this hypothetical change to best practice, and then the items are regenerated to reflect this change.

Fourth, the models used for AIG are compact and easy to store. Current item writing practices are grounded in the item. We noted earlier that each item is individually written, reviewed, revised, edited, and banked. If, for instance, an SME wants to have 2,000 operational test items in her bank, then she must write, review, revise, edit, and bank 2,000 items. What we did not describe earlier, however, is that these 2,000 items must also be tagged and stored. For instance, the format (e.g., multiple choice, numeric response, linked items), content (e.g., learning outcomes, blueprint categories, source of item), and SME attributes (e.g., year the item was written, item writer name, item writer demographics) are often coded for each item in the bank. The amount of work required for item-level tagging can be considerable. The physical space within a database required for storing item-level information can also be considerable. The 109,300 Common Core State Standards in Mathematics items described by Gierl et al. (2015), for instance, required 1 GB of hard drive space to store. By way of contrast, AIG is ground in the item model. If, for instance, an SME intends to have 30,360 algebra items and 6,240 surgery items, she may only require five algebra and five surgery item models (see our example in the previous section "Automatic Item Generation: A Three-Step Process"). Our review process is efficient because it is focused on reviewing 10 models rather than the 36,600 test items. The tagging task also focuses on model-level rather than item-level information. For example, these 10 models can readily be tagged for their format, content, and SME attributes. The bank may also contain coded information at the model-level variables such as item model ID, item model stem type, item model option type, number of constraints for the model, the number of elements in the model, and the number of generated items. But even with the addition of these model-level variables, the physical storage requirements for 10 models are comparatively small relative to 36,600 items. When a sample of the items is required, they are generated for test construction. But the items themselves are not stored within the database. As a result, small banks of models, which represent large numbers of items, can readily be stored, imported, and exported, thereby permitting the efficient transfer of large amounts of assessment content between users, agencies, and institutions to support our contemporary modes of test delivery and design.

Statistical Properties of Generated Items

AIG can be used to generate hundreds of new test items. The process we described offers a method for evaluating item quality by reviewing a small number of models that provide the core instructions for the generation task rather than reviewing large numbers of generated items. In addition to content, the statistical quality of the generated items must also be evaluated. Statistical item quality is typically determined through a field testing process where a small number of items are given to a sample of students so that indices of difficulty and discrimination can be computed. But the traditional approach to item analysis is neither feasible nor cost-effective when large numbers of generated items must be field tested. An alternative method for estimating the properties of the generated items is with psychometric methods that permit item *precalibration*. With precalibration, the statistical properties of the items can be estimated as part of the item generation process. One promising approach for precalibrating items is with skills-based methods, where the generated items are calibrated based on an analysis of the cognitive skills required to solve the items (e.g., Bradshaw & Templin, 2014; Embretson & Poggio, 2013). Skills-based item precalibration requires that an item-by-skill matrix, called the Q-matrix, be defined so that the knowledge and cognitive skills required to solve each item are explicitly identified. The items are then calibrated systematically using the statistical parameters associated with each cognitive skill outlined in the Q-matrix. Another promising approach uses the concept of item families to develop statistical models for calibrating generated items (e.g., Geerlings, Glas, & van der Linden, 2011; Luecht, 2013a; Sinharay & Johnson, 2013, 2013b; Sinharay, Johnson, & Williamson, 2003). The siblings (i.e., generated items) share common features within a family. Indices of within-family item variation can then be computed and included as part of the statistical model, thereby ensuring that the generated items contain similar content in addition to similar statistical characteristics. Future research is therefore warranted on expanding the methods and procedures that can be used to evaluate the quality of the statistics for generated items in addition to the quality of the content.

Acknowledgments

We would like to thank the Medical Council of Canada and ACT Inc. for their support with this research. However, the authors are solely responsible for the methods, procedures, and interpretations expressed in this study. We would also like to thank Dr. Howard Everson and two anonymous reviewers for their constructive feedback on an earlier version of this manuscript. The authors were presented with the 2016 National Council on Measurement in Educational Annual Award. This manuscript was prepared and presented at the annual meeting as part of the recognition for this award.

Notes

[1]Our example is adapted from a parent item used by a large testing company. The models in our example, however, have been simplified to highlight the basic concepts that underlie AIG. Operational AIG models are much larger, typically containing five to seven features with 2–10 variables per element.

[2]This is also a simplified example adapted from a parent item used in a testing program.

References

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium*. ETS Research Report 96-13. Princeton, NJ: Educational Testing Service.

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp.199–217). Hillsdale, NJ: Lawrence Erlbaum.

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. Available from http://www.jtla.org.

Bradshaw, L., & Templin, J. (2014). Combining scaling and classification: A psychometric model for scaling ability and diagnosing misconceptions. *Psychometrika*, *79*, 403–425.

Breithaupt, K., Ariel, A., & Hare, D. (2010). Assembling an inventory of multistage adaptive testing systems. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 247–266). New York, NY: Springer.

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 471–516). Washington, DC: American Council on Education.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219–250). Mahwah, NJ: Lawrence Erlbaum.

Embretson, S., & Poggio, J. (2013, April). *Some considerations for psychometric standards for generated items*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

Geerlings, H., Glas, C. A. W., & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, *76*, 337–359.

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

Gierl, M. J., & Lai, H. (2012). Using item models for automatic item generation. *International Journal of Testing*, *12*, 273–298.

Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36–50.

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

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., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *Journal of Technology, Learning, and Assessment*, *7*(2), 1–51. Available from http://www.jtla.org.

Graesser, A., Li, H., & Forsyth, C. (2014). Learning by communicating in natural language with conversational agents. *Current Directions in Psychological Science*, *23*, 374–380.

Gütl, C., Lankmayr, K., Weinhofer, J., & Höfler, M. (2011). Enhanced Automatic Question Creator – EAQC: Concept, development and evaluation of an automatic test item creation tool to foster modern e-education. *Electronic Journal of e-Learning*, *9*, 23–38.

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

Haladyna, T., & Shindoll, R. (1989). Items shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, *12*, 97–106.

Higgins, D. (2007). *Item distiller: Text retrieval for computer-assisted test item creation*. Educational Testing Service Research Memorandum (RM-07-05). Princeton, NJ: Educational Testing Service.

Higgins, D., Futagi, Y, & Deane, P. (2005). *Multilingual generalization of the Model Creator software for math item generation*. Educational Testing Service Research Report (RR-05-02). Princeton, NJ: Educational Testing Service.

Hively, W., Patterson, H. L., & Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, *5*, 275–290.

Hoijtink, H., Béland, S., & Vermeulen, J. (2014). Cognitive diagnostic assessment via Bayesian evaluation of informative diagnostic hypotheses. *Psychological Methods*, *19*, 21–38.

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

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.

Lane, S., Raymond, M., Haladyna, R., & Downing, S. (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.

Luecht, R. M. (2013a). An introduction to assessment engineering for automatic item generation. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 59–101). New York, NY: Routledge.

Luecht, R. M. (2013b). Automatic item generation for computerized adaptive testing. In M. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 196–216). New York, NY: Routledge.

Minsky, M. (1974). *A framework for representing knowledge*. MIT-AI Laboratory Memo 306.

Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design. In S. M. Downing & T. Haladyna (Eds.), *Handbook of test development* (pp. 61–90). Mahwah, NJ: Lawrence Erlbaum.

National Council on Measurement in Education. (2015). *Newsletter*, *23*, 1–23.

Norman, G., Eva, K., Brooks, L., & Hamstra, S. (2006). Expertise in medicine and surgery. In K. A. Ericsson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 339–353). Cambridge, UK: Cambridge University Press.

Patel, V. L., Arocha, J., & Zhang, J. (2010). Medical reasoning and thinking. In K. Holyoak & R. G. Morrison (Eds.), *Oxford handbook of thinking and reasoning* (pp. 736–754). Oxford, UK: Oxford University Press.

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.

Pugh, D., DeChamplain, A., Gierl, M. J., Lai, H, & Touchie, C. (2016). Using cognitive models to develop quality multiple-choice questions. *Medical Teacher*, *38*. doi:10.3109/0142159X.2016.1150989

Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. *Science*, *323*(5910), 75–79.

Reckase, M. (2016). Moving from art to science. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement* (pp. 174–178). New York, NY: Routledge.

Rudner, L. (2010). Implementing the graduate management admission test computerized adaptive test. In W. van der Linden & C. Glas (Eds.), *Elements of adaptive testing* (pp. 151–165), New York, NY: Springer.

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.

Shermis, M., Burstein, J., Brew, C., Higgins, D., & Zechner, K. (2016). Recent innovations in machine scoring of student- and test taker-written and spoken responses. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 335–254). New York, NY: Routledge.

Singley, M. K., & Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361–384). Mahwah, NJ: Lawrence Erlbaum.

Sinharay, S., & Johnson, M. S. (2013). Statistical modeling of automatically generated items. In M. J. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 183–195). New York, NY: Routledge.

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*, 337–359.

Sireci, S., & Zenisky, A. (2016). Computerized innovative item formats: Achievement and credentialing. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 313–334). New York, NY: Routledge.

van der Kleij, F. M., Eggen, T. J., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers and Education*, *58*, 263–272.

van der Kleij, F. M., Vermeulen, J. A., Schildkamp, K., & Eggen, T. J. (2015). Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment. *Assessment in Education: Principles, Policy and Practice*, *22*(3), 324–343.

Webb, M., Gibson, D., & Forkosh-Baruch, A. (2013). Challenges for information technology supporting educational assessment. *Journal of Computer Assisted Learning*, *29*, 451–462

Zenisky, A., & Hambleton, R., (2016). A model and good practices for score reporting. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 585–602). New York, NY: Routledge.