Using Automated Processes to Generate Test Items And Their Associated Solutions and Rationales to Support Formative Feedback

Mark J. Gierl[1], Hollis Lai[2]

[1]Faculty of Education, University of Alberta, Edmonton, AB, Canada T6G 2G5
[2]School of Dentistry, University of Alberta, Edmonton, AB, Canada T6G 2G5
{gierl, mark}@ualberta.ca

Abstract. Automatic item generation is the process of using item models to produce assessment tasks using computer technology. An item model is similar to a template that highlights the elements in the task that must be manipulated to produce new items. The purpose of our study is to describe an innovative method for generating large numbers of diverse and heterogeneous items along with their solutions and associated rationales to support formative feedback. We demonstrate the method by generating items in two diverse content areas, mathematics and nonverbal reasoning.

Keywords: Automatic item generation; Item writing; Formative assessment; Technology-enhanced assessment

1 Introduction

Profound global and economic changes are shaping how we develop and deliver educational tests. These changes can be attributed to the growing emphasis on knowledge services, information, and communication technologies. To thrive in this new environment, countries require skilled workers who can think, reason, solve complex problems as well as quickly adapt to novel situations, communicate, and collaborate. Educational tests, once developed almost exclusively to satisfy demands for accountability and outcomes-based summative assessment, are now expected to provide teachers and students with timely, detailed, formative feedback to directly support the teaching and learning of these new 21st century skills [1]. To meet these teaching and learning directives, formative assessment principles are beginning to guide our educational testing practices. Formative principles include any assessment-related activities—including administering tests more frequently—that yield constant and specific feedback to modify teaching and improve learning. But when testing occurs frequently, more tests are required and these tests must be created both efficiently and economically. Fortunately, this requirement for frequent and timely educational testing coincides with the dramatic changes occurring in educational technology. Developers of local, national, and international educational tests are now implementing internetbased computerized tests at an extraordinary rate [2]. Computerized testing offers many important benefits to support and promote key principles in formative assessment. For instance, computers permit testing on-demand thereby allowing students to take the test at any time during instruction and as often as they choose; items on computerized tests are scored immediately thereby providing students with instant and detailed feedback; computers support the development of multimedia item types that allows educators to measure more complex performances as well as a broader variety of knowledge and skills. In short, computers are helping educators to infuse their formative testing principles into their assessment practices to support teaching and learning.

Despite these important benefits, the advent of computerized educational testing has also raised formidable challenges, particularly in the area of test item development [3]. Hundreds or even thousands of new test items are needed to create the banks necessary for computerized testing because items are continuously administered and, therefore, exposed to students. A bank is a repository of test items, which includes both the individual items and data about their content and psychometric characteristics (e.g., difficulty level of the item) as well as usage (e.g., item exposure rate). These banks must be frequently replenished with new test items to ensure that students receive a continuous supply of unique, content-specific, items while, at the same time, limiting item exposure within the testing environment to maintain security so the testing process is fair for all students. Unfortunately, educational test items, as they are currently produced, are expensive and time consuming to create because *each individual item* is developed, initially, by a content specialist and, later, reviewed, edited, and revised by committees of content specialists to ensure the item yields reliable and valid information. Because educators are now faced with the daunting task of creating large numbers of new items for computerized tests, alternative methods of item development are needed. One method that may be used to address this challenge is through *automatic item generation* (AIG) [4], [5], [6].

AIG is a rapidly evolving research area where cognitive theories, computer technologies, and psychometric practices establish a process that can be used to generate test items. AIG can be described as the process of using models to generate items with the aid of computer technology. It requires two general steps. First, content specialists create item models that highlight the elements in the assessment task that can be manipulated. An item model is similar to a template that specifies the variables or elements in the task that must be manipulated to produce new items. Second, the elements in the item model are varied using computer-based algorithms to generate new items. The purpose of this study is to describe and illustrate a method where one item model can be used to generate many test items. We also present an innovative new method associated with the AIG process were the solution and rationale for each item is also produced as part of the generative process. Hence, the AIG method we present provides a way to generate test items along with the solutions and rationales thereby providing educators with a method for creating large numbers of test items and the feedback associated with these items to support formative feedback systems. The method will be applied in two diverse content areas—K-12 mathematics and nonverbal reasoning—to demonstrate the feasibility and generalizability of our method.

2 Overview of Automatic Item Generation

Item modeling provides the foundation for AIG [7], [8]. An item model is comparable to a template, mould, or rendering that highlights the elements in an assessment task that must be manipulated to produce new items. Elements can be found in the stem, the options, and/or the auxiliary information. The stem is the part of an item model that contains the context, content, and/or the question the student is required to answer. Options are the alternative answers that include one correct and one or more incorrect option. Auxiliary information includes any additional content, in either the stem or option, required to generate an item. Auxiliary information can be expressed as images, tables, diagrams, sound, or video. The stem and options are further divided into elements. Elements are denoted as strings which are non-numeric and integers which are numeric.

Content specialists play a critical role in AIG. They engage in the creative task of developing item models using design guidelines discerned from a combination of experience, theory, and research [3]. Often, the starting point is to use an existing test item. Existing items, also called parent items, can be found by reviewing previously administered tests, by drawing on existing items from a bank, or by creating the parent item directly. The parent item highlights the structure of the model thereby providing a point-of-reference for creating alternative items. Then, content specialists identify elements in the parent that can be manipulated to produce new items. They also specify the content (i.e., string and integer values) for these elements. When a relatively small number of elements are manipulated, the generated items may appear comparable to one another. Generated items that are homogeneous are often called clones. Conversely, when a relatively large number of elements are manipulated, the generated items may be different from one another. Generated items that are heterogeneous are called variants.

To illustrate these concepts, a parent item from an Grade 6 mathematics achievement test is shown in Figure 1. This simple example is presented for the purpose of illustrating the basic concepts that underlie AIG. It will also be used throughout our study to demonstrate key concepts. The stem in this example contains two numeric elements (E1, E2). The E1 element includes Ann’s payment. It ranges from \$1525 to \$1675 in increments of \$75. The E2 element includes the size of the lawn, as either 30/m^2 or 45/m^2. The options, labelled A to D, are generated using formulas that include the integer values E1 and E2.

Once the item model is developed by content specialists, AIG can begin. AIG is a way of processing item models using computer technology to generate test items. The role of the content specialist is critical for the creative task of designing and developing meaningful item models. The role of computer technology is critical for the algorithmic task of systematically combining elements specified by the content specialists in each model to produce items. If we return to the math example in Figure 1, the logic underlying the generative process can be demonstrated. The generative task for this example involves producing six new items with the following E1, E2 combinations: E1=\$1525 and E2=30/m^2; E1=\$1600 and E2=30/m^2; E1=\$1675 and E2=30/m^2; E1=\$1525 and E2=45/m^2; E1=\$1600 and E2=45/m^2; E1=\$1675 and E2=45/m^2. The items generated from this example would be expected to have the same difficulty level and measure the same underlying mathematical construct. Gierl, Zhou, and Alves [8] developed a JAVA-based computer program called IGOR (which stands for **I**tem **G**enerat**OR)** that automatically generates items using the string and integer combinations specified in the item model. IGOR, research software developed by Gierl et al., is just one of many linear programming method can be used to solve the type of combinatorial problem found within AIG.

Parent Item	Ann has paid $1525 for planting her lawn. The cost of lawn is $45/m2. Given the shape of her lawn is square, what is the side length of Ann's lawn? A. 5.8 B. 6.8 C. 4.8 D. 7.3
Stem	Ann has paid $E1 for planting her lawn. The cost of lawn is $E2/m^2. Given the shape of her lawn is square, what is the side length of Ann's lawn?
Elements	E1 Value Range: 1525-1675 by 75 E2 Value Range: 30 or 45
Options	[A] = $\sqrt{E1/_{E2}}$ * *-denotes correct option [B]= $\sqrt{E1/_{E2}} + 1$ [C]= $\sqrt{E1/_{E2}} - 1$ [D]= $\sqrt{E1/_{E2}} + 1.5$

Fig. 1. A simple 2-element mathematics item model.

3 Method

The methods are described in four sections. First, we describe the mathematics and the nonverbal reasoning item models. Second, we present the procedures used to implement item modeling. Third, we explain how formative rationales can be incorporated into the generation process. Fourth, we summarize the generated outcomes by describing the IGOR software program.

3.1 Mathematics and Nonverbal Reasoning Item Models

Item models were developed by content specialists in two different areas. The first content area is mathematics. We used the simple example in Figure 1. To solve this problem, the formula for the side length of a square is provided in option A. The incorrect responses use the same formula for the side length except each option includes a computational error: option B adds 1, option C subtracts 1, and option D adds 1.5.

The second content area is nonverbal reasoning. The nonverbal reasoning item format called "middle of the sequence" was used. To solve this item type, students are required to reorder five figures to form the most logical sequence. Then, they select the alternative that is in the middle of the sequence. This task is based on sequences of shapes designed to assess students' ability to reason in the abstract and to solve problems in non-verbal contexts. An example of a middle of the sequence nonverbal reasoning item is shown in Figure 2. To solve this item, students are first required to rotate the subfigure from each corner or vertex of the triangle to the middle position in the base image. Then, students are required to identify the most systematic order for the figures so the middle of the sequence can be specified. For this example, the order follows a clockwise rotation beginning in the bottom left corner of the triangle. Therefore, the correct sequence is CADBE and the middle of the sequence is figure D.

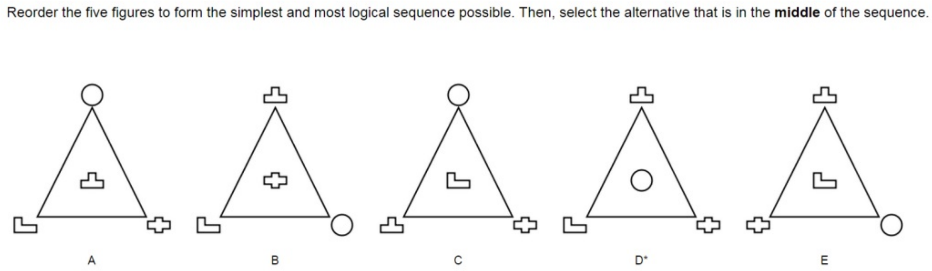

Reorder the five figures to form the simplest and most logical sequence possible. Then, select the alternative that is in the **middle** of the sequence.

Fig. 2. A "middle of the sequence" nonverbal reasoning item.

3.2 Item Modeling Procedure

An item model was created by content specialists using the parent item presented in Figures 1 and 2. For mathematics, two elements were identified, E1 and E2. These values were selected for illustrative purposes only. The generative capacity of this model could be increased substantially by increasing the range and decreasing the increment value within the range. It could also be increased by including related calculations for other geometric shapes. Taken together, the mathematics item model has the element structure of E1(3)*E2(2) which produces 6 generated items.

For nonverbal reasoning, six elements were identified and manipulated for item generation. The elements are summarized in Figure 3. Element 1 is the base image for the nonverbal reasoning item which corresponds to the central figure. Our example contains five base images (i.e., E1=5). Element 2 defines the number of positions for the subfigures located around the base image. Our example has two positions (E2=2). Element 3 specifies the number and shape of each subfigure. Our example has eight subfigures (E3=8). Element 4 specifies the type of rotation permitted by each subfigure around the base image. Our example allows for 12 rotational positions (E4=12). Element 5 highlights the shading pattern for the subfigures. We have nine shading patterns in our example (E5=9). Element 6 is the step logic required to rotate the subfigures from one base figure to the next in the sequence. Our example includes four different step logic sequences (E6=4). Taken together, our nonverbal reasoning item model has the element structure of E1(5)*E2(2)*E3(8)*E4(12)*E5(9)*E6(4) which yields 34,560 generated items.

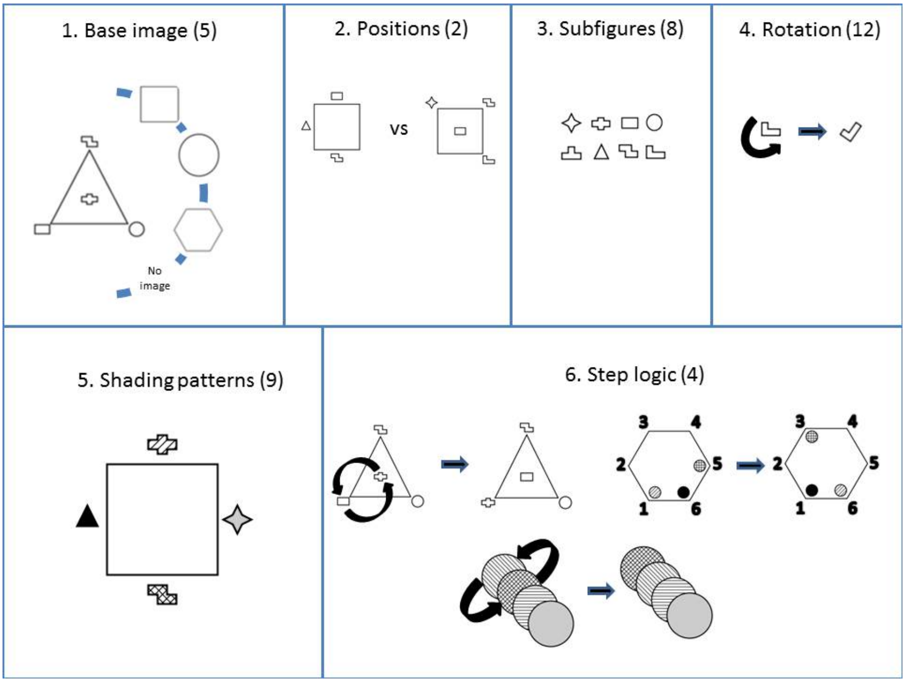


Fig. 3. A 6-element nonverbal reasoning item model.

3.3 Generation of the Solutions and Rationales

Feedback can also be included as part of the generative process. For mathematics, the solution and rationale for each set of generated correct and incorrect options can be included algorithmically in the item model. Hence, the generation process can produce both the items and the corresponding rationales unique to each item. Rationales and reasoning can be systematically generated as items are logically represented in the item model. Therefore, the adjustments made in each item model can be explained and rationalized systematically in the model. For nonverbal reasoning, the rationale for the correct option can be specified in the item model (see Table 1). Additional information could also be generated as part of the solution and rationale such as correct examples, although this type of information is not included in our study.

Table 1. Text and elements used to produce feedback for the item models in two content areas.

Item Model	Option	Feedback
Mathematics	Correct Option	The value of [] is the correct answer. It is calculated using the formula for a side length of a square given as $\sqrt{\mathrm{E1}/_{\mathrm{E2}}}$.
	Incorrect Options	Distractor 1: The value of [] is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{\mathrm{E1}/_{\mathrm{E2}}}$ but with a calculation mistake of adding 1 to the solution. Distractor 2: The value of [] is an incorrect answer. It is calculated using the formula for a side length of a square$\sqrt{\mathrm{E1}/_{\mathrm{E2}}}$but with a calculation mistake of subtracting 1 to the solution. Distractor 3: The value of [] is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{\mathrm{E1}/_{\mathrm{E2}}}$ but with a calculation mistake of adding 1.5 to the solution.
Nonverbal Reasoning	Solution	The base pattern is a [hexagon:E1] with [same colored:E5] [different subfigure shapes:E3] placed [in all the inner corners:E2]. There are [] given subfigures: [circle:E3], [star:E3] and [rectangle:E3]. The subfigures change their position [in the hexagon:E1]. The [circle:E3] moves by [] position [clock-wise:E4] each step. The [star:E3] moves by [] position [clock-wise:E4] each step. The [rectangle:E3] moves by [] positions [counter clock-wise:E4] each step. Therefore, the correct logical sequence is [DAEBC:E6] and the middle of the sequence is [].

3.4 Item Generation with IGOR

After the model is created, items and feedback are generated using IGOR. IGOR produces all possible combinations of elements based on the definitions within the model. To generate items, solutions, and rationales, a model must be expressed in an XML format that IGOR can interpret. Once a model is expressed in XML, IGOR computes the necessary information and outputs items in either HTML or Word format.

4 Examples of Generated Items

IGOR generated 6 items from the 2-element mathematics model. This result serves as a simple example to demonstrate the logic of the generation process. IGOR generated 34,560 items from the 6-element nonverbal reasoning model. This result provides a more realistic outcome highlighting the complexity and capacity that would be expected in an operational item generation situation. A sample of two mathematics (33% of the total sample) and two nonverbal reasoning (0.006% of the total sample) items is presented in Figure 4.

Ann has paid $1600 for planting her lawn. The cost of lawn is $30/m^2. Given the shape of her lawn is square, what is the side length of Ann's lawn?

A. 7.3 B. 8.3 C. 6.3 D. 8.8

OPTION A: The value of 7.3 is the correct answer. It is calculated using the formula for a side length of a square given as $\sqrt{{}^{1600}/_{30}}$.

OPTION B:: The value of 8.3 is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{{}^{1600}/_{30}}$. but with a calculation mistake of adding 1 to the solution.

OPTION C: The value of 6.3 is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{{}^{1600}/_{30}}$. but with a calculation mistake of subtracting 1 to the solution.

OPTION D: The value of 8.8 is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{{}^{1600}/_{30}}$. but with a calculation mistake of adding 1.5 to the solution.

Ann has paid $1675 for planting her lawn. The cost of lawn is $45/m^2. Given the shape of her lawn is square, what is the side length of Ann's lawn?

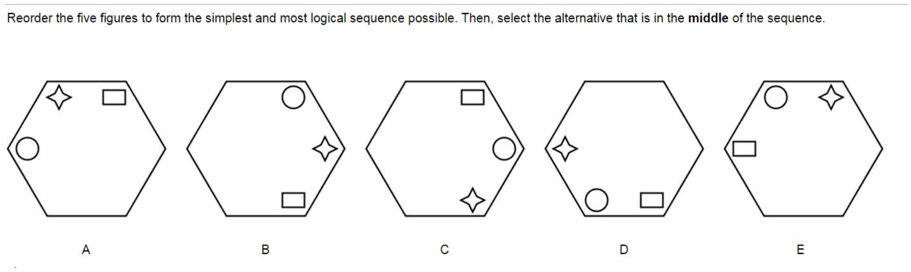
A. 6.1 B. 7.1 C. 5.1 D. 7.6

OPTION A: The value of 6.1 is the correct answer. It is calculated using the formula for a side length of a square given as $\sqrt{{}^{1675}/_{45}}$.

OPTION B:: The value of 7.1 is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{{}^{1675}/_{45}}$. but with a calculation mistake of adding 1 to the solution.

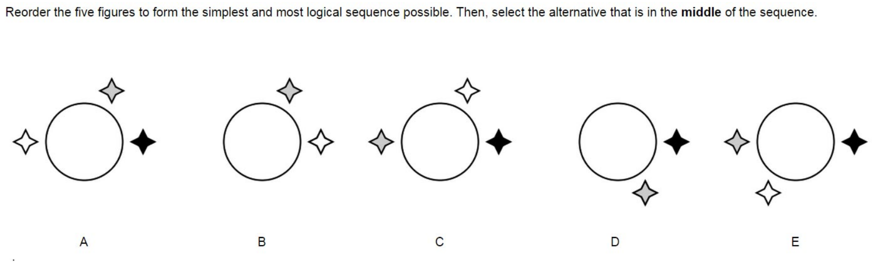
OPTION C: The value of 5.1 is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{{}^{1675}/_{45}}$. but with a calculation mistake of subtracting 1 to the solution.

OPTION D: The value of 7.6 is an incorrect answer. It is calculated using the formula for a side length of a square $\sqrt{{}^{1675}/_{45}}$. but with a calculation mistake of adding 1.5 to the solution.



Reorder the five figures to form the simplest and most logical sequence possible. Then, select the alternative that is in the **middle** of the sequence.

RATIONALE: The base pattern is a hexagon with different subfigure shapes placed in all the inner corners. There are 3 given subfigures: circle, star and rectangle. The subfigures change their position in the hexagon. The circle moves by 1 position clock-wise each step. The star moves by 1 position clock-wise each step. The rectangle moves by 2 positions counter clock-wise each step. Therefore, the correct logical sequence is DAEBC and the middle of the sequence is E.



Reorder the five figures to form the simplest and most logical sequence possible. Then, select the alternative that is in the **middle** of the sequence.

RATIONALE: The base pattern is a circle with different colored stars placed in 6 equally distributed positions on the outside of the circle. There are 3 given subfigures: white, grey and black. The subfigures change their position on the circle. The white star moves by 1 position counter clock-wise each step. The grey star moves by 2 positions counter clock-wise each step. The black star moves by 0 positions each step. Therefore, the correct logical sequence is AEDBC and the middle of the sequence is D.

Fig. 4. A sample of four generated items along with the solution and rationale from the mathematics and nonverbal reasoning item models.

5 Conclusions

Administering a test more frequently so that students receive timely, specific feedback serves as one important formative assessment principle. But when testing occurs more frequently, a constant supply of unique, content-specific test items is needed. This item supply must also be produced in a cost-effective manner. One approach that may help address these challenges is with automatic item generation (AIG). AIG is the process of using models to generate items using computer technology. It requires two steps. First, content specialists create item models. Second, the elements in the item model are manipulated with computer-based algorithms. With this two-step process, hundreds or even thousands of new items can be created from a single item model. The purpose of our study was to present an innovative new method for creating solutions and rationales as part of the item modeling and generation process. The method was demonstrated in two diverse content areas, mathematics and nonverbal reasoning.

5.1 Directions for Future Research

For the multiple-choice format, items include a stem with a correct option and a corresponding set of incorrect options. Presumably, the incorrect options are designed from a list of plausible but incorrect alternatives linked to common misconceptions or errors in thinking, reasoning, and problem solving. All three incorrect options we selected for the mathematics item model in Figure 1 were based on simple computational errors. Hence, only inferences about students' computational skills can be made from their responses to these options. But other types of errors could also be used to create the incorrect options (e.g., using side length calculations for geometric shapes other than squares). These errors could yield different types of inferences about students' thinking and reasoning skills that, in turn, could be used to produce different types of rationales and feedback as part of the item generation process. Effective formative assessment permits instructors to identify students' problem-solving strengths and weaknesses so they can adjust their instruction to overcome the weaknesses [10]. But to achieve this outcome, the assessment must contain a relatively large number of items with carefully selected incorrect options related to a single concept in order to pinpoint different types of weaknesses, problems, and/or misconceptions. Future research on the generation of plausible but incorrect options should now be conducted so a more robust feedback system can be developed for students and their teachers.

Also, the items generated must be evaluated using both substantive and statistical outcomes. Substantive analyses focus on the judgement of item quality from content experts. We used an iterative process with content specialists who specified the content used for the models in the current study in order to ensure that the generated items were judged to be acceptable. However, a more thorough review should be conducted using a independent panel of content experts who were not involved in the item development process to ensure the generated items meet the appropriate standards of quality. Statistical analyses focus on the psychometric characteristics of the generated items. The generated items should be administered to a sample of examinees so the psychometric properties can be evaluated. Item analyses from both classical test theory and item response theory can then be used to analyze the performance of the correct and incorrect options to ensure they are functioning properly. Statistical analyses can also be conducted to evaluate the difficulty and discrimination levels across the generated items to determine their comparability.

Acknowledgement. The authors would like to thank the Australian Council for Educational Research (ACER) for their support. However, the authors are solely responsible for the methods, procedures, and interpretations expressed in this study. Our views do not necessarily reflect those of ACER.

References

1. Mayrath, M. C. (Ed.). (2012). Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research. IAP.
2. Beller, M. (2013). Technologies in large-scale assessments: New directions, challenges, and opportunities. In M. von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), The role of international large-scale assessments: Perspectives from technology, economy and educational research (pp. 25–45). New York: Springer.

3. 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.
4. Gierl, M.J., & Haladyna, T. (2013). Automatic item generation: Theory and practice. New York: Routledge.
5. Irvine, S. H., & Kyllonen, P. C. (2002). Item generation for test development. Hillsdale, NJ: Erlbaum.
6. Foulonneau, M., & Ras, E. (2013, June). Assessment item generation, the way forward. Paper presented at the annual meeting of the International Computer Assisted Assessment (CAA) Conference. Zeist, The Netherlands.
7. 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.
8. 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.
9. 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.
10. Quellmalz, E. S., & Pellegrino, J. W. (2009). Technology and testing. Science, 323(5910), 75-79.