Methods for Creating and Evaluating the Item Model Structure Used In Automatic Item Generation

Mark J. Gierl
Hollis Lai
Centre for Research in Applied Measurement and Evaluation
University of Alberta

Submitted to:
Dr. Krista Breithaupt
Director, Research and Development
Medical Council of Canada

March 31, 2012

INTRODUCTION

Automatic item generation (AIG; Drasgow, Luecht, & Bennett, 2006; Embretson & Yang, 2007; Gierl & Haladyna, in press; Irvine & Kyllonen, 2002) represents a relatively new but rapidly evolving research area where cognitive theories, computer technologies, and psychometric practices are used to generate items. In its most ambitious form, AIG can be described as the process of using models to generate statistically calibrated items with the aid of computer technology. Significant developments in AIG research and practice have occurred in the last decade, with a particularly strong wave of development occurring in the last several years. Important areas of AIG growth included cognitive model development (Gierl & Lai, in press-a; Gierl, Lai, & Turner, in press), item model development (Gierl, Alves, & Zhou, 2008; Gierl & Lai, in press-b), test design (Embretson & Yang, 2007; Huff, Alves, Pellegrino, & Kaliski, in press; Lai & Gierl, in press; Luecht, in press), statistical modeling (Embretson, 1999; Geerling, van der Linden, & Glas, 2011; Glas & van der Linden, 2003; Sinharay & Johnson, 2008; Sinharay, Johnson, & Williams, 2003), and computer technology (Gierl et al., 2008; Mortimer, Stroulia, & Yazdchi, in press).

Automatic item generation requires three general steps. First, content and test development specialists create item models that highlight the features or elements in the assessment task that can be manipulated. Second, the elements in the item model are varied to generate new items with the aid of computer-based algorithms. Third, statistical models are used to estimate the psychometric properties of the generated items based on the combination of elements used in item assembly. The focus of our study is on steps 1 and 2, item model development and item generation.

ITEM MODELS AND AUTOMATIC ITEM GENERATION

Item models provide the foundation for AIG. Item models (Bejar, 1996, 2002; Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003; LaDuca, Staples, Templeton, & Holzman, 1986) have been described using different terms, 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). Item models contains the variables in an assessment task that can be manipulated and used for generation. The elements include the stem, the options, and the auxiliary information. The stem is the part of an item model which contains the context, content, item, and/or the question the examinee is required to answer. The options include the alternative answers with one correct option and one or more incorrect options or distracters. For multiple-choice item models, both stem and options are required. For constructed-response item models, only the stem is created. Auxiliary information includes any additional content, in either the stem or option, required to generate an item. Auxiliary information can be expressed in text, images, tables, diagrams, sound, or video. The stem and options can be further divided into elements. Elements are denoted as strings, which are non-numeric content, and integers, which are numeric content.

Drasgow, Luecht, and Bennett (2006) claimed that items models can be created using either a weak or a strong theory approach. With weak theory, a combination of outcomes from research, theory, and experience provide the guidelines necessary for identifying and manipulating the elements in an item model that yield generated assessment tasks. If the goal is to pre-calibrate the generated items using statistical methods (i.e., step 3 in the three step process we described in our Introduction), then the models should be designed so they generate items that yield comparable levels of difficulty when they are administered to examinees. The weak theory approach is well suited to broad content domains where few theoretical descriptions exist on the knowledge and skills required to solve test items.

With strong theory, a cognitive model of item difficulty serves as the principled basis for identifying and manipulating those elements that yield generated items with predictable psychometric characteristics. To date, the use of strong theory AIG has focused on the psychology of specific response processes, such as spatial reasoning (Bejar, 1990) and abstract reasoning (Embretson, 2002), where articulate cognitive models of task performance exist. For most educational achievement tests, few

comparable cognitive theories exist to guide our item development practices (Leighton & Gierl, 2011) or to account for test performance in broad content areas (Schmeiser & Welch, 2006). Hence, weak theory approaches to item modeling currently prevail.

The goal of automatic generation using an item model cast within a weak theory framework is to produce new assessment tasks by manipulating a relatively small number of elements in the model. Often, the starting point is to use a parent item whose psychometric characteristics are known. The parents can be found by reviewing items from previously administered tests, by drawing on an inventory of existing test items, or by creating the parent item directly. The parent item highlights the underlying structure of the model, thereby providing a point-of-reference for creating alternative items. Then by drawing on their experiences, intuitions, theories, and luck, content specialists create the model by identifying characteristics of the parent that can be manipulated to produce new items. If the purpose of AIG is to generate statistically calibrated items, then the content specialist's task is to manipulate those elements in the parent that yield generated items with similar psychometric characteristics (i.e., item difficulty). Generated items with comparable psychometric characteristics are called isomorphs (Bejar, 2002). Alternatively, if the purpose is to generate items that are not statistically calibrated (in this case, the generated items will need to be field tested), then the content specialist's task is to manipulate those elements that yield large numbers of instances of the parent item through the generative process. These items are often called variants.

One serious drawback of using a weak theory item model is that relatively few elements can be manipulated, regardless of whether or not statistical pre-calibration is a goal. The manipulations are limited because the number of potential elements in any one item model is, typically, small. For example, if a parent item contains 20 words in the stem, then the maximum number of elements that can be manipulated is 20, assuming that all words can be made into elements. One important consequence of manipulating only a small number of element is that the generated items may be

overtly similar to one another. In our experience, this type of item modeling poses a serious problem in the current application of AIG because most content specialists view this process negatively and often refer to it pejoratively as item cloning.

Cloning, in a biological sense, refers to any process where a population of identical units is derived from the same ancestral line. Cloning helps characterize weak theory item modeling if we consider it to be a process where specific content (e.g., nuclear DNA) in a parent item (e.g., currently or previously existing animal) is manipulated to generate a new item (e.g., new animal). Through this process, instances are created that are identical (or, at least, very similar) to the parent because information is purposefully transferred from the parent to the offspring. Unfortunately, current approaches to generating isomorphic tasks from a weak theory item model yield outcomes that are described by content specialists and test developers as “clones”, “ghost” items, or “Franken-items”. Clones are perceived by content specialists to be generated items that are easy to produce, unlike more traditional items. Clones are often seen as a simplistic product from an overly simple item development process, compared to a more sophisticated traditional test item which is a complex product from a more sophisticated item development process. Most importantly, clones are believed to be easily recognized by coaching and test preparation companies which limits their usefulness in operational testing programs. In short, items generated from weak theory item models are viewed by many content specialists as easily produced, overly simplistic, and clearly detectable. As a result, content specialists are rarely impressed with items produced from weak theory models, particularly when the underlying model is thought to be discernible through the generated items. From our summary, it is also easy to understand why AIG researchers are not warmly received in most test development committees.

PURPOSE OF STUDY

The purpose of our study is to introduce and illustrate a new method for generating assessment tasks using the n-layer item model. The n-layer model serves as a generalization of the current item modeling approach by permitting a relatively larger number of elements to be manipulated during generation. As a result, the generated items are more heterogeneous and, therefore, less susceptible to the label "item cloning". We also describe a measure for evaluating item similarity and we demonstrate how this measure can be used to describe item models.

n-LAYER ITEM MODELING: A GENERALIZED APPROACH FOR STRUCTURING AND GENERATING TEST ITEMS

Recall, the goal of automatic generation using a weak theory item model is to produce new assessment tasks by manipulating a relatively small number of elements at one layer in a parent model. This approach will now be referred to as *1-layer item modeling*. A generalization of the 1-layer item model is the *n-layer item model*. The goal of automatic generation using the n-layer model is to generate items by manipulating a relatively large number of elements at two or more layers in a parent model. Much like the 1-layer item model, the starting point for the n-layer model is to use a parent item. But unlike the 1-layer model where the manipulations are constrained to a linear set of generative operations using a small number of elements at a single level, the n-layer model permits manipulations of a nonlinear set of generative operations using elements at multiple levels. As a result, the generative capacity of the n-layer model is substantially increased.

The concept of n-layer item generation is adapted from the literature on syntactic structures of language where researchers have reported that sentences are typically organized in a hierarchical manner (e.g., Higgins, Futagi, & Deane, 2005). This hierarchical organization, where elements are embedded within one another, can also be used as a guiding principle to generate large numbers of meaningful test items. The use of an n-layer item model is therefore a flexible template for expressing different syntactic structures thereby permitting the development of many different but feasible

combinations of embedded elements. The n-layer structure can be described as a model with multiple layers of elements, where each element can be varied simultaneously at different levels to produce different items. In the computational linguistic literature, our n-layer structure could be characterized as a generalized form of template-based natural language generation, as described by Reiter (1995).

A comparison of the 1- and n-layer item model is presented in Figure 1. For this example, the 1-layer model can provide a maximum of four different values for element A. Conversely, the n-layer model can provide up to 64 different values using the same four values for elements C and D embedded within element B. Because the maximum generative capacity of an item model is the product of the ranges in each element (Lai, Gierl, & Alves, 2010), the use of an n-layer item model will always increase the number of items that can be generated relative to the 1-layer structure.

One important benefit of using the n-layer item structure is that more elements can be manipulated within the model simultaneously resulting in generated items that appear to be different from one another. Hence, n-layer item modeling can be used to directly address the problem of cloning that concerns many content specialists. The drawback of using an n-layer structure is that the models are challenging to create given the complexity of combining elements in an embedded fashion. Also, the effect of embedding elements in multiple levels, while useful for generating large numbers of items, may make it challenging to predict the psychometric characteristics of the generated items. Hence, n-layer item modeling may yield items that are not possible to pre-calibrate and, therefore, will need to be field tested using conventional administration procedures.

COSINE SIMILARITY INDEX

To measure and compare the similarity of items created using a 1- and an n-layer model, the intra-model differences, meaning items generated within the same model, must be evaluated. Because fewer elements are manipulated with the 1-layer approach, similarity should be higher for items generated with this model compared with the n-layer model. Similarity is quantified in the current study using the

cosine similarity index (CSI). The CSI is a measure of similarity between two vectors of co-occurring texts. It is computed using the cosine of the angle between the two vectors in a multidimensional space of unique words. The CSI can be expressed as

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|},$$

where A and B are two items expressed in a binary vector of word occurrences. For example, if A is a list of three words (e.g., dog, walk, talk) and B is a list of three words (e.g., cat, walk, mock), then the length of both binary vectors is the number of unique words used across both lists (i.e., dog, walk, talk, cat, mock). To vectorize A and B so the words can be compared, the occurrence of each word in the corresponding list is quantified with a value of 1. The resulting vectors for A and B in our example are [1,1,1,0,0] and [0,1,0,1,1]. The CSI has a minimum value of 0, meaning that no word overlapped between the two vectors, and a maximum of 1, meaning that the text represented by the two vectors are identical. The CSI is one among many types of word comparison methods. The lack of a word net or common corpus in the content area of medicine (which is one of the content areas used in the current study) limits the use of semantic distances (see Belov & Knezevich, 2008) and n-gram comparison is not appropriate for short text lengths such as test items (Lin, 2004). As a result, the CSI was selected as our measure of similarity.

METHODS

Item Model Development

To compare the word similarity of items created using a 1- and an n-layer model, items were generated in two contrasting content areas. The first content area is SAT Mathematics. An example of a 1-layer SAT Mathematics item model is presented in Figure 2. This example was developed by three College Board Mathematics content specialists. The item model is used to generate Number and Operations word problems that measure the concepts of ratio, proportion, and rate. The stem contains

three integers elements (I1 to I3). An example of an n-layer SAT Mathematics item model is presented in Figure 3. This item model serve as a generalization of the example presented in Figure 2. Our n-layer item model example, in addition to manipulating the integer values, now includes two string values (i.e., Occupation; Task) where both the integers and strings are embedded within one another to facilitate the generative process. That is, by embedding elements within elements, different occupations and tasks can be combined with the integer values initially presented in the 1-layer model to generate more diverse and heterogeneous items.

The second content area is surgery from a medical licensure test. An example of a 1-layer surgical item model is presented in Figure 4. This example was created by two Medical Council of Canada surgery content specialists. The item model in Figure 4 requires diagnosing complications associated with hernias. The stem contains one integer (Age) and six strings (Gender; Pain; Location; Acuity of Onset; Physical Findings; White Blood Cell count). An example of a n-layer surgical item model is presented in Figure 5. This item model serve as a generalization of the example presented in Figure 4. Our n-layer item model example includes integers and strings embedded within one another to facilitate the generative process. That is, by embedding elements within different situations, test findings, and question prompts, more heterogeneous surgery items can be generated.

Item Generation

After the four models were created, items were generated using IGOR (Gierl, Zhou, & Alves, 2008). IGOR, the acronym for Item GeneratOR, is a software program written in JAVA that instantiates all possible combinations of elements into items based on the definitions within the model. To generate items, models need to be expressed in an XML format that IGOR can interpret. Once a model is expressed in an XML form, IGOR computes the necessary information and outputs items in either a HTML or a Word format.

Measuring Similarity Among Generated Items

After generation, the similarity of the assessment tasks can be analyzed and compared. The hypothesis of our study is that the intra-item model differences should be higher with n-layer modeling. After the items were generated, a porter stemming algorithm was used. Porter stemming is a process for removing common morphological (e.g., a, is, this) and inflexional endings (e.g., bladders into bladder) so words can be compared more directly. Then, to compare the word similarity among the generated items, a sample of 100 items for each model was randomly selected and analyzed.

To compute the CSI, the items were compiled into a matrix of word occurrences for each item model, where each row represents a vector of a generated item, each column represents a unique word in the pool of generated items, and each row-by-column cell is numerated dichotomously to determine whether a given item contains a given word. The CSI was calculated for each unique item pair within the same item model. The outcome is a CSI mean and standard deviation for each model that, in turn, can be compared between models.

RESULTS

Item Generation

IGOR generated 177 and 3,906 items for the 1– and n-layer item model in SAT Mathematics, respectively. A random sample of four items from each model is presented in Table 1. IGOR generated 256 and 12,287 items for the 1- and n-layer models in surgery, respectively. A random sample of four items from each model is presented in Table 2.

Item Word Similarity

From the 100 randomly-selected items generated for each item model, a total of 4,950 pair-wise comparisons were conducted. The CSI was calculated for each unique pair of items. The summary statistics are presented in Table 3. For SAT Mathematics, the 1-layer model produced more similar items than the n-layer model. The CSI values ranged from 0.77 to 1.00 for the 1-layer model, with a high

overall mean of 0.87 and a low standard deviation of 0.05 indicating that the generated items are quite similar and homogeneous. By comparison, the n-layer model produced CSI values ranging from 0.00 to 1.00, with a relatively low mean of 0.37 and a high standard deviation of 0.27. When the CSI means were compared between the 1- and n-layer models, the independent samples *t*-test produced statistically different results, $t(9004) = 126.05$, $p < 0.05$.

For Surgery, the 1-layer model, again, produced more similar items than the n-layer model. The CSI values ranged from 0.55 to 0.98 for the 1-layer model, with a high overall mean of 0.74 and a low standard deviation of 0.11 indicating that the items are quite similar and homogeneous. The n-layer model produced CSI values ranging from 0.17 to 1.00, with a comparatively low mean of 0.53 and a higher standard deviation of 0.16. When the CSI means were compared between the items generated with the 1- and n-layer models, the independent samples *t*-test produced statistically different results, $t(9004) = 77.18$, $p < 0.05$. Taken together, these results allow us to conclude that the n-layer item models do, in fact, produce more heterogeneous and diverse items compared to the items generated from 1-layer item models using a measure of word similarity.

CONCLUSIONS AND DISCUSSION

Modern testing programs require large numbers of high-quality items that are produced in both a timely and cost-effective manner. One approach that may help address these challenges is through automatic item generation. Automatic item generation requires three general steps. First, content specialists create item models that specify the elements in the assessment task that must be manipulated. Second, the elements in the model are manipulated with computer-based algorithms to generate new items. Third, statistical models are used to estimate the psychometric properties of the generated items. *With this three-step process, hundreds or even thousands of new items can be created from a single item model.* Not surprising, automatic item generation is seen by many executives and managers in testing agencies as a “dream come true”, given the laborious processes and high costs

required for traditional item development. Unfortunately, many content specialists are not so enamored by this dream because they find the quality of the generated items is still lacking.

Hence, this study was motivated by our desire to improve the quality of generated items, in light of our interactions and discussions with content specialists. Simply put, test development specialists and content experts dislike item cloning. Biological cloning could serve as an analogy for 1-layer item modeling, particularly when the generated items are designed to emulate the statistical properties of the parent. While item cloning has an important role to play in some AIG research (e.g., Embretson, 1999; Geerling, van der Linden, & Glas, 2011; Glas & van der Linden, 2003; Sinharay & Johnson, 2008; Sinharay, Johnson, & Williams, 2003), it is also important to recognize that these types of generated items (i.e., clones, ghost items, Franken-items) may have limited value in operational testing programs, according to many content specialists, because they are deemed to be easily produced, overly simplistic, and readily detectable.

In the current study, we generated items for both mathematics and surgery using two different item modeling approaches. The first approach, we called 1-layer item modeling, generated tasks by manipulating a relatively small number of elements in the parent. This represents the current standard-of-practice in most AIG applications. We also introduced a second approach called n-layer item modeling where elements are embedded and manipulated in different layers to generate tasks. This represents a new approach to AIG where more diverse and heterogeneous items are generated.

Directions for Future Research

The n-layer model is a flexible structure for item generation thereby permitting many different but feasible combinations of embedded elements. It can be used with any type of template-based item generation method. It can be used to generate different item types. And, as was illustrated in our study, the n-layer models can accommodate a wide range of elements found in diverse content areas. In addition to generating more diverse and heterogeneous items, one possible application of n-layer

modeling may be in generating multi-lingual test items. Different languages require a different grammatical structure and word order (Higgins, Futagi, & Deane, 2005). With a 1-layer model, the grammatical structure and word order cannot be easily or readily manipulated because the generative operations are constrained to a linear set using a small number elements at a single level. However, with the use of an n-layer model, the generative operations are expanded dramatically to include non-linear sets using a large number of elements at multiple levels. Language, therefore, can serve as an additional level or layer that is manipulated during item generation. Table 4 contains a random sample four items selected from a set of 12,287 items that was generated with the surgery item model in Figure 5, except that Spanish was added as a language layer in the model. The embedded element structure used to generate the items in Table 4 is shown in Figure 6. One important direction for future research, then, is to use n-layer item modeling to generate tasks in multiple languages by adding language as an additional layer in the model.

Our research can also be used to support AIG methods currently focused on item cloning. We demonstrated how the CSI can serve as a measure for summarizing item similarity. But item models, like items, require descriptive measures for their proper use. Sinharay et al. (2003) and Sinharay and Johnson (2008), for example, used the concept of item families to develop a statistical model for calibrating generated items. The siblings (i.e., generated items) in their statistical model must share some common features to be considered part of the family. To-date, however, there are no empirical methods available for quantifying commonality and, hence, sibling membership must be established more subjectively using judgements and ratings from content reviews. As an empirical measure of sibling commonality, the mean CSI among the generated item pairs within an item model could be used to describe the similarity among the generated items. The outcomes from the CSI could then be used as evidence to decide whether an item model has generated clones or isomorphs (i.e., high mean CSI and low standard deviations) or whether it has generated variants (i.e., low mean CSI and high standard

deviation). Hence, future studies could also be conducted to evaluate the effectiveness of using the CSI for establishing family membership for calibration methods that require item families.

REFERENCES

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14,* 237-245.

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

Belov, D., & Knezevich, L. (October, 2008*). Automatic prediction of item difficulty based on semantic measures.* Research Report, 08-04. Law School Admissions Council, Newtown, PA.

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. (1999). Generating items during testing: Psychometric issues and models. *Psychometrika, 64,* 407-433.

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: Erlbaum.

Embretson, S. E., & Yang, X. (2007). Automatic item generation and cognitive psychology. In C. R. Rao & S. Sinharay (Eds.) *Handbook of Statistics: Psychometrics, Volume 26* (pp. 747-768). North Holland, UK: Elsevier.

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

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

Gierl, M. J., & Lai, H. (in press-a). Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples. In M. J. Gierl & T. Haladyna (Eds.),). *Automatic item generation: Theory and practice.* New York: Routledge.

Gierl, M. J., & Lai, H. (in press-b). Using item models for automatic item generation. *International Journal of Testing.*

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

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).* Retrieved [date] from http://www.jtla.org.

Glas, C. A. W., & van der Linden, W. J. (2003). Computerized adaptive testing with item cloning. *Applied Psychological Measurement, 27,* 247-261.

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., 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.

Huff, K., Alves, C., Pellegrino, J. & Kaliski, P. (in press). Using evidence-centered design task models in automatic item generation. In M. J. Gierl & T. Haladyna (Eds.),). *Automatic item generation: Theory and practice.* New York: Routledge.

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

Lai, H., & Gierl, M. J. (in press). Generating items under the assessment engineering framework. In M. J. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and practice.* New York: Routledge.

Lai, J., Gierl. M. J., & Alves, C. (2010, April). *Using item templates and automated item generation principles for assessment engineering.* In R. M. Luecht (Chair), Application of assessment engineering to multidimensional diagnostic testing in an educational setting. Paper presented at the annual meeting of the National Council on Measurement in Education, Denver, CO.

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

Lin, C., (June, 2004). *ROUGE: A package for automatic evaluation of summaries.* Proceedings of the ACL-04 Workshop. Philadelphia, PA.

Luecht, R. (in press). An introduction to assessment engineering for automatic item generation. In M. J. Gierl & T. Haladyna (Eds.),). *Automatic item generation: Theory and practice.* New York: 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: Erlbaum.

Mortimer, T., Stroulia, E., & Yazdchi, Y. (in press). IGOR: A web-based item generation tool. In M. J. Gierl & T. Haladyna (Eds.),). *Automatic item generation: Theory and practice.* New York: Routledge.

Reiter, E. (1995). *NLG vs. templates.* Proceedings of the Fifth European Workshop on Natural Language Generation (pp. 95-105). Leiden, The Netherlands.

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.

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: Erlbaum.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing, 8,* 209-236.

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,* 295-313.

Table 1.

A Random Sample of Five Generated Items from the 1- and n-layer Item Model in SAT Mathematics

1-Layer Item Model

38. Yesterday a veterinarian treated 3 mice, 6 cats, 5 dogs, and no other animals. What was the ratio of the number of cats treated to the total number of animals treated by the veterinarian?

a. 1 to 6
b. 5 to 8
c. 1 to 11
d. 6 to 14*

55. Yesterday a veterinarian treated 4 mice, 5 cats, 2 dogs, and no other animals. What was the ratio of the number of cats treated to the total number of animals treated by the veterinarian?

a. 5 to 11*
b. 1 to 7
c. 2 to 6
d. 1 to 2

139. Yesterday a veterinarian treated 7 mice, 5 cats, 8 dogs, and no other animals. What was the ratio of the number of cats treated to the total number of animals treated by the veterinarian?

a. 1 to 14
b. 1 to 8
c. 8 to 15
d. 5 to 20*

26. Yesterday a veterinarian treated 3 mice, 2 cats, 5 dogs, and no other animals. What was the ratio of the number of cats treated to the total number of animals treated by the veterinarian?

a. 5 to 8
b. 1 to 7
c. 1 to 5*
d. 1 to 6

*-correct option

n-Layer Item Model

1094. Last week a mechanic fixed, 7 transmissions, 5 tires and 8 brakes. What is the total number of cars the mechanic fixed during that time?

a. 15
b. 20*
c. 11
d. 14

3300. Yesterday a veterinarian treated, 4 mice and 8 cats. What is the number of mice compared to the total number of animals the veterinarian treated in that period?

a. 1 to 15
b. 4 to 12*
c. 12 to 19
d. 1 to 7

1419. In the past three days, a doctor diagnosed 5 colds, 2 fevers and 6 infections. What is the number of colds the doctor diagnosed in that given time period?

a. 1 to 6
b. 1 to 10
c. 5 to 13*
d. 1 to 8

744. Yesterday a mechanic fixed 5 transmissions, 7 tires and 6 brakes. What is the number of transmissions and tires the mechanic fixed in that period?

a. 12 to 18*
b. 1 to 6
c. 1 to 10
d. 1 to 13

*-correct option

Table 2.

A Random Sample of Five Generated Items from the 1- and n-layer Item Model in Surgery

1-Layer Item Model

11. A 35-year-old woman presented with a mass in the left groin. It occurred a few months ago. On examination, the mass is protruding but with no pain and lab work came back with normal results. Which of the following is the next best step?

a. ice applied to mass*
b. exploratory surgery
c. reduction of mass
d. hernia repair

50. A 30-year-old man presented with a mass in an area near a recent surgery. It occurred a few months ago. On examination, the mass is protruding but with no pain and lab work came back with normal results. Which of the following is the next best step?

a. ice applied to mass*
b. exploratory surgery
c. reduction of mass
d. hernia repair

175. A 55-year-old woman presented with a mass and severe pain in the umbilicus. It occurred a few days ago. On examination, the mass is tender and exhibiting redness and lab work came back with elevated white blood cell count. Which of the following is the next best step?

a. ice applied to mass
b. exploratory surgery
c. reduction of mass
d. hernia repair*

137. A 25-year-old woman presented with a mass and severe pain in the left groin. It occurred a few days ago. On examination, the mass is tender and exhibiting redness and lab work came back with elevated white blood cell count. Which of the following is the next best step?

a. ice applied to mass
b. exploratory surgery
c. reduction of mass
d. hernia repair*

*-correct option

n-Layer Item Model

5326. A 50-year-old man presented with a mass and mild pain in the left groin. It occurred a few days ago after moving a piano. Upon further examination, the patient had normal vitals and the mass is tender and reducible. Which one of the following is the best treatment?

a. exploratory surgery
b. reduction of mass
c. hernia repair*
d. ice applied to mass

4610. Patient complaints of a mass in the left groin which has been a problem since a few months ago. On examination, the mass is protruding but with no pain and lab work came back with normal vitals. Which one of the following is the best treatment?

a. reduction of mass
b. exploratory surgery
c. hernia repair
d. ice applied to mass*

12010. Patient complaints of a mass and mild pain in the umbilicus which has been a problem since a few days ago after moving a piano. There is tender and reducible in the umbilicus and the patient had normal vitals. Given this information, what is the best course of action?

a. exploratory surgery
b. ice applied to mass
c. reduction of mass*
d. hernia repair

7325. A 45-year-old man presented with a mass and severe pain in right groin. It occurred a few days ago. There is tender and exhibiting redness in the right groin and the patient had elevated white blood cell count. Which one of the following is the best treatment?

a. reduction of mass
b. hernia repair*
c. ice applied to mass
d. exploratory surgery

*-correct option

Table 3.

Summary of Cosine Similarity Index as a Function of Content Area and Model

	Min	Max	Mean	SD
Mathematics				
1-layer	0.77	1.00	0.87	0.05
N-layer	0.00	1.00	0.37	0.27
Surgery				
1-layer	0.55	0.98	0.74	0.11
N-layer	0.17	1.00	0.53	0.16

Table 4.

A Random Sample of Four Generated Spanish Items Using the n-layer Model in Surgery with Language as a Layer

684546. Un varon de 55 años presenta una masa en la ingle izquierda hace unos meses. Tras un examen posterior, el paciente presenta signos vitales normales y la masa es es sobresaliente pero sin dolor. ¿Cuál de las siguientes es la mejor opción?

a. aplicar hielo sobre la masa
b. cirugía exploratoria
c. reducir la masa*
d. reparación de hernia

684553. Un varon de 25 años presenta una masa en la ingle izquierda hace unos meses. Tras un examen posterior, el paciente presenta signos vitales normales y la masa es es sobresaliente pero sin dolor. ¿Cuál de las siguientes es la mejor opción?

a. aplicar hielo sobre la masa
b. cirugía exploratoria
c. reducir la masa*
d. reparación de hernia

768953. La paciente presenta una masa en la ingle izquierda. El paciente es una mujer de 40 años. Hay es sobresaliente pero sin dolor en la ingle izquierda y el paciente presenta signos vitales normales. ¿Cuál de las siguientes es la mejor opción?

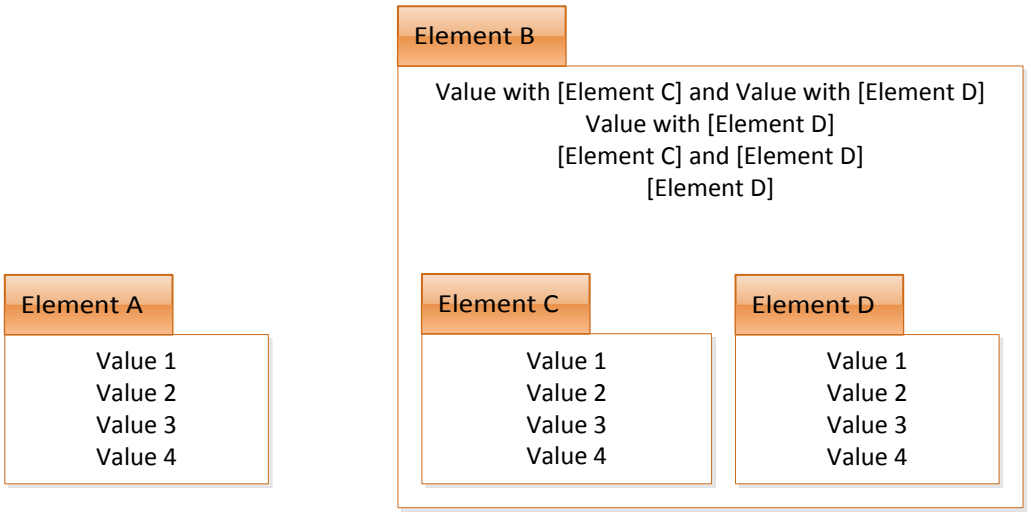
a. aplicar hielo sobre la masa*
b. cirugía exploratoria
c. reducir la masa
d. reparación de hernia

805775. Una mujer de 40 años presenta una masa en la ingle izquierda hace unos meses. Aunque el paciente presenta signos vitales normales y es sobresaliente pero sin dolor en la zona afectada, el paciente no presenta otras complicaciones. ¿Cuál es el mejor paso a seguir?

a. aplicar hielo sobre la masa
b. cirugía exploratoria
c. reducir la masa
d. reparación de hernia*

*-correct option

Figure 1. An comparison of the elements in a 1-layer and n-layer item model.



Example of a 1-layer element

Example of an n-layer element, with two layers

Figure 2. One-layer SAT Mathematics item model in Number and Operations.

Parent Item:

Yesterday a veterinarian treated 2 mice, 3 cats, 6 dogs, and no other animals. What was the ratio of the number of cats treated to the total number of animals treated by the veterinarian?

(A) 1 to 4
(B) 1 to 6
(C) 1 to 13
(D) 3 to 8
(E) 3 to 11

Item Model:

Stem	Yesterday a veterinarian treated [I1] mice, [I2] cats, [I3] dogs, and no other animals. What was the ratio of the number of cats treated to the total number of animals treated by the veterinarian?
Elements	[I1] Range: 2 to 8 by 1 [I2] Range: 2 to 8 by 1 [I3] Range: 2 to 8 by 1
Options	(A) [I2] to \[[I1] + [I2] + [I3] \] (B) 1 to \[2* [I1] \] (C) 1 to [I3] (D) [I3] to \[[I3] + [I1] \] (E) 1 to \[[I2] + [I3] \]
Key	(A)

Figure 3. N-layer SAT Mathematics item model in Number and Operations.

Item Model:

Stem	[Occupation statement], [Tasks Involved]. What is the [Problem]?

Elements:

Layer 1	[Occupation statement]: Last week a mechanic fixed, The doctor diagnosed, Yesterday a veterinarian treated [Task involved]: [I1] [jobA] and [I2] [jobB].,[I1] [jobA] , [I2] [jobB] and [I3] [jobC].,[I1] patients with [jobA] and [I2] patients with [jobB]. (only with doctors) [Problem]: number of [jobA] the [Occupation] [task] today?, number of [jobA] and [jobB] the [Occupation] [task] today?, total number of [Jobs] the [Occupation] [task] today? [IKey] Range: [I1],[I2],[I3]
Layer 2	[I1] Range: 2 to 8 by 1 [I2] Range: 2 to 8 by 1 [I3] Range: 2 to 8 by 1 [IKey] Range: [I1],[I2],[I3] [Occupation]: mechanic, doctor, veterinarian [jobA]: transmissions, colds, mice [jobB]: alignments, fevers, cats [jobC]: brakes, infections, dogs [task]: fixed, diagnosed, treated [Jobs]: cars, patients, animals
Options	(A) [IKey] to \[[I1] + [I2] + [I3] \] (B) 1 to \[2* [I1] \] (C) 1 to [I3] (D) [IKey] to \[[I3] + [I1] \] (E) 1 to \[[I2] + [I3] \]
Key	(A)

Figure 4. One-layer surgery item model for generating surgery test items.

Parent Item:

A 24-year-old man presented with a mass in his left groin. It appeared suddenly 2 hours ago while lifting a piano. On examination he has a tender firm mass in the left groin. Which one of the following is the next best step?

(A) Immediate hernia repair
(B) Needle aspiration
(C) Ice packs to groin
(D) Reduction of mass
(E) Ultrasound of groin

Item Model:

Stem

A [AGE]-year-old [GENDER] presented with a mass [PAIN] in [LOCATION]. It occurred [ACUITYOFONSET]. On examination, the mass is [PHYSICALFINDINGS] and lab work came back with [WBC]. Which of the following is the next best step?

Elements

[AGE] (Integer): From 25.0 to 60.0, by 5.0

[GENDER] (String): 1: man 2: woman

[PAIN] (String): 1: 2: and intense pain 3: and severe pain 4: and mild pain

[LOCATION] (String): 1: the left groin 2: right groin 3: the umbilicus 4: an area near a recent surgery

[ACUITYOFONSET] (String): 1: a few months ago 2: a few hours ago 3: a few days ago 4: a few days ago after moving a piano

[PHYSICALFINDINGS] (String): 1: protruding but with no pain 2: tender 3: tender and exhibiting redness 4: tender and reducible

[WBC] (String): 1: normal results 2: normal results 3: elevated white blood cell count 4: normal results

Options

exploratory surgery; reduction of mass; hernia repair; ice applied to mass

Key

Conditional

Figure 5. N-layer surgery item model for generating surgery test items.

Item Model:

Stem	[Situation TestFindings QuestionPrompt]

Elements:

Layer 1	QuestionPrompt (Text): 1: What is the best next step? 2: Which one of the following is the best prognosis? 3: Given this information, what is the best course of action? TestFindings (Text): 1: On examination, the mass is [[PhysicalFindings]] and lab work came back with [[WBC]]. 2: Upon further examination, the patient had [[WBC]] and the mass is [[PhysicalFindings]]. 3: With [[WBC]] and [[PhysicalFindings]] in the area, the patient is otherwise nominal. 4: There is [[PhysicalFindings]] in the [[Location]] and the patient had [[WBC]]. Situation (Text): 1: A [[AGE]]-year-old [[Gender]] presented with a mass [[Pain]] in [[Location]]. It occured [[AcuityofOnset]]. 2: Patient presents with a mass [[Pain]] in [[Location]] from [[AcuityofOnset]]. The patient is a [[AGE]]-year-old [[Gender]]. 3: Patient complaints of a mass [[Pain]] in [[Location]] which has been a problem since [[AcuityofOnset]]. 4: A [[Gender]] was admitted with pain in the [[Location]] from [[AcuityofOnset]].
Layer 2	[AGE] (Integer): From 25.0 to 60.0, by 5.0 [GENDER] (String): 1: man 2: woman [PAIN] (String): 1: 2: and intense pain 3: and severe pain 4: and mild pain [LOCATION] (String): 1: the left groin 2: right groin 3: the umbilicus 4: an area near a recent surgery [ACUITYOFONSET] (String): 1: a few months ago 2: a few hours ago 3: a few days ago 4: a few days ago after moving a piano [PHYSICALFINDINGS] (String): 1: protruding but with no pain 2: tender 3: tender and exhibiting redness 4: tender and reducible [WBC] (String): 1: normal results 2: normal results 3: elevated white blood cell count 4: normal results

Options	exploratory surgery; reduction of mass; hernia repair; ice applied to mass

Key	Conditional

Figure 6. N-layer surgery item model with language as a layer.