frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 04 May 2022
doi: 10.3389/feduc.2022.853578

Using Content Coding and Automatic Item Generation to Improve Test Security

Mark J. Gierl[1], Jinnie Shin[2], Tahereh Firoozi[1] and Hollis Lai[3]*

[1] Department of Educational Psychology, University of Alberta, Edmonton, AB, Canada, [2] Research and Evaluation Methodology, University of Florida, Gainesville, FL, United States, [3] Department of Dentistry & Dental Hygiene, University of Alberta, Edmonton, AB, Canada

Automatic item generation (AIG) is the process of using models to generate items using computer technology. AIG is a scalable content development method because it relies on the item model as the unit of analysis which means that it is more efficient and economical compared to traditional item development. But to use the generated items effectively, they must be managed properly. Managing a bank that could include millions of items results in problems related to identifying, organizing, and securing the content. As a result, the challenges inherent to managing item models and generated items warrant a shift in banking methodology where the management task must be accomplished using content coding at the model level. The purpose of our paper is to describe and illustrate methods that use content coding to organize and secure generated items in a bank.

Keywords: automatic item generation, content coding, technology and assessment, item development, test construction

OPEN ACCESS

Edited by:
Jennifer Randall,
University of Massachusetts Amherst, United States

Reviewed by:
Andrew Spielman,
New York University, United States
Audra Kosh,
Edmentum, United States
Andrea Gotzmann,
Medical Council of Canada, Canada

****Correspondence:***
Mark J. Gierl
mark.gierl@ualberta.ca

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *12 January 2022*
Accepted: *21 March 2022*
Published: *04 May 2022*

Citation:
Gierl MJ, Shin J, Firoozi T and Lai H (2022) Using Content Coding and Automatic Item Generation to Improve Test Security. Front. Educ. 07:853578. doi: 10.3389/feduc.2022.853578

INTRODUCTION

Testing organizations require large numbers of high-quality items to support innovations in test delivery and test design. Test delivery is one important source of innovation. Computer-based testing (CBT) has replaced traditional paper-based testing because the time, effort, and expense required to print, score, and report paper-based tests are prohibitive. Hence, paper-based testing is no longer feasible, nor desirable. CBT is a more efficient and economical method for administering tests. It also provides organizations with a range of new and desirable test administration options. CBT permits testing organizations to expand their test delivery services to include, for example, testing on-demand which allows examinees to write their exams on a more frequent and flexible schedule. CBT permits testing organizations to provide examinees with instant feedback thereby allows assessment to serve both formative and summative purposes. CBT permits testing organizations to create and deliver multimedia item types which can be used to measure more complex performances. Test design is another important source of innovation. Test design allows organizations to create assessment products that can be used to satisfy different purposes. For instance, CBTs can be designed to identify examinee's cognitive-problem solving strengths and weaknesses (e.g., Molnár and Csapó, 2019; Ran et al., 2021). CBTs can be designed to engage examinees in conversational dialogs thereby measuring speaking skills (Crossley and McNamara, 2013; Lu et al., 2018). CBTs can be designed to score examinees' written responses thereby measuring writing skills (Astanina et al., 2017; Mohamadi, 2018). CBTs can be designed to provide examinees with instant feedback using dynamic multimedia score reports thereby

increasing the interpretability of complex test performance (Attali and van der Kleij, 2017; Bulut et al., 2019; Horák and Gandini, 2019). In short, innovations in test delivery and test design allow organizations to offer a wide range of new and innovative products and services. But this expanded list of products and services also requires more testing. To address this need, large banks of items are needed. A bank is a repository of test items. These banks must be initially be created and then frequently replenished to ensure that examinees receive a continuous supply of unique, content-specific, items while limiting exposure to maintain test security.

Traditional paper-based tests designed for one specific purpose typically required a small number of items because tests were delivered in fixed length forms across a small number of administrations. 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, often using different test designs intended to serve different purposes. In most testing organizations, the large, content-specific, item banks required for modern educational testing are not readily available because the current approach used to create the content for these banks relies on a method where a subject-matter expert (SME) creates each item individually. Traditional item development is viewed as a process where SMEs use their experiences and expertise to produce each new test item, one at a time. Then, after the items are created, they are edited, reviewed, and revised—one item at a time—until they meet the required standards of quality (Lane et al., 2016). SMEs are solely responsible for identifying, organizing, and evaluating the content required for item development. Hence, this approach relies exclusively on human judgment acquired through extensive training and practical experiences. While the traditional item development approach can be used to create the banks needed for modern CBTs, it will always be a costly and time-consuming process due to the human effort needed to create large numbers of new items. As a result, it is challenging to meet the content demands required to satisfy emerging innovations in test delivery and test design because it is difficult to scale the traditional item development approach. Hence, item development is a critical bottleneck to innovation in modern educational testing.

One method that may be used to overcome this bottleneck is with automatic item generation (AIG). AIG is a new but rapidly evolving research area where cognitive theory and psychometric practice guide the production of items that are created with the aid of computer technology. AIG can be used to produce hundreds of items from a single item model. Hence, it serves as a method that can be used to scale the item development process. Gierl and Lai (2013, 2016) described a three-step approach to AIG. In step 1, the content for item generation is identified. SMEs identify and structure the content required to generate new test items. One framework that can be used to organize and structure the content is called a cognitive model for AIG (Gierl et al., 2012). A cognitive model for AIG highlights the knowledge and skills required to solve a problem in a specific content area. This model also organizes the cognitive- and content-specific information thereby presenting a structured representation of how the SME expect that examinees will think about and solve problems in a specific content area (see Gierl et al., 2021, Chapter 2). In step 2, an item model is developed to specify where the content from the cognitive model must be placed to generate new items. The item model (LaDuca et al., 1986) identifies which parts of the test item can be manipulated for 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., Singley and Bennett, 2002; Higgins et al., 2005; Gierl et al., 2008; Gütl et al., 2011; Khan et al., 2021). Logical constraints provide one straight-forward approach for generating items when item modeling is used (Gierl et al., 2021, Chapter 4). The generation process can be described as an iterator that permutes through all combinations of values and, in the process, eliminates combinations that do not meet the constraints defined by the SME in the cognitive model.

An example helps illustrate the logic required to generate test items. **Table 1** contains an item model created for differentiating the common cold from the seasonal influenza. This example was selected because it is easy to understand. It will be used throughout the manuscript to demonstrate key AIG concepts. The first panel at the top of the table contains the parent item. The examinee is presented with a vignette that contains information on both the patient history and the patient examination results. The examinee's task is to use this information to diagnose Acute Nasopharyngitis (Common Cold). The item model is presented in the second panel in **Table 1**. It contains five variables: Cough type, body aches, headache, throat, and fever. The values for the variables are provided in the third panel. Cough type, for instance, contains three values: mild, hacking, severe. The fourth panel contains the correct options for this model. In this example, the model contains content required to produce two correct options: Acute Nasopharyngitis and Seasonal Influenza. The fifth panel at the bottom of the table contains the incorrect options: Acute Nasopharyngitis, Bronchitis, Hay Fever, Acute Sinusitis, Bacterial Pneumonia, Acute Tonsillitis, Acute Laryngitis, Acute Tracheitis, Seasonal Influenza, Streptococcal infection, Nasal Diphtheria, Listeriosis, and Blastomycosis. The cognitive model is constrained so that only presentations such as mild and hacking cough, slight body aches and slight body pains, a sore throat, a mild headache, and 37.6 and 37.9°C are associated with Acute Nasopharyngitis. Incorrect options 2–9 are used as distractors for Acute Nasopharyngitis. This simple model generated 528 items. A sample of 4 items from the model is presented in **Table 2**.

To summarize, AIG is an item development method that supports innovation in educational testing because it can be used to overcome the scalability problem inherent to the traditional

TABLE 1 | Medical item model example.

Parent item:	
	A 31-year-old male sees his doctor and reports that he has been experiencing a mild cough with slight body aches. He has a mild headache and says he does not have a sore throat. Upon examination, the patient presents with an oral temperature of 37°C. What is the most likely diagnosis? (1) Hay fever (2) Seasonal flu (3) Otitis media (4) Acute nasopharyngitis*
Item model:	
Stem	A 31-year-old male sees his doctor and reports that he has been experiencing a (Cough Type) cough and (Body Aches). He has (Headache) headache and says he (Throat) a sore throat. Upon examination, he presents with an oral temperature of (Fever). What is the most likely diagnosis?
Variable: value	Cough type: (1) mild, (2) hacking, (3) severe Type of body aches: (1) slight body aches, (2) slight body pains, (3) severe body aches, (4) severe body pains Headache: (1) a mild, (2) a serious Throat: (1) has, (2) does not have Fever: (1) 37.6°C; (2) 37.9°C; (3) 38.9°C; (4) 39.2°C
Correct option	Acute nasopharyngitis; seasonal influenza
Incorrect option	Acute nasopharyngitis Bronchitis Hay fever Acute sinusitis Bacterial pneumonia Acute tonsillitis Acute laryngitis Acute tracheitis Seasonal influenza Streptococcal infection Nasal diphtheria Listeriosis Blastomycosis

item development approach. AIG is an efficient and cost-effective method because it integrated SME expertise with computer technology rather than relying solely on human expertise. AIG treats the model as the fundamental unit of analysis where a single model is used to generate many items compared with the traditional approach where the item is treated as the unit of analysis where each item is created individually. As a result, the number of required items is no longer tied to the number of SMEs who can write and review items. Rather, content creation is linked to the number of available models, where a small number of SMEs can create the models that produce large numbers of new items. AIG can therefore be used to meet the content demands required to satisfy emerging innovations in test delivery and test design because it is a scalable item development approach that can be used to produce large numbers of items. With an abundance of content, the next challenge becomes managing this resource.

MANAGING GENERATED ITEMS WITH CONTENT CODES

The AIG can be used to scale the item development process thereby addressing the challenge of producing large numbers of items. Management is the process of organizing and administering tasks in order to reach a goal. The management of a bank requires that items be appended with information so they can be identified and differentiated. Once differentiated, the items can be used to address a specific purpose or to achieve a particular goal within a testing organization. To create a digital assessment with a flexible and frequent administration schedule that serves multiple purposes, thousands of items are needed. This item volume can easily be created using AIG. However, to accommodate this volume, two management challenges must be overcome because a much larger number of items needs to be organized and managed. The first challenge stems from the sheer volume of items produced using AIG. Possessing a large bank is a new situation for most testing organizations (Cole et al., 2020). Managing a bank containing hundreds of items is complex. But when the bank is expanded dramatically to include millions of items, problems related to storage limits, search criteria, and content review quickly arise. The second management challenge occurs when shifting the unit of analysis from the item to the model. Testing organizations are familiar with developing and organizing items. But AIG creates another new challenge that many organizations have not likely experienced because the generating models must also be created, organized, and managed. While traditional development relies on processes where items are written, reviewed, and revised individually, AIG

uses processes where models are written, reviewed, and revised in order to generate items. Hence, testing organizations must manage the generating models in addition to the generated items in their bank.

To address these challenges, a shift in banking is required. With a traditional approach, items are managed at the item level. With AIG, items are managed at the model level. Model management is accomplished with the use of content coding. Content coding is a method used to describe data. For example, digital image files from a smartphone contain descriptive data such as image resolution, smartphone model, and color depth. Similarly, a bank contains descriptive data about each item that can be defined at many levels of specificity ranging from specific to general. For example, the format of the item can be described. Item formats include multiple choice, numeric response, written response, passage based, and multimedia based. The purpose for the item can be described. The purpose can be highlighted using learning objectives and blueprint categories. The item development attributes can be described. These attributes include year the item was written, SME name, SME demographics, development status, and review status. The statistical characteristics for the item can be described. Statistical characteristics include readability indices, classical item statistics, and item response theory parameters.

Identifying Content Codes

The type of descriptive data that is used in an AIG item can include conventional labels at the item level such as format, purpose, item development attributes, and statistical characteristics. But additional information in the form of content codes can also be produced. Content codes are created for different data elements (i.e., parts of the item model). This information is then appended to every item during the generation process thereby rendering each item as unique because every item has a unique set of content codes. Content codes can be created using two different methods. The first method is based on an *ad hoc* review of the model and/or the item where an SME decides on the content descriptors. This method does not require any advanced planning and therefore is quite flexible. The disadvantage of *ad hoc* coding is that the content definitions may be overly broad, subject to change and, challenging to interpret. *Ad hoc* coding is particularly problematic when different SME are responsible for creating the codes because SMEs often describe content differently. For example, a medical SME could describe the common cold as Nasopharyngitis, Rhinopharyngitis, or Acute Coryza. When different content codes are used to describe the same data element, searching for generated items with these codes is challenging, particularly as more item models are created to measure, in this example, respiratory illness because the conditions are not specific and the content codes that differentiate each medical condition are not defined which means that different search terms can yield different sets of generated items. Appending content codes to each generated item can also be a tedious and time-consuming task, especially when large numbers of items are created. SMEs must review the item, decide on the content codes, and then map the codes onto each generated item. This process may also require an additional review by an independent group of SMEs to ensure that the appropriate content code is selected and that the content codes are consistently applied to items. But perhaps the most important limitation of *ad hoc* content coding is that the links among the data elements are both precarious and uncertain because coding proceeds *ad hoc* rather than *a priori*.

The second method is based on applying a predefined content coding nomenclature found in a taxonomy. That is, content coding occurs *a priori*. A taxonomy contains a structure where the data elements are given additional meaning because of their position and relationship with other data elements (Gartner, 2016, Chapter 6). For example, K-12 science educators can use a science taxonomy from the National Assessment of Educational Progress (NAEP) to describe content areas (Physical Science; Life Science; Earth and Space Science) and practices (identifying science principles; using science principles; using scientific inquiry; using technological design) required by students to solve items on NAEP exams. Test items on NAEP are written by SMEs to measure these content and practices for students in grades 4, 8, and 12 (National Assessment of Educational Progress [NAEP], 2019). Science test items can be content coded and then located in the NAEP Science Framework in order to make inferences about students' knowledge and skills in a particular content area of science. Medical educators also use taxonomies. For instance, the Royal College of Physicians and Surgeons of Canada developed CanMEDS which contains a comprehensive list of competencies physicians are expected to acquire and demonstrate during their training (Frank et al., 2015). These competencies are organized thematically according to seven roles (i.e., medical expert; communicator; collaborator; leader; health advocate; scholar; and professional) where a competent physician is expected to integrate and achieve all of the competencies across all of the roles. Medical test items can be coded in reference to CanMEDS Standards in order to make inferences about physicians' knowledge and skills.

One important advantage of content coding *a priori* using a taxonomy is that it allows SMEs to have a specific set of existing content codes for describing AIG models and generated items. These codes have an established meaning that can be used to describe a specific data element thereby ensuring the coded outcome is interpretable. Another important advantage of using a taxonomy is that the data elements can be linked to other types of content both internal to and external from the taxonomy thereby creating sources of data about the content coded data to further enhance the meaning of the test items. Data about data is called metadata (Gartner, 2016). Metadata is meaningful when one data element can be directly associated with a second data element, where the second element provides information about the first. This relationship can be interpreted and, therefore, the relationship between the data elements has meaning. Metadata can be used to describe the characteristics of each generating model or generated item thereby allowing the SME to differentiate AIG content. Metadata can also be used to link diverse kinds and diverse sources of information together to create new types of information that, in turn, form meaningful structures of knowledge about the content in the bank. Libraries provide an important analogy for understanding the importance

TABLE 2 | Sample of generated items from the medical example.

1. A 31-year-old male sees his doctor and reports that he has been experiencing a mild cough and slight body aches. He has a mild headache and says he has a sore throat. Upon examination, he presents with an oral temperature of 37.6°C. What is the most likely diagnosis?

   A. Bronchitis
   B. Hay fever
   C. Acute sinusitis
   D. Acute nasopharyngitis*

2. A 31-year-old male sees his doctor and reports that he has been experiencing a hacking cough and slight body pains. He has a mild headache and says he has a sore throat. Upon examination, he presents with an oral temperature of 37.9°C. What is the most likely diagnosis?

   A. Acute tracheitis
   B. Acute laryngitis
   C. Seasonal influenza
   D. Acute nasopharyngitis*

3. A 31-year-old male sees his doctor and reports that he has been experiencing a severe cough and severe body aches. He has a serious headache and says he does not have a sore throat. Upon examination, he presents with an oral temperature of 39.2°C. What is the most likely diagnosis?

   A. Listeriosis
   B. Seasonal influenza*
   C. Streptococcal infection
   D. Acute nasopharyngitis

4. A 31-year-old male sees his doctor and reports that he has been experiencing a severe cough and severe body pains. He has a serious headache and says he does not have a sore throat. Upon examination, he presents with an oral temperature of 38.9°C. What is the most likely diagnosis?

   A. Seasonal influenza*
   B. Bacterial pneumonia
   C. Streptococcal infection
   D. Acute nasopharyngitis

of metadata. Libraries contain data that are used to describe and locate individual books in a collection. As a result, each book can be identified and located. Metadata can also be created to link one book by a specific author, as an example, to all books by that author which, in turn, can be linked to all books by that author on a specific topic, such as illness of the upper respiratory tract. The topic of illness of the upper respiratory tract can be described by an area such as the respiratory system. Hence the identity of one book in a collection can be used to link authors, topics, and areas through the use of metadata. This example demonstrates that data about data initiates a process of linking information where the links form relationships that increase the meaningfulness of the data, specifically, along with the content in the AIG bank, more generally. It also demonstrates that metadata can be used to address specific purposes (i.e., creating a library) or to solve specific problems (i.e., finding a particular book in a library on illness of the respiratory tract). In other words, the links that are created and the knowledge produced from these links is purposeful and intentional. Metadata permits the SME to organize, track, and manage large amounts of information that is a defining characteristic of AIG. In our example, banks can be likened to libraries, a generated item is comparable to a book, and the authors, topics, and areas are the content codes. The main disadvantage of using a taxonomy is that it may hamper or even restrict the content coding task. Coding is limited to the content in the taxonomy. A novel AIG cognitive model could be created to produce unique items that fall outside the range of the existing content codes as described in a taxonomy resulting in items that cannot be coded and classified. To address this limitation, more than two or one taxonomies can be used to code the generating models and the generated items where each taxonomy contain different levels of information so that one taxonomy, for instance, can be used to describe more specific information while another taxonomy can be used to describe more general information.

Locations for Content Coding in the Item Modeling Process

Content coding the item model can be characterized as adding a new dimension of information to this model. The location of this information can reside in three different places. The first location requires content coding at the model level. This type of data is the most general. With model-level coding, specific codes that describe all of the generated items from a particular item model are used. For example, an item model that is designed to generate items for diagnosing Nasopharyngitis could contain a "Disease of the Respiratory System" content code. As a result, all items generated from this model will be coded as a "Disease of the Respiratory System." The second location requires content coding at the value level. This type of data is the most specific. A cognitive model contains variables and values. Variables include values that will be manipulated during item generation. In our **Table 1** example, the variable Cough contains three values, mild, hacking, and severe. Hence, Cough is the variable and mild,

hacking, and severe are the values. Each variable and value can be content coded. Defining content codes at the variable and value level has important benefits for organizing and managing a bank because each generated item will have a unique value content code meaning that all of the generated items in the bank can be differentiated from one another. The third location requires content coding at the option level. This type of content code is unique for the selected-response item type. This type of code is often the most descriptive. Model- variable- and value-level content coding can be used to describe any type of generated item with a stem. But when the item type includes options, content coding can be applied to this information source as well. The correct and incorrect options can each have a unique content code. The correct option also highlights the purpose of the test item. Taken together, content coding at the model, variable, value, and option levels provides the SME with a comprehensive and flexible approach for adding content codes to the generating models as well as each generated item. Content codes embedded within a taxonomy allow SMEs to have a common understanding of the content that will be used to describe a specific domain thereby ensuring the coded items are interpretable. But for this benefit to be realized, an appropriate content coding system must exist. The codes must be applicable to the items that will be created using the AIG. The SMEs must also be trained to use the content coding system reliably and to interpret the content codes consistently. Hence, content coding will add additional time to the AIG process because it requires that specific information be coded for all of the correct and incorrect options. This is time consuming because the content codes must be applied to all the values in the model. The coding should be independently verified—as least for a sample of the models—by a second SME to ensure the codes are applied reliably. The computer program must also assemble the content codes in addition to creating the items during the generation process. Hence, content coding requires additional time and effort to implement during the three-step AIG workflow.

Applying Content Codes

Applying the content codes to the models and items adds an extra dimension of data to the item model in step 2 of the AIG process. Fortunately, content coding the model is substantially simplified compared to content coding the item because the unit of analysis shift means that the SME appends the codes to the model. These codes are then assembled across all of the content coding locations (e.g., model, variable, value, and option) during item generation in step 3 to produce a unique content code list for each item. Logical constraints provide a straight-forward approach for generating items when item modeling is used. The generation process permutes through all combinations of values and eliminates combinations that do not meet the constraints defined by the SME in the cognitive model. The outcome is a set of generated items. To produce descriptive data for each generated item, the content codes that are used for each model are added together to produce a list. This list, in turn, includes all of the codes that were used in the item model coding process. As a result, multiple content codes serve as the data that can be used to describe each generated item. Because different models, variables, values, and options contain different content codes, a unique list of content codes is compiled for each generated item.

EXAMPLE OF CONTENT CODING IN AUTOMATIC ITEM GENERATION USING A TAXONOMY

We return to the example in **Table 1** to illustrate a content coding method for AIG. The International Classification of Diseases (ICD-11) is the eleventh revision of the World Health Organization's classification system for content coding health information (World Health Organization [WHO], 2019). ICD-11 provides a common language that is used throughout the world for defining and reporting on diseases and other health-related problems. The foundational taxonomy is called the ICD-11 MMS which standards for Mortality and Morbidity Statistics (herein referred to as the ICD-11). This taxonomy contains more than 85,000 entities, where entities can be chapters, blocks or categories. ICD-11 consists of 28 chapters. A chapter is the top-level entity of the taxonomy. Within each chapter, a block is used to group related categories. A category is presented within a block where a category can be anything that is relevant to health care. All categories have a unique ICD code. The ICD-11 is an example of a contemporary medical taxonomy that can be used to describe medical outcomes such as disease. Hence, the ICD chapters, blocks, and categories can each be used as content codes for the model-level descriptors as well as the correct and incorrect options in an item model. In addition, the ICD-11 chapters, blocks, and categories are structured as a taxonomy and therefore have added meaning because of their position in the hierarchy as well as their relationships to one another.

SNOMED, the acronym for Systematized Nomenclature of Medicine, is an international classification system for medical terms that provides content codes needed for clinical documentation and reporting (National Library of Medicine, 2021). SNOMEDS provides a common language for defining and reporting on healthcare processes. It contains more than 350,000 concepts, where a concept is an entry that describes a clinical term. Each concept has a unique ID. Concepts are organized in hierarchies. As a result, concepts can be related to one another using more than 1.3 million links within the classification system. Concepts are also described by different clinical terms and phrases called descriptions thereby providing elaborated information for each category code. SNOMED serves as an example of a contemporary medical taxonomy that can be used to describe medical inputs such as healthcare processes. Hence, the SNOMED concepts and descriptions can be used as content codes for the variables and values in an item model. The SNOMED concepts and descriptions are structured as a taxonomy that contain millions of links that provide additional meaning to the categories and descriptions because of their position in the hierarchy as well as their relationships to one another. Taken together, the ICD-11 and SNOMED serve as two comprehensive classification systems for describing medical outcomes and processes, respectively. Both classification systems contain an extensive list of content codes that can be used

TABLE 3 | Medical item model with content codes.

Stem	A 31-year-old male sees his doctor and reports that he has been experiencing a (Cough Type) cough and (Body Aches). He has (Headache) headache and says he (Throat) a sore throat. Upon examination, he presents with an oral temperature of (Fever). What is the most likely diagnosis? (12)
Variable: value	Cough type (49727002): (1) mild (11833005), (2) hacking (59994004), (3) severe (43025008). Type of body aches (82991003): (1) slight body aches (82991003), (2) slight body pains (82991003), (3) severe body aches (76948002), (4) severe body pains (76948002). Headache (25064002): (1) a mild (44538002), (2) a serious (162299003). Throat (162397003): (1) has (267102003), (2) does not have (162387007). Fever (386661006): (1) 37.6°C (87273009); (2) 37.9°C (87273009); (3) 38.9°C (10151000132103); (4) 39.2°C (10151000132103).
Correct option	Acute nasopharyngitis (CA00); seasonal influenza (1E30)
Incorrect option	Acute nasopharyngitis (CA00) Bronchitis (CA20.Z) Hay fever (CA08.00) Acute sinusitis (CA01) Bacterial pneumonia (CA40.0Z) Acute tonsillitis (CA03.Z) Acute laryngitis (CA05.0) Acute tracheitis (CA05.1) Seasonal influenza (1E30) Streptococcal infection (1B51) Nasal diphtheria (1C17.1) Listeriosis (1C1A.Y) Blastomycosis (1F22)

to describe the model, variables, values, options in a medical item model. Both classification systems are also structured as a hierarchy which means that the content codes can be used to describe and to link data in a medical item model.

ICD-11 and SNOMED were used to content coding the item model in **Table 1**. The medical model example with content codes is presented in **Table 3**. These codes were used to describe all 528 generated items. Because content coding is conducted at the model level, it means that the string of data for each item is produced during the generation process. As a result, a unique string of content codes is created for each generated item. The content coding outcome for one generated item is shown in **Figure 1**. The first sample generated item from **Table 2** is provided at the top of **Figure 1**. The outcome from the item model in this example includes the model, correct option, and incorrect options. ICD-11 was used to code these three outputs. The first code is defined at the model level for the correct option. In our example, the correct option is Acute Nasopharyngitis. Acute Nasopharyngitis is located in Chapter 12 (Diseases of the Respiratory System) of ICD-11 (see **Figure 1**, column 1). The second column includes the correct option code. The correct option is Acute Nasopharyngitis. It contains the category content code of CA00 (**Figure 1**, column 2). The third column includes the incorrect option code. The incorrect options in our example are all selected from Chapter 12. They include Bronchitis (CA20.13), Hay Fever (CA08.00), Sinusitis (CA01), and Pneumonia (CA40.0Z) (**Figure 1**, column 3). The inputs from the example in **Figure 1** include the variables and values from the item model. SNOMED was used to code these two inputs. The fourth column is defined at the variable level in the item model. Our item model contains five variables specified at the concept level in SNOMED. The variables and category codes are cough (49727002), generalized aches and pains (82991003), pain in throat (162397003), headache (25064002), and fever (386661006) (**Figure 1**, column 4). The fifth column includes the values for the variables. The values for the sample item are mild; slight body aches; has; a mild; 37.6°C. Hence, the values are codes as dry cough (11833005), generalized aches and pains (82991003), sore throat symptom (267102003), nasal headache (44538002), and temperature normal (87273009) (**Figure 1**, column 5). A summary for the model, correct option, incorrect options, variables, and values for one generated item is provided at the bottom of **Figure 1**.

A bank of generated items that contain content codes from two existing taxonomies, as presented in our example, has two immediate benefits. The first benefit is classification consistency. Content codes from an existing taxonomy can be used to create models and items that are meaningful because the content codes can be applied consistently. The ICD-11 chapters, for instance, provide a comprehensive list of diseases and other health-related outcomes. The chapters could therefore be used to create a bank of items that measures *all disease-related healthcare outcomes*. The models and items in this bank are interpretable because the content codes describe specific outcomes in the ICD-11. Similarly, SNOMED concepts can be used to describe healthcare processes. Hence, the concepts can be used to create a bank of items that measure specific medical processes and procedures. The models and items in this bank are interpretable because the content codes contain specific health care terms that are used throughout the world for clinical reporting. Taken together,

A 31-year-old male sees his doctor and reports that he has been experiencing a [MILD] cough and [SLIGHT BODY ACHES]. He has [A MILD] headache and says he [HAS] a sore throat. Upon examination, he presents with an oral temperature of [37.6°C]. What is the most likely diagnosis?

A. Bronchitis
B. Hay Fever
C. Acute Sinusitis
D. Acute Nasopharyngitis*

ICD-11			SNOMED	
Model	Correct Option	Incorrect Option	Variable	Value
12 Diseases of the respiratory system Upper respiratory tract disorders	CA00 Acute nasopharyngitis	Bronchitis 12 Diseases of the respiratory system Certain lower respiratory tract diseases CA20 Bronchitis CA20.Z Bronchitis, unspecified Hay Fever 12 Diseases of the respiratory system Upper respiratory tract disorders CA08 Vasomotor or allergic rhinitis CA08.0 Allergic rhinitis CA08.00 Allergic rhinitis due to pollen Acute Sinusitis 12 Diseases of the respiratory system Upper respiratory tract disorders CA01 Acute sinusitis	49727002 Cough 82991003 Generalized aches and pains 25064002 Headache 162397003 Pain in throat 386661006 Fever	1. mild, 11833005 Dry cough 1. slight body aches, 2. slight body pains, 82991003 Generalized aches and pains 1. a mild, 44538002 Nasal headache 1. has, 267102003 Sore throat symptom 1. 37.6°C; 2. 37.9°C; 87273009 Temperature normal
12	CA00	CA20.Z CA08.00 CA01	49727002 82991003 162397003 25064002 386661006	11833005 82991003 44538002 267102003 87273009

FIGURE 1 | ICD-11 and SNOMED content codes for one generated item with metadata.

models and items can be coded using the chapters and concepts from ICD-11 and SNOMED thereby providing the foundation for creating a *global medical testing standard* because the items on any medical test can be described using the content codes in these taxonomies. This means that every generated item could be described with the same medical concepts and terms in order to convey the same meaning about the content that is measured by the items on the tests. The items could also be described in 43 different languages using the ICD-11 and seven different languages using SNOMEDS.

The second benefit is practicality. Content codes from an existing taxonomy have practical value because they can be used to create and validate the AIG models and items. Distractor development serves as one example. Creating plausible but incorrect distractors in AIG is a challenging task. Selected-response items require the examinees to distinguish among options that differ in their relative correctness where subtle but meaningful distinctions exist among the distractors. Distractors are often generated based on their relationship to the correct option where the distractors are related to but still distinct from the correct answer. To create distractors, the features required to produce the correct option are first identified and then these features are used again to construct the distractors. The variables and values in an AIG model describe these features. Hence, the features of the correct and incorrect options must be identified. Added to this challenge is that large numbers of plausible but incorrect options must be created. A selected-response item always contains one stem and one correct option. But it also requires three (i.e., 4-option item) or four (i.e., 5-option item) distractors for each correct

option. Hence the challenge with creating effective distractors is selecting three or more plausible but incorrect options that are related to but distinct from the correct option. Taxonomies are helpful for creating plausible distractors because content codes can be used to identify related concepts for an item model. In our example, one of the correct options is Acute Nasopharyngitis. Acute Nasopharyngitis contains the chapter code 12 (Diseases of the Respiratory System) and the category code CA (Upper Respiratory Tract Disorders) in the ICD-11. Hence, potential distractors can be found in chapter 12 and category CA because the diseases in this taxonomic category are all related to but different from Acute Nasopharyngitis. Hay Fever (CA08), as an example, could be a plausible distractor for Acute Nasopharyngitis because it comes from 12/CA but it contains a different code (Hay Fever is CA08). Twenty-two different upper respiratory tract disorders are identified in ICD-11. Hence, distractors can be identified by their content codes.

In addition, each disease is described in ICD-11. These descriptions can be used as a source of validity evidence to justify the selection of a particular distractor. Acute Nasopharyngitis (CA00) is described as: "A disease of the upper respiratory tract, caused by an infection with rhinovirus. This disease is characterized by pharyngitis, runny nose, stuffy nose, or cough. Transmission is by inhalation of infected respiratory secretions, or direct contact." Hay Fever is a good distractor for Acute Nasopharyngitis because it shares some but not all of the features of the Common Cold. Hay Fever is described in the ICD-11 as: "Rhinitis is inflammation of the nasal mucosa clinically characterized by major symptoms: sneezing, nasal pruritus, running nose, and stuffy nose. Allergic rhinitis is an inflammation of nasal airway triggered by allergens to which the affected individual has previously been sensitized." These two descriptions reveal that Acute Nasopharyngitis and Hay Fever both share the symptoms of a runny nose and stuffy nose. As a result, Hay Fever could be used as a plausible but incorrect options because it shares some but not all of the features of Acute Nasopharyngitis. Seven of the distractors (i.e., Bronchitis, Hay Fever, Acute Sinusitis, Bacterial Pneumonia, Acute Tonsillitis, Acute Laryngitis, and Acute Tracheitis) used with the correct option Acute Nasopharyngitis in our **Table 1** example were identified and validated using ICD-11 content codes and descriptions.

Automatic item generation requires a novel approach to banking because large numbers of items must be managed. Content coding is an effective method for adding a descriptive list of data to the generating model and the generated items in order to manage the content in a bank. By applying codes to the model as an additional layer of information, content can be generated for each item during the generation step. Our examples demonstrate how content codes can be appended to each model and item as a list. This data describes each generated item at the model, correct option, incorrect option, variable, and value level. Taxonomies provide content codes that can be used to describe the models and items consistently. Taxonomies also contain metadata that can be used to create and validate the content in AIG models.

ITEM SECURITY AND AUTOMATIC ITEM GENERATION

Item security is unquestionably a topic of importance in educational testing. Security breaches threaten the validity of the testing process and of the test score interpretations. When operational items are available to examinees prior to a test administration, pre-knowledge of the compromised items provides an unfair advantage for those examinees who have gained prior access. The statistical characteristics of the compromised items are also affected. Psychometric methods are available to detect aberrant test performance (e.g., Belov and Armstrong, 2010; Sinharay, 2017; Liu et al., 2019). However, these methods are limited to *post hoc* analyses conducted at the item level after the test has been administered. Hence, these methods are retrospective and can only inform item security policies and practices that are applied after the test administration.

Security strategies can also be prospective. These strategies are created to prevent item exposure prior to the exam administration, to prevent examinees from sharing operational items with one another, and to prevent examinees from preparing for the exams using breached items. Solutions for limiting item exposure prior to the exam administration can be addressed by implementing electronic credentials. Credentials mean that only registered examinees are permitted to sit for the exam which controls and restricts access during the test administration. Solutions to prevent examinees from sharing items and from preparing for exams using previously administered items require strategies applied to the test items and test administrations. To address these risks, site-specific, form-specific, and item-specific strategies can be used. Site-specific strategies involve securing the location of testing by invigilating and by monitoring examinees during the test administration. An example of a site-specific strategy would be a policy that prevents examinees from taking written information in to or out of an exam. This strategy focuses on limiting the examinee's ability to record information during the exam, but it does not prevent examinees from memorizing the items. Form-specific strategies are intended to prevent examinees from remembering information about test items. Randomizing the order of the items within a single form of the exam is a strategy intended to limit an examinee's ability to memorize items. While this strategy varies the presentation of the items thereby decreasing an examinee's ability to memorize the content, the problem still remains that item exposure is increasing proportional to the number of examinees who viewed the items. Item-specific strategies are intended to limit item exposure. The use of multiple test forms is one strategy that can be used to reduce the item exposure rate (Wendler and Walker, 2016). However, the most effective item-specific strategy for limiting exposure and enhancing security is to increase the size of the bank. Increasing bank size decreases item exposure because the bank contains a large number of unique items. AIG is a method for scaling the item development process in order to produce large numbers of new items efficiently and economically thereby increasing the size of the item bank. Scaling the item

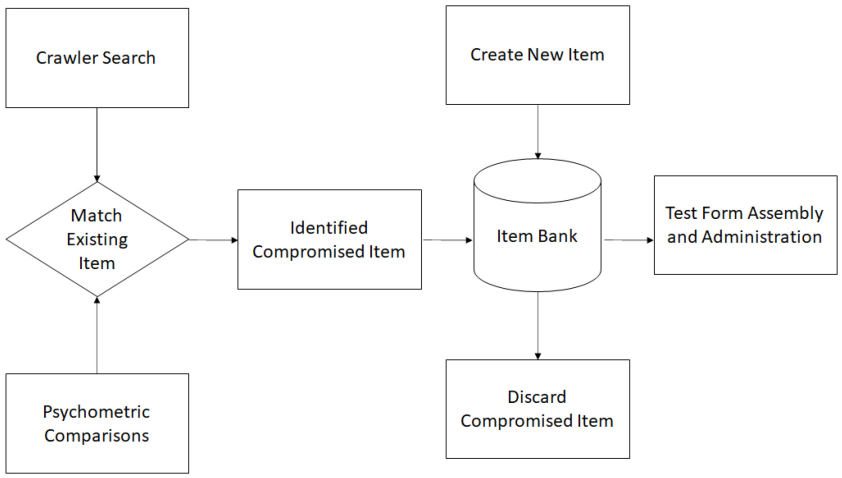

FIGURE 2 | Illustration of a simplified process for detecting compromised items.

development process still does not completely alleviate the issue of item security.

A METHOD FOR IMPROVING SECURITY USING ITEM MODELING WITH CONTENT CODING

Specifications and blueprints guide the test construction process by identifying the items required in each content area and cognitive skill category. As we demonstrated, item can be described using content codes. Content coding is the method of annotating data to each item that, in turn, can be used to describe the items as well as link the items to other items to content descriptors inside and outside of the bank. Content coding requires the identification and application of a predefined nomenclature or taxonomy of content codes. Attaching content codes to each individual item is both laborious and time consuming. Content coding the item model, on the other hand, simplifies the coding process because the SME appends codes to specific parts of the model and then a specific set of codes is assembled for each item based on which part of the model was used. The outcome of the assembly process is a unique content code for each generated item.

Incorporating content coding into the item model also provides a novel approach for securing each item. **Figure 2** illustrates a scenario for detecting a compromised test. Web crawlers can be used to identify exposed test items. A web crawler is software that searches websites for specific types of information. Web crawlers are instructed to locate content on forums or websites and then to compare this content to the items in a bank. The process typically involved comparing text. When an item is compromised or exposed, an investigation ensues but in most cases the item is discarded and a new, parallel item is created and then used in future test administrations. This scenario, albeit simplistic, highlights two issues in the current approach to managing compromised items. First, when an item is compromised, replacement items are developed as part of a separate process. This process requires tracking because the content area for the compromised and for the new test item must be identified. Hence, items require content codes. Second, the current approach is reliant on information from the item. This implies that the security protocol must be conducted at the item level and provided for every individual item in the bank. With AIG, information is also available at the model level. Model-level information can provide new strategies to enhance item and test security.

Item models contain the parameters, constraints, instructions, and content codes required to produce test items (Gierl et al., 2021). This characterization can be expanded. For example, the production information such as which item is created from the model and the psychometric information such as the item difficulty level can also be captured in the item model thereby increasing the utility of the model to include information that can be used to create items as well as track items after they have been administered. A modern approach to item modeling would append models with three additional sources of information: item production specifications, item production logs, and item administration logs. An item production specification is the current standard of practice, as illustrated in **Table 1**. It contains the parameters, constraints, instructions, and content codes required to produce test items. This instruction set is an important aspect of the model because it changes the unit of analysis from the item to the model, meaning that each generated item is no longer viewed as a unique, individual item but rather as a traceable expression from a generating model. When the item production specification has been completed and reviewed, the model is placed into production where the item production log is used to track the generated items from the model. Item production logs contain information such as which combination of content was used for generation, which set of associated content codes applies to each generated item, and which exams used each generated item. After the generated items are administered, the item administration log is created.

This log stores information about the psychometric properties (e.g., student responses, item difficulty levels, and the distractor discrimination indices) of the generated items.

A modern item modeling approach with incorporates content coding has immediate benefits. First, inferences about the testing process can now be made at the model level. Information such as how the item was created, where the items were used, what statistical outcomes the item produced are all recorded and are all available in a single source. This information can be mined in order to pursue new approaches to field testing such as bootstrapping and extrapolating psychometric results across items that share item production specifications. Second, item models can be evaluated in three different ways. The text of the model can be evaluated. The content codes of the model can be evaluated. The statistical characteristics of the model can be evaluated. Third and, perhaps, most importantly, item security strategies can now be implemented at the model level. For example, if an item is compromised, there are three data sources that can be used to evaluate the scope and the nature of the problem. A text search can be conducted to identify items already in production and to prevent similar items from being administered. A content code search of the compromised items can be conducted to identify existing operational items that share those codes. Generated items with similar content codes could also be removed from the bank and replacement items could be generated. A statistical search of the generated items can be conducted to monitor the psychometric characteristics of the generated items. For example, the statistical performance of the items in a model can be monitored and evaluated to determine if performance was outside of the expected range suggesting a possible security breach. In short, the item model definition can be expanded to include item production specifications, item production logs, and item administration logs. This information can be used to guide the generation process in addition to tracking the items, mining data contained in the model, and implementing new methods to monitor and address item security issues.

USING ITEM MODELS TO ENHANCE ITEM SECURITY: THREE EXAMPLES

To demonstrate how item models can be used to enhance security, we return to the medical model presented in **Table 1**. In example 1, content coding at the model level is used for parallel forms construction in order to limit item exposure. When the item is the unit of analysis, content coding is applied to the item. For instance, items can be coded for respiratory illness. However, large numbers of items in a bank might contain the respiratory illness content code meaning the items are challenging to differentiate. When the model is the unit of analysis, content coding is applied to the model. As a result, items can now be queried and requested using the model, variable, value, and option content codes. For instance, sample item 4 in **Table 2** can be requested by content code as follows: Infectious Disease (01–Model), Seasonal Influenza (1E30—Correct Option), Bacterial Pneumonia (CA40.0Z—Incorrect Option 1), Streptococcal Infection (1B51—Incorrect Option 2), Acute Nasopharyngitis (CA00—Incorrect Option 3), Cough Type (43025008—Severe), Body Aches (76948002—Severe Body Pains), Headache (162299003—Serious), Throat (162387007—No Sore Throat), Fever (10151000132103—38.9°C). Test assembly can then be conducted using content codes by selecting new items with the same content code at the model, variable, value, and/or option levels from the previous test administration. This example demonstrates an item-specific strategy for item security. Item-specific strategies limit item exposure. The use of different forms that measure the same content but with different (i.e., parallel) items reduce item exposure.

In example 2, content coding at the model level is used to identify compromised items. A web crawler can be used by a testing organization to identify compromised test items. The crawler locates items and then compares the content in the items to the content in a bank. With the item as the unit of analysis, matching algorithms identify whether the content from an outside source is similar to the text used in an existing item. The comparison is conducted at the item level meaning every item must to be evaluated individually. When the model is the unit of analysis, similarity can be evaluated in three different ways. First, the generated items used on an operational test form can be indexed and searched using text comparisons, as with the item-level approach. Second, the content used in the item model can direct the search and comparison. For instance, sample item 4 in **Table 2** can be searched using the string values in asterisks as follows: A 31-year-old male sees his doctor and reports that he has been experiencing *a severe cough* and *severe body pains*. He has a *serious headache* and says he does *not have a sore throat*. Upon examination, he presents with an oral temperature of *38.9°C*. What is the most likely diagnosis? *Seasonal Influenza* *Bacterial Pneumonia* *Streptococcal Infection* *Acute Nasopharyngitis*. This approach allows for the search to be conducted much more quickly because the comparisons are based on the most meaningful segments of text rather than the entire text. In addition, the results from the model, variable, value, and options search can be indexed to the corresponding content code to further enhance search findings. For instance, if the first item presented in **Table 2** was found to be compromised, then the text that corresponds to the content code such Acute Nasopharyngitis (CA00), Bronchitis (CA20.13), Hay Fever (CA08.00), Sinusitis (CA01), Pneumonia (CA40.0Z), cough (49727002), headache (25064002), and fever (386661006) could be used to expand the search criteria to determine if other related items were exposed. Feedback from the web crawler search which uses the item text as well as the indexed model, variable, value, and option content codes can also help gauge the security of the model in the bank. If a significant amount of the content in a model has been exposed, then a new model can be created to measure the outcomes in a test specification or blueprint category. But it is important to note that replacing items using the multiple test forms method can be a complex task. For example, before the compromised items on the test can be replaced with alternative items from a new AIG model, the item position and the item

options must be considered in the selection of the replace items because these factors can affect the test score equating results (Kolen and Brennan, 2004, Chapter 8; Wendler and Walker, 2016). AIG provides a method for creating content. But additional considerations specific to the test design and to the equating methods must be taken into account when selecting the replacement items.

In example 3, results from the compromised items are used as feedback to prevent the generation of content that has been disclosed. Recall that content in AIG models is controlled through the use of constraints. The use of different constraints results in the generation of different items. Once a generated item has been compromised, this information can be tracked in the item administration log and items with the same content codes can be removed from the test administration. For example, if item 4 in **Table 2** was compromised, then the model can be adjusted using constraint coding so that no item with the same content codes as item 4 will be generated in the future. This restriction will limit the model so that it cannot be used to generate exposed content. In this example, items with the following content codes will not be generated: Infectious Disease (01–Model), Seasonal Influenza (1E30—Correct Option), Bacterial Pneumonia (CA40.0Z—Incorrect Option 1), Streptococcal Infection (1B51—Incorrect Option 2), Acute Nasopharyngitis (CA00—Incorrect Option 3), Cough Type (43025008—Severe), Body Aches (76948002—Severe Body Pains), Headache (162299003—Serious), Throat (162387007—No Sore Throat), and Fever (10151000132103—38.9°C). Because each security breach is logged as an event, each request to limit generation can then be added to the item production specification and, as a result, items with the same content codes or items with text that is similar to the compromised items are not generated.

SUMMARY AND IMPLICATIONS

Innovation is occurring rapidly in test delivery and test design. There is also a noteworthy shift in testing policies and practices due to the COVID pandemic. Amid these changes, organizations must instill trust in their testing process where security is of paramount importance. Large numbers of new, content-specific, high-quality items are needed to support testing innovation and to enhance test security. AIG is an item development method that can be used to produce large numbers of test items in order to promote innovation and improve security. AIG is the process of using models to create test items using computer technology. AIG integrates human judgment and computer technology. It also relies on the model rather than the item as the unit of analysis. As a result, content creation is no longer tied to the number of SMEs who can write and review items. Instead, it is related to the number of available models, where a small number of SMEs can create the models needed to produce large numbers of items thereby scaling the item development process. But with an abundant supply of content, managing this resource becomes the next challenge. Management relies on appending items with information so the items can be identified and differentiated in a bank. An effective management approach is used to describe the data and then to link the data to other sources of information. In this paper, we presented and illustrated a modern approach to item modeling using content coding. This approach requires items to be managed at the model level in the bank. It requires content coding using a taxonomy. It requires collecting and analyzing information in the item production specifications, item production logs, and item administration log in order to enhance item security.

Implications for Innovation

A traditional item bank serves as an electronic repository for storing, maintaining, and managing information on each item. The maintenance task focuses on item-level information. For example, the format of the item, the purpose of the item, the developer attributes of the item, and the statistical characteristics of the item can be described. Many different people in a testing organization are involved in the item development process including the test development specialists, SMEs, psychometricians, editors, graphic designers, and document production specialists. Procedures must therefore be created and implemented to decide who has access to the bank and when items can be created, added, modified, or removed.

Models, rather than items, serve as the unit of analysis in an AIG model bank. A model bank is an electronic repository for storing, maintaining, and managing information on each model. Each model—which is individually written, reviewed, revised, edited, and banked—can be used to generate many items. Because the AIG model serves as the unit of analysis, the banks contains information on every model as well as every item. This information can be used to address three purposes (Gartner, 2016). The first purpose is descriptive. Descriptive data can be used to locate information. Descriptive meta includes relationships among the content codes in an AIG model bank. Descriptive data can describe and locate models in the bank. The second purpose is administrative. Administrative data can be used to store, access, and preserve information. Administrative data can include the item production logs in an AIG model bank. These logs contain information such as which combination of content was used for generation, which set of associated content codes applies to each generated item, and which exams used each generated item. The third purpose is structural. Structural data can be used to link smaller pieces of information in order to produce larger and more meaningful pieces of information. Structural data can include the item production specification in an AIG model bank. These specifications contain the parameters, constraints, instructions, and content codes required to produce test items.

But unlike the traditional item banking approach, where many different people in a testing organization are involved, modeling banking must be its own specialization (Gierl and Lai, 2012). This specialist is skilled in test development, but also in computer programming and database management. The *AIG model bank*

developer helps bridge the gap between the SME who creates the cognitive and item models and the required programming tasks needed to constrain the models, format output, and generate the items. The model developer is also responsible for entering the models into the bank, coding the models, maintaining the contents of the bank, and managing the use of the model bank. In other words, the AIG model bank developer curates information in the content development system. Gartner (2016) explains:

> *"Curation is often confused with preservation, but there is much more to it than this alone. Curation involves identifying those elements of a culture that particularly define it and choosing which ones are important; it then describes and adds context to these, making connections between them, so that they can be understood by all who have an interest in them. Finally, it involves disseminating a culture, making it accessible. All of these are in addition to ensuring that these elements will continue to exist for a long time in the future. Going through these steps ensures above all that a culture can be understood when it is transmitted between generations." (p. 12)*

The culture for a testing organization is based on the content in their exams. Item development therefore provides the context needed to initially create and then expand this culture. Content coding can be used to describe items, but it can also be used to link information in order to create new types of information that, in turn, form meaningful structures of knowledge about the testing organization. In the same way that libraries provide a useful example for demonstrating the importance of content coding and metadata in an information system, model banks provide an example for demonstrating the importance of content coding and metadata for the culture in a testing organization. Model banks contain data that are used to describe and locate items in a bank. As a result, one particular item can be identified and found in the bank (see **Table 2**, Item 2). But content coding can yield metadata that links one item to one model (see **Table 1**), as an example, to all items from the same model (see **Table 2**, Items 1–4) which, in turn, can be linked to all models in a specific content area such as "Certain lower respiratory tract diseases" or "Diseases of the respiratory tract" (see **Figure 1**). These content areas can be located in a taxonomy—such as ICD-11 and SNOMED—that provides definitions and descriptions for the outcomes and processes in a discipline which could be used to create a universal testing standard because every generated item could be described with the same medical concepts and terms in order to convey the same meaning about the content that is measured by the items on the tests. Hence, the identity of one item in a bank can be used to link items, models, and content areas through the use of content coding and metadata. In short, content coding is a powerful method for appending metadata to the generated items because the content that can be used to describe the data in a single generated item is practically limitless. For this reason, items produced from AIG models appended with metadata could address any purpose or solve any problem within a testing organization because the number of meaningful relationships that could exist between the data elements is extraordinarily large.

Directions for Additional Research

The taxonomies introduced and illustrated in our manuscript were from the content area of medicine. The taxonomies were also described at a fine-grain size thereby yielding great specificity in the content. Additional research is required to discern if taxonomies in other content areas, such as the science taxonomy from the National Assessment of Educational Progress, can be used to content code generated items, as described in our manuscript. Additional research is also required to evaluate whether taxonomies with different levels of granularity (e.g., CanMEDS competencies) can be used for AIG content coding. While we described AIG content coding methods and applications, the generalizability of our results still need to be evaluated with different types of taxonomies that contain different levels of content coding. Hence, the applicability of content coding using different classification systems remains an important area of future research.

In this manuscript we described content coding newly generated items in a bank. But many different types of item banks exist (Vale, 2006). Moreover, there is no research on the use of different types of existing item banks (e.g., commercial and custom-made) as it applies to AIG and content coding. Hence, important item banking issues remain to be addressed. For example, are existing banks capable of importing AIG items in bulk? Are commercial item banks compatible with different AIG classification systems? Are existing banks capable of importing different types of content codes and metadata, as described in our section titled "A Method for Improving Security Using Item Modeling with Content Coding"? How can content coding be executed in existing banks that contain items created using both traditional and modern item development methods? What are the best strategies for classifying anomalous items that do not adhere to a specific content coding system? These questions highlight practical issues that reside at the intersection between item banking, content coding, and item development that require additional investigation.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

MG contributed to the ideas in the manuscript, conducted analyses, and wrote the manuscript. JS and HL contributed to the ideas in the manuscript and wrote the manuscript. TF conducted analyses and reviewed the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

Astanina, A., Beliaeva, I., and Rasskazova, T. (2017). "Providing computer-based testing for low-level students at a Russian university," in *Proceedings of the The International Scientific Conference eLearning and Software for Education*, Vol. 1, (Bucharest: Carol I" National Defence University), 132–139.

Attali, Y., and van der Kleij, F. (2017). Effects of feedback elaboration and feedback timing during computer-based practice in mathematics problem solving. *Comput. Educ.* 110, 154–169. doi: 10.1016/j.compedu.2017.03.012

Belov, D., and Armstrong, R. (2010). Automatic detection of answer copying via kullback-leibler divergence and K-index. *Appl. Psychol. Meas.* 34, 379–392. doi: 10.1177/0146621610370453

Bulut, O., Cutumisu, M., Aquilina, A. M., and Singh, D. (2019). Effects of digital score reporting and feedback on students' learning in higher education. *Front. Educ.* 4:65. doi: 10.3389/feduc.2019.00065

Cole, B. S., Lima-Walton, E., Brunnert, K., Vesey, W. B., and Raha, K. (2020). Taming the firehose: unsupervised machine learning for syntactic partitioning of large volumes of automatically generated items to assist automated test assembly. *J. Appl. Test. Technol.* 21, 1–11. doi: 10.1093/oso/9780190941659.003.0001

Crossley, S., and McNamara, D. (2013). Applications of text analysis tools for spoken response grading. *Lang. Learn. Technol.* 17, 171–192.

Frank, J. R., Snell, L., and Sherbino, J. (2015). *CanMEDS 2015 Physician Competency Framework*. Ottawa, ON: Royal College of Physicians and Surgeons of Canada.

Gartner, R. (2016). *Metadata: Shaping Knowledge from Antiquity to the Sematic Web*. New York, NY: Springer.

Gierl, M. J., and Lai, H. (2012). Using item models for automatic item generation. *Int. J. Test.* 12, 273–298. doi: 10.1080/15305058.2011.635830

Gierl, M. J., and Lai, H. (2013). Using automated processes to generate test items. *Educ. Meas.* 32, 36–50. doi: 10.1111/emip.12018

Gierl, M. J., and Lai, H. (2016). "Automatic item generation," in *Handbook of Test Development*, 2nd Edn, eds S. Lane, M. Raymond, and T. Haladyna (New York, NY: Routledge), 410–429.

Gierl, M. J., Lai, H., and Tanygin, V. (2021). *Advanced Methods in Automatic item Generation*. New York, NY: Routledge.

Gierl, M. J., Lai, H., and Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Med. Educ.* 46, 757–765. doi: 10.1111/j.1365-2923.2012.04289.x

Gierl, M. J., Zhou, J., and Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *J. Technol. Learn. Assess.* 7, 1–50.

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

Higgins, D., Futagi, Y., and Deane, P. (2005). *Multilingual Generalization of the Model Creator Software for Math Item Generation (Research Report No. RR-05–02)*. Princeton, NJ: Educational Testing Service.

Horák, T., and Gandini, E. (2019). "Improving Feedback through Computer-Based Language Proficiency Assessment," in *Innovative Language Teaching and Learning at University: a Look at New Trends*, eds N. Becerra, R. Biasini, H. Magedera-Hofhansl, and A. Reimão (Wuhan: Research-publishing.net), 95–103. doi: 10.14705/rpnet.2019.32.906

Khan, S. M., Hamer, J., and Almeida, T. (2021). "Generate: a NLG system for educational content creation," in *Proceedings of the 14th International Conference on Educational Data Mining*, Paris.

Kolen, M. J., and Brennan, R. L. (2004). *Test Equating, Scaling, and Linking: Methods and Practices*, 2nd Edn. New York, NY: Springer.

LaDuca, A., Staples, W. I., Templeton, B., and Holzman, G. B. (1986). Item modelling procedures for constructing content-equivalent multiple-choice questions. *Med. Educ.* 20, 53–56. doi: 10.1111/j.1365-2923.1986.tb01042.x

Lane, S., Raymond, M., and Haladyna, R. (2016). "Test development process," in *Handbook of Test Development*, 2nd Edn, eds S. Lane, M. Raymond, and T. Haladyna (New York, NY: Routledge), 3–18.

Liu, C., Han, K., and Li, J. (2019). Compromised item detection for computerized adaptive testing. *Front. Psychol.* 10:829. doi: 10.3389/fpsyg.2019.00829

Lu, Z., Zheng, C., and Li, Z. (2018). Effects of embedded summary writing on EFL learners' anxiety and oral production in a computer-based testing environment. *J. Comput. Educ.* 5, 221–241. doi: 10.1007/s40692-018-0105-1

Mohamadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Stud. Educ. Evaluat.* 59, 29–40. doi: 10.1016/j.stueduc.2018.02.003

Molnár, G., and Csapó, B. (2019). "Technology-based diagnostic assessments for identifying early mathematical learning difficulties," in *International Handbook of Mathematical Learning Difficulties*, eds A. Fritz, V. G. Haase, and P. Räsänene (Cham: Springer), 683–707. doi: 10.1007/978-3-319-97148-3_40

National Assessment of Educational Progress [NAEP] (2019). *Science Framework for the 2019 National Assessment of Educational Progress*. Washington, DC: National Assessment Governing Board.

National Library of Medicine (2021). *SNOMED CT*. Bethesada, MD: National Library of Medicine.

Ran, H., Kim, N. J., and Secada, W. G. (2021). A meta-analysis on the effects of technology's functions and roles on students' mathematics achievement in K-12 classrooms. *J. Comput. Assist. Learn.* 38, 258–284. doi: 10.1111/jcal.12611

Singley, M. K., and Bennett, R. E. (2002). "Item generation and beyond: applications of schema theory to mathematics assessment," in *Item Generation for Test Development*, eds S. H. Irvine and P. C. Kyllonen (Mahwah, NJ: Lawrence Erlbaum), 361–384.

Sinharay, S. (2017). Which statistic should be used to detect item preknowledge when the set of compromised items is known? *Appl. Psychol. Meas.* 41, 403–421. doi: 10.1177/0146621617698453

Vale, C. D. (2006). "Computerized item banking," in *Handbook of Test Development*, 1st Edn, eds S. Downing and T. Haladyna (New York, NY: Routledge), 261–286.

Wendler, C. L. W., and Walker, M. E. (2016). "Practical issues in designing and maintaining multiple test forms," in *Handbook of Test Development*, 2nd Edn, eds S. Lane, M. Raymond, and T. Haladyna (New York, NY: Routledge), 433–449.

World Health Organization [WHO] (2019). *International Statistical Classification of Diseases and Related Health Problems*, 11th Edn. Geneva: World Health Organization.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.