Mesure et évaluation en éducation

A Methodology for Multilingual Automatic Item Generation

Mark J. Gierl et Hollis Lai

Volume 37, numéro 3, 2015

URI : https://id.erudit.org/iderudit/1036327ar
DOI : https://doi.org/10.7202/1036327ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Gierl, M. J. & Lai, H. (2015). A Methodology for Multilingual Automatic Item Generation. *Mesure et évaluation en éducation, 37*(3), 39–61. https://doi.org/10.7202/1036327ar

Résumé de l'article

Les agences d'évaluation ont besoin d'un grand nombre d'items de première qualité produits de façon rapide et économique, et de plus en plus souvent dans différentes langues. Dans cet article, une méthodologie de génération automatique d'items (AIG) multilingues est proposée. L'AIG correspond au processus d'utilisation de modèles d'items dans le but de générer les items d'un test à l'aide de la technologie informatique. Une approche AIG en trois étapes est décrite, dans laquelle les spécialistes en développement de test doivent d'abord identifier le contenu qui sera utilisé pour générer les items. Par la suite, ces spécialistes créent des modèles d'items afin de préciser le contenu de la tâche d'évaluation qui doit être manipulée pour produire de nouveaux items. Enfin, les éléments du modèle d'items sont manipulés à l'aide d'algorithmes informatiques pour générer de nouveaux items. L'ajout des langues désirées à l'étape de création des modèles d'items permet d'effectuer une génération automatique d'items multilingues. Cette méthode est illustrée en générant 360 items en français et 360 items en anglais dans le domaine de la formation médicale. L'importance de créer des banques d'items lors du développement de tests multilingues est également discutée.

A Methodology for Multilingual Automatic Item Generation

Mark J. Gierl
Hollis Lai
University of Alberta

Keywords: automatic item generation, test development, technology and testing

Testing agencies require large numbers of high-quality items that are produced in a cost-effective and timely manner. Increasingly, these agencies also require items in different languages. In this paper we present a methodology for multilingual automatic item generation (AIG). AIG is the process of using item models to generate test items with the aid of computer technology. We describe a three-step AIG approach where, first, test development specialists identify the content that will be used for item generation. Next, the specialists create item models to specify the content in the assessment task that must be manipulated to produce new items. Finally, elements in the item model are manipulated with computer algorithms to produce new items. Language is added in the item model step to permit multilingual AIG. We illustrate our method by generating 360 English and 360 French medical education items. The importance of item banking in multilingual test development is also discussed.

Mots clés : génération automatique d'items, développement de test, technologie et évaluation

Les agences d'évaluation ont besoin d'un grand nombre d'items de première qualité produits de façon rapide et économique, et de plus en plus souvent dans différentes langues. Dans cet article, une méthodologie de génération automatique d'items (AIG) multilingues est proposée. L'AIG correspond au processus d'utilisation de modèles d'items dans le but de générer les items d'un test à l'aide de la technologie informatique. Une approche AIG en trois étapes est décrite, dans laquelle les spécialistes en développement de test doivent d'abord identifier le contenu qui sera utilisé pour générer les items. Par la suite, ces spécialistes créent des modèles d'items afin de préciser le contenu de la tâche

d'évaluation qui doit être manipulée pour produire de nouveaux items. Enfin, les éléments du modèle d'items sont manipulés à l'aide d'algorithmes informatiques pour générer de nouveaux items. L'ajout des langues désirées à l'étape de création des modèles d'items permet d'effectuer une génération automatique d'items multilingues. Cette méthode est illustrée en générant 360 items en français et 360 items en anglais dans le domaine de la formation médicale. L'importance de créer des banques d'items lors du développement de tests multilingues est également discutée.

Palavras-chave: geração automática de itens, desenvolvimento de testes, tecnologia e avaliação

As agências de avaliação precisam de um grande número de itens de primeira qualidade produzidos de forma rápida e económica, e, cada vez mais, em diferentes línguas. Neste artigo, é proposta uma metodologia para a geração automática de itens (AIG) multilingues. A AIG é o processo de utilização de modelos de itens com a finalidade de gerar itens de um teste com o apoio da tecnologia informática. Descreve-se uma abordagem AIG em três etapas, na qual os especialistas em desenvolvimento de testes devem identificar, desde logo, o conteúdo que será utilizado para gerar os itens. De seguida, estes especialistas criam os modelos de itens para especificar o conteúdo da tarefa de avaliação que deve ser manipulado para produzir novos itens. Finalmente, os elementos do modelo de itens são manipulados usando algoritmos informáticos para gerar novos itens. Adicionando as línguas desejadas à etapa de criação de modelos de itens é possível efetuar a geração automática de itens multilingues. Este método é ilustrado através da geração de 360 itens em francês e 360 itens em inglês no campo da formação médica. Discute-se também a importância da criação de bancos de itens no desenvolvimento de testes multilingues.

Authors' Notes: Invited paper appearing in special issue on Methodological Advances in Assessment, Eric Frenette and François Vachon (Guest Editors). The authors would like to thank Vasily Tangyin, Christina Rinaldi, and the Medical Council of Canada for their contributions to this research. However, the authors are solely responsible for the methods, procedures, and interpretations expressed in this study.

For correspondence: Mark J. Gierl, 6-110 Education North, Faculty of Education, University of Alberta, Edmonton, AB, Canada T6G 2G5, Phone: 780-492-2396, E-mail: [mark.gierl@ualberta.ca]; Hollis Lai, 5-533 Edmonton Clinic Health Academy, University of Alberta, Edmonton, AB, Canada, T6G 1C9, Phone: 780-492-7429, E-mail: [hollis.lai@ualberta.ca].

Introduction

Automatic item generation (AIG; Embretson & Yang, 2007; Gierl & Haladyna, 2013; Irvine & Kyllonen, 2002) is a test development method where the procedures and practices found in cognitive psychology, psychometrics, linguistics, and computer science are used to create large numbers of test items. Items can be generated for any educational or psychological test where content representation is required to evaluate the examinees' domain knowledge, including content areas in K-12 education such as mathematics, science, social studies, and language arts or in a licensure and certification testing context such as medical education. AIG can be characterized as the process of using models to generate items with the aid of computer technology. AIG serves as a technology-enhanced approach to test development that allows educators to produce large numbers of high-quality items in any content area both rapidly and efficiently. (For a historical overview of AIG, see Haladyna, 2013.)

There are many practical reasons why large numbers of items must be developed and then banked to operate a modern testing program. A flexible administration schedule is now required in most programs because examinees expect continuous, on-demand testing. Similarly, decision makers want immediate access to information about examinees. To meet this demand, large item banks must be developed to support a computer-based testing system to ensure that examinees receive a constant supply of new assessment tasks. These banks must also be frequently replenished with new items to ensure that exposure rates are minimized so test security can be maintained.

A modern testing program must also accommodate multilingual testing. Educational tests are now developed and administered to examinees in different languages throughout the world. As a result, large numbers of items are not only required to promote flexible administration schedules with adequate security procedures but these items must also be developed in multiple languages to accommodate the linguistic diversity that can now be expected when tests are administered to heterogeneous samples of examinees. Many examples can be cited. For instance, the Organization

for Economic Co-operation and Development (OECD) develops and administers the Programme for International Student Assessment (PISA). In this testing programme, 101 different "national versions" of PISA are created, validated, and administered in 45 different languages. The European Personnel Selection Office (EPSO) administers the exams used to select civil servants for various positions within the European Union. EPSO administers their selection exams in all 23 official European Union languages. Bilingual countries like Canada also require multilingual assessments because all national exams must be developed and administered in both official languages. For instance, all medical students in Canada seeking entry into supervised clinical practice in postgraduate training programmes must write the Medical Council of Canada qualifying exam. This one-day computer-based exam is offered in both English and French.

Multilingual Automatic Item Generation: A Three-step Approach

Gierl, Lai, and Turner (2012; see also Gierl & Lai, 2013) described a three-step process for AIG. In Step 1, the content required for the generated items is identified by test development specialists. In Step 2, an item model is developed by the test development specialists to specify where content is placed in each generated item. In Step 3, computer algorithms are used to place the content specified in Step 1 into the item model developed in Step 2. Gierl, Lai, Fung, and Zheng (in press) recently demonstrated how Step 2 could be expanded to permit multilingual AIG.

An item model is a template that highlights the features in the assessment task that must be manipulated to produce new items. Hence, the AIG approach we describe is *template based*, meaning that an item model is used to guide the generative process. Two types of item-model templates can be created (Gierl & Lai, 2012, 2013). A *one-layer* item model manipulates a relatively small number of elements at one level in the model. An *n-layer* item model manipulates a relatively large number of elements at two or more levels in the model. Gierl et al. (in press) demonstrated how n-layering could be used to link language elements within an item model, thereby permitting multilingual AIG. However, they did not describe how *n*-layer item modeling could serve as a generalizable method for multilingual AIG. Hence, the purpose of this paper is to describe a method that can be used for multilingual AIG. We also

demonstrate how this method can be used to generate medical education test items in English and French. But before the method is presented, we briefly describe each step in the AIG process.

Step 1: Identify Content for Generated Test Items

To begin, test development specialists identify the content required to produce new items. Gierl et al (2012) introduced the concept of a cognitive model for AIG. Figure 1 contains a cognitive model for AIG required to treat infection during pregnancy. A cognitive model is a representation that highlights and helps organize the knowledge, skills, and content required to make a medical diagnosis. The model also organizes this cognitive and content-specific information into a coherent whole, thereby presenting a succinct yet structured organization of the content relationships and sources of information used in formulating medical diagnoses.

This cognitive model was created by two medical content specialists who were experienced test developers and practicing physicians. It serves as a representation of how physicians think about and solve problems related to infection during pregnancy. The cognitive structure in Figure 1 is presented in three panels. The top panel identifies the problem and its associated scenarios. In this example, five different drug types could be prescribed to treat specific infections during pregnancy [i.e., penicillin (P), cephalosporin (C), macrolides (M), sulfa (S), furantoin (F)]. The brand names for each drug type are also identified (e.g., penicillin G, amoxicillin, and ampicillin). The middle panel specifies the relevant sources of information. Two sources of information are specified in this example: type of infection and patient characteristics. The bottom panel highlights the features. Five features (i.e., urinary tract infection, pneumonia, cellulitis, gestation period, allergy) are identified across the two sources of information. Each feature contains elements and constraints. Elements identify the content specific to each feature that can be manipulated for item generation. The pneumonia feature in the cognitive model, for example, contains the element "present" (i.e., the pregnant patient with infection has pneumonia). The second component for a feature is the constraint. Each element is constrained by the scenarios specific to this problem. Cephalosporin (C) and macrolides (M) are the drugs "very likely" to be used to treat infection during pregnancy when the type of infection is pneumonia. This constraint also implies that the other drugs are not likely to be used.

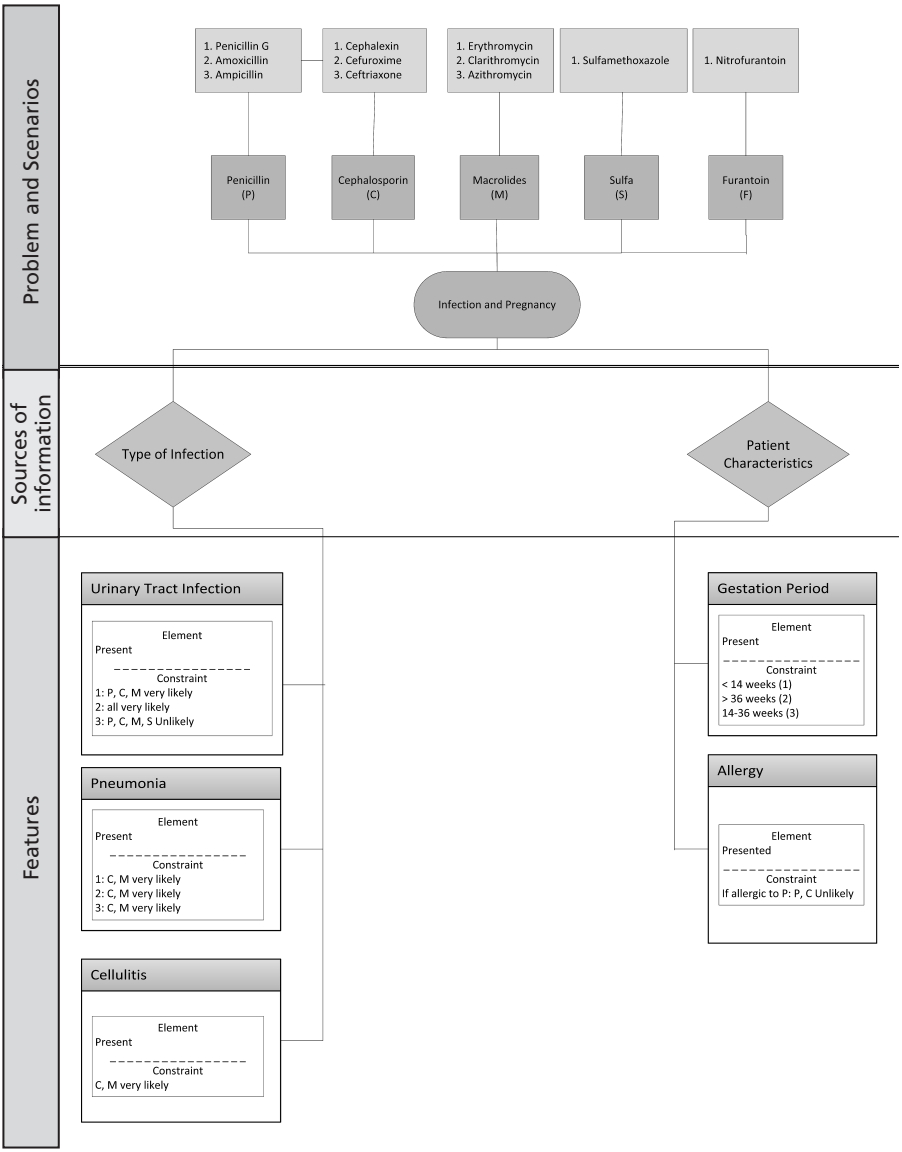

Figure 1. ***Cognitive model for AIG using the infection-during-pregnancy scenario***

Step 2: Item Model Development

Once the content is identified in the cognitive model, it must be placed into the template required for item generation. Item models provide this template-based format. Item models contain the components in an assessment task that require content. These components include the stem, the options, and the auxiliary information. The stem contains context, content, and/or the 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 additional content, in either the stem or option, required to generate an item, including text, images, tables, graphs, and diagrams. A sample parent item and the associate item model for the infection during pregnancy example is presented in Figure 2. This example contains a stem and options but no auxiliary information.

PARENT ITEM:

A 24-year-old pregnant female at 24 weeks gestation presents with clinical and radiological signs and symptoms consistent with a left lower lobe pneumonia. Which one of the following antibiotics is the most appropriate?

1. Levofloxacin.
2. Tetracycline.
3. Clarithromycin.
4. Doxycycline.
5. Azithromycin.

ITEM MODEL:

A [[AGE]] -year-old pregnant female at [[TRIMESTER]] [[ALLERGY]] presents with clinical and radiological signs and symptoms consistent with [[TYPE OF INFECTION]]. Which one of the following antibiotics is the most appropriate?

Figure 2. ***Parent item (top) and the associated item model (bottom) used to evaluate an examinees knowledge about treatments for infection during pregnancy***

Two types of item models can be created for AIG. The first type is a one-layer item model. The goal of item generation using the one-layer approach is to produce new assessment tasks by manipulating elements at one level. An example of a one-layer item model for infection during pregnancy is presented in Figure 3. The stem contains one integer [AGE] and three strings [TRIMESTER; ALLERGY; TYPE OF INFECTION]. The second type is an *n*-layer item model. The goal of AIG using the *n*-layer item model is to produce items by manipulating elements at two or more levels. The *n*-layer structure can therefore be described as a model with multiple layers of elements, where each element can be varied simultaneously at different levels to produce different items. An example of an *n*-layer item model for infection during pregnancy is presented in Figure 4. In this example, two layers are used. The first layer is sentence structure. The first sentence is "A [AGE]-year-old pregnant female at [TRIMESTER]; [ALLERGY]; presents with clinical and radiological signs and symptoms consistent with [TYPE OF INFECTION]." The second sentence is "Suppose a pregnant woman [ALLERGY] was admitted with signs consistent with [TYPE OF INFECTION]. She was in her [TRIMESTER] ." The second layer includes the same elements specified in the one-layer model, except that trimester is expressed as both an integer and string and the options are presented as both drug types and brand names to increase the generative capacity of the model. The item model also contains the options which specify the correct alternative and one or more incorrect alternatives or distracters. For the infection-during-pregnancy example, the possible treatment options are Furantoin, Sulfa, Macrolide, Cephalosporin, and Penicillin. The correct option serves as the key and the remaining options serve as the distractors for every generated item.

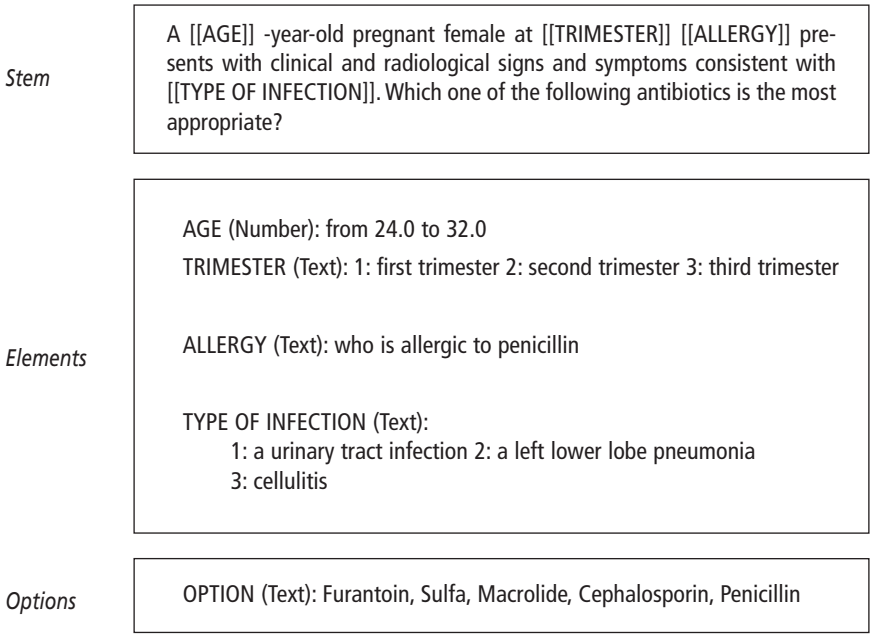

Figure 3. ***1-layer item model for the infection-during-pregnancy example***

The *n*-layer model is a flexible structure for item generation because it allows for many different combinations of embedded elements. Gierl et al. (in press) illustrated how language could be used as a layer within an item model to permit multilingual AIG. They introduced the concept of a *linked element* as a way to facilitate the computer programming task in Step 3 of the AIG process. Layered elements permit content to be embedded within content in an item model, thereby serving a "vertical" function for item content (i.e., content within content). Layered elements are presented in the *n*-layer example in Figure 4. Linked elements expand the capabilities of item modeling by permitting content to be transformed within an item model. For multilingual AIG, the transformation is from one language to another. Linked elements, therefore, have a "horizontal" function for item content (i.e., content in source language is transformed to content in target language). Gierl et al. (in press) used linked elements in English, French, and Chinese to create an *n*-layer model for generating medical test items requiring examinees to treat complications with hernias.

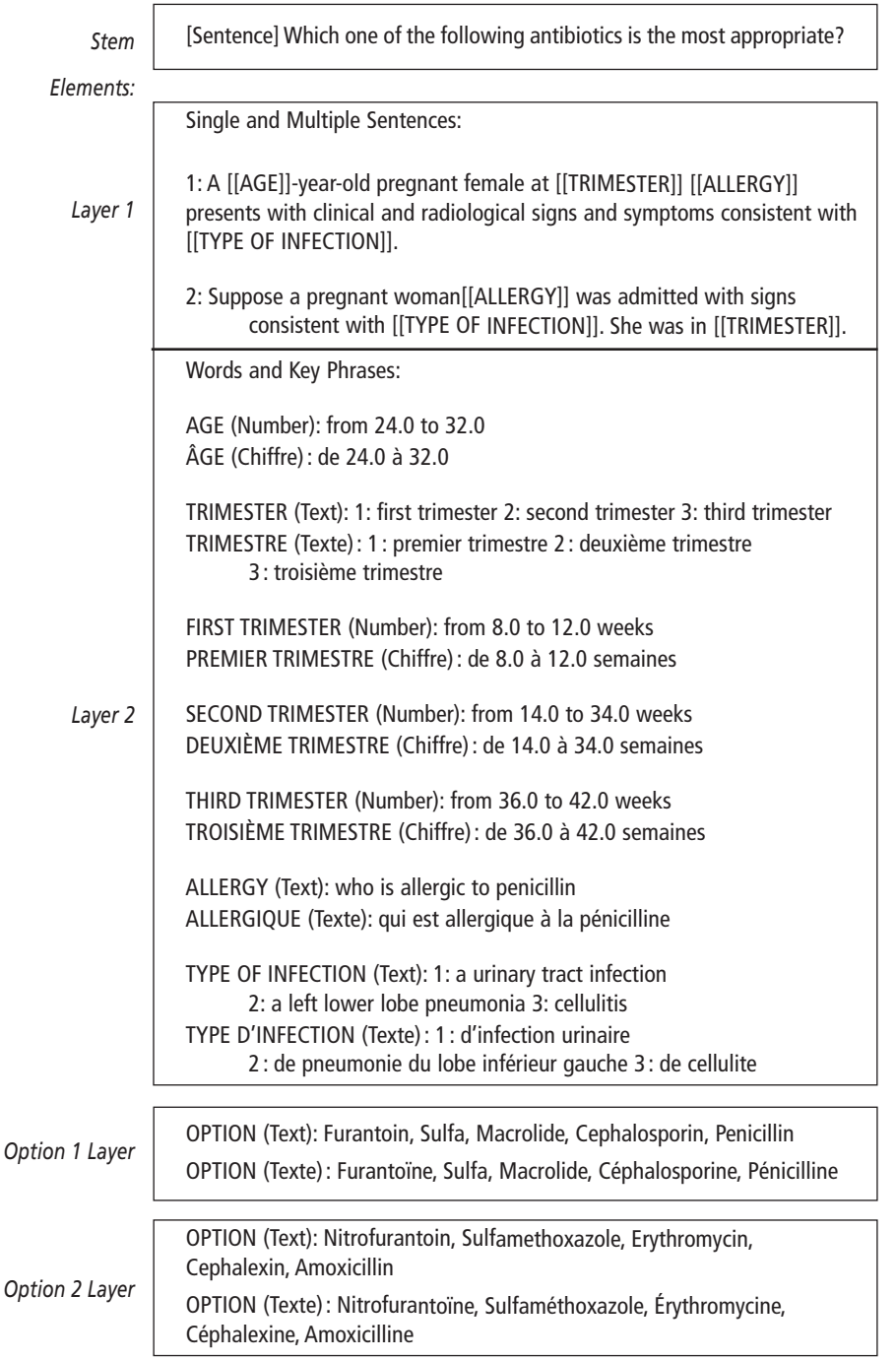

Figure 4. ***N-layer item model for the infection-during-pregnancy example***

Step 3: Item Generation Using Computer Technology

Once the content has been specified and the item models created, this information must be assembled to generate new items. This assembly task is conducted with a computer algorithm because it is often a complex combinatorial problem. Gierl, Zhou, and Alves (2008) created a computer program called IGOR to solve this test assembly problem. IGOR, which stands for Item GeneratOR, is a Java-based program designed to assemble the content specified in the item model, subject to the elements and constraints identified in the cognitive model. While we use IGOR to demonstrate Step 3 in the AIG process, it should also be noted that any linear programming method could be used to solve this type of combinatoric problem (van der Linden, 2005).

A Methodology for Multilingual Automatic Item Generation

While Gierl et al. (in press) illustrated how linked elements could be used with *n*-layer modeling to generate items in multiple languages, they did not present either an explicit or a generalizable method for their approach. Rather, they only demonstrated how *n*-layer modeling could permit multilingual AIG. Hence, the purpose of this paper is to describe a method that can be used by researchers and practitioners for multilingual AIG. We also demonstrate how this method can be used with the infection-during-pregnancy item model in Figure 4 to generate items in English and French. To accomplish this goal, we begin with a description of key concepts and procedures drawn from the literature on classical machine translation (Isabelle & Foster, 2006; Jurafsky & Martin, 2009; Koehn, 2010). Machine translation refers to a body of methods by which computers can perform the operations required to translate a human *source language* to a human *target language*. These concepts establish the foundation necessary for our newly proposed method.

The Vauquois Triangle and Classical Machine Translation

Classical machine translation is often explained using the Vauquois triangle (Vauquois, 1968). The triangle, shown in Figure 5, highlights different levels of abstraction required when translating written text from a source to a target language. The bottom level is described as the direct approach to machine translation where words are directly translated from one language to another using automated, computer-based methods. Direct approaches are considered to be the least abstract (or the most concrete). In the most literal application of this approach, each word in the source language is translated to a corresponding word in the target language, one word at a time. The direct approach assumes that one language can be transformed to another language at the word level.

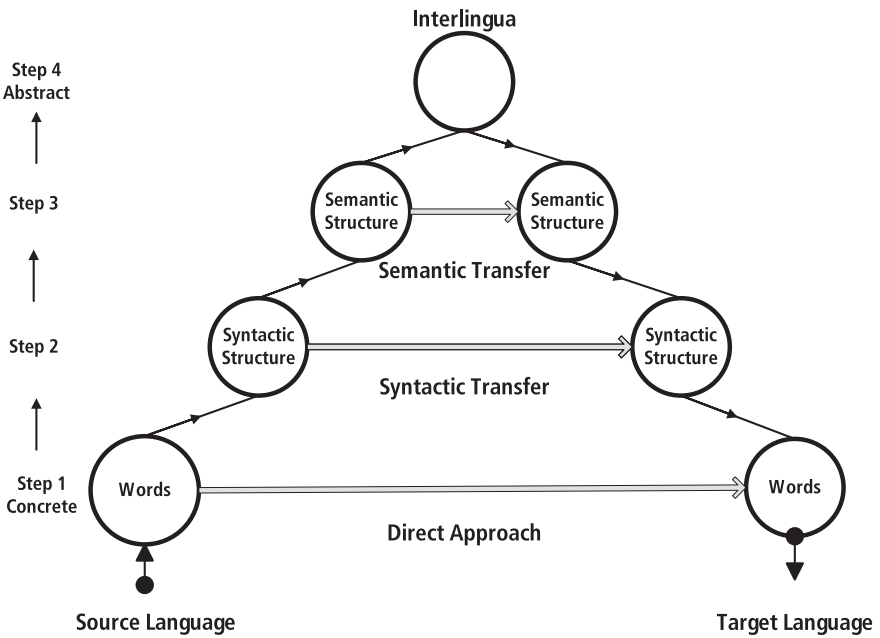

Figure 5. ***The Vauquois triangle (adapted from Jurafsky & Martin, 2009)***

The next two levels can be characterized as transfer approaches to machine translation where the structural differences between languages are identified and adjusted using automated, computer-based methods. Transfer approaches are moderately abstract. At these two levels, rules designed to address syntactic (e.g., grammatical) differences between the source and target languages are implemented so that semantic (i.e., meaning) concordance between the languages can be achieved. In other words, transfer rules are used to adjust the structure of the source language so it conforms to the structure of the target language. Then, additional transfer rules are implemented so the meaning between the languages is aligned across the newly structured texts. This approach is guided by the logic that transfer rules can be used to align structural differences across languages, thereby producing meaningful text translations.

The top level is called the interlingua approach. It is the most abstract level. With this approach, the meaning of the source language text is expressed as a representation called the interlingua. Then, the target language translation is produced using the interlingua representation. This approach is guided by the logic that translation is mediated by meaning. Hence, successful translation entails the extraction of meaning from the source language which is then expressed in the target language.

Senellart, Dienes, and Váradi (2001) described how levels in Vauquois's triangle could be combined to guide rule-based machine translation systems. They proposed a combined approach that incorporates the direct and transfer levels. We adopt this approach for our multilingual AIG method. The approach has three components: Analysis, transfer, and synthesis. We describe each component and illustrate how the component can be used to translate the infection-during-pregnancy item model from English to French, thereby permitting the generation of multilingual test items. A French-speaking content specialist who was fluent in both English and French served as our medical test developer and translator. The source language is English and the target language is French for this demonstration.

Analysis Component

To begin, the linked elements are translated directly from the source to the target language in two forms. The first form is at the *word level*, meaning that each word element in the item model is translated from English to French. The second form is the *key phrase level*, meaning that

each key phrase elements in the item model is translated from English to French. The analysis component is therefore akin to the direct approach, which is the first step in Vauquois's triangle. The bilingual content specialists directly translated the words and key phrases from the source to the target language for all elements in the *n*-layer infection-during-pregnancy item model. A summary of the translations is presented in Figure 4.

Transfer Component

Next, the linked elements are translated from the source to the target language at the *sentence level*. For this component, the adequacy of the word and key phrase translation is considered by the bilingual content specialist, as in the analysis component, but now the syntactic structure of the words and key phrases as they form sentences is also considered. The process of reordering occurs in this component to capture important structural differences that differentiate the source and target languages. The reordering process is often called alignment (Jurafsky & Martin, 2009). The transfer component corresponds to the syntactic level in the transfer approach, which is the second step in Vauquois's triangle. Take, for example, the sentence "Which one of the following antibiotics is the most appropriate?". The French translation is not a literal word-by-word transformation. Instead, the syntactic structure from English must be adapted to French, as shown in Figure 6, so that the two phrases are aligned. The resulting French item model is given as "Lequel des antibiotiques suivants serait le plus indiqué?". In short, rules are used in the transfer component to adjust the structure of the source language so it conforms to the structure of the target language.

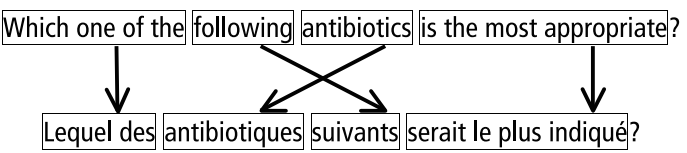

Figure 6. ***Reordering required between English and French needed to address the syntactic differences between the languages***

Synthesis Component

Finally, the meaningfulness of the translated items, given the outcomes from the analysis and transfer components, is addressed. The linked elements are translated and aligned from the source to the target language at the *multiple sentence level* to produce models that can be used to generate translated test items. For this component, the word and key phrase translation is synthesized with the newly aligned sentences so items can be generated from the model. Once the items are generated, a judgment is made about the adequacy of the semantic structure for the translated items. In our multilingual AIG method, this judgment is made by a human translator who evaluates a sample of the generated items to determine if they are meaningful. The synthesis component corresponds to the semantic level in the transfer approach, which is the third step in Vauquois's triangle. For example, our item model initially generates the following stem: "Supposons qu'une femme enceinte a été hospitalisée parce qu'elle présentait des signes de cellulite. Elle était au 1er trimestre de sa grossesse. Lequel des antibiotiques suivants serait le plus indiqué?". But upon review by the bilingual medical content specialist, the item was edited and revised, as follows, to produce a more meaningful stem: "Supposons qu'une femme enceinte est hospitalisée parce qu'elle présente des signes de cellulite. Elle est dans le 1er trimestre de sa grossesse. Lequel des antibiotiques suivants serait le plus indiqué?".

If the analysis and transfer components are implemented successfully, then the elements across the source and target item models are correctly synthesized, thereby producing meaningful test items. Conversely, if either the analysis or the transfer component is implemented unsuccessfully, then the elements across the source and target item models are incorrectly synthesized, producing problematic test items. The root of the problem can be addressed by scrutinizing the analysis and transfer results. Once the potential problems are identified and corrected, the results can again be synthesized and the meaningfulness of the generated items can be determined. Iterating through the analysis and transfer components should occur until the item models can consistently generate meaningful (i.e., semantically appropriate) test items. A summary of our multilingual AIG approach is presented in Figure 7.

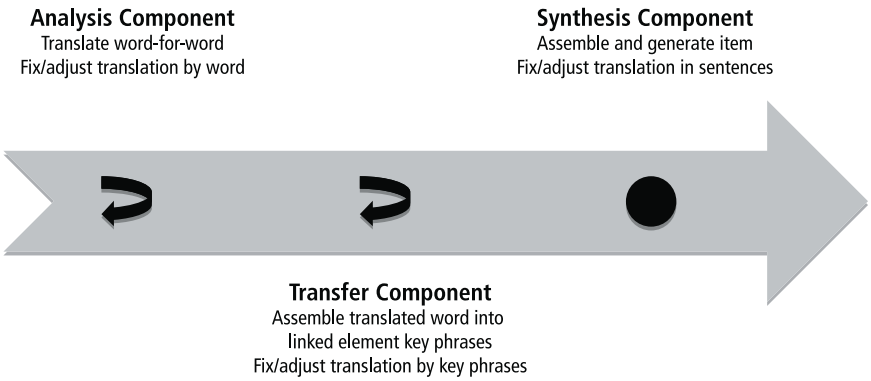

Figure 7. ***Summary of the multilingual AIG process.***

We described our multilingual AIG method and then illustrated the method using two languages, English and French. But this method can also be used with three or more languages using the same logic to permit simultaneous multilingual item generation across all of the languages specified in the item modeling step. Once the analysis and transfer components are completed, a multilingual AIG *linking map* is produced. The map contains the necessary text and translation required to link the words, key phrases, single sentences, and multiple sentences across the language groups. Then, we implement computer assembly algorithms found in programs like IGOR as part of Step 3 in the AIG process so that multilingual items can be generated. Using our multilingual AIG methodology, a total of 720 infection-during-pregnancy items were generated—360 English and 360 French items using the model in Figure 4. A random-sample of four items in English and French is presented in Tables 1 and 2, respectively.

Table 1
Random-sample of four English items generated using the infection-during-pregnancy cognitive model

40. A 32-year-old pregnant female in her second trimester presents with clinical and radiological signs and symptoms consistent with a urinary tract infection. Which one of the following antibiotics is the most appropriate?
    1. Cephalosporin.
    2. Furantoin.
    3. Isotretinoin.
    4. Macrolide.
    5. Penicillin.
48. A 32-year-old pregnant female in her third trimester presents with clinical and radiological signs and symptoms consistent with cellulitis. Which one of the following antibiotics is the most appropriate?
    1. Penicillin.
    2. Macrolide.
    3. Sulfa.
    4. Isotretinoin.
    5. Furantoin.
69. Suppose a pregnant woman was admitted with signs consistent with cellulitis. She was in her first trimeter. Which one of the following antibiotics is the most appropriate?
    1. Isotretinoin.
    2. Cephalosporin.
    3. Sulfa.
    4. Penicillin.
    5. Furantoin.
74. Suppose a pregnant woman was admitted with signs consistent with cellulitis. She was in her second trimester. Which one of the following antibiotics is the most appropriate?
    1. Furantoin.
    2. Isotretinoin.
    3. Sulfa.
    4. Penicillin.
    5. Cephalosporin.

NB These items have been changed as per author's instructions to use the TRIMESTER parameter rather than weeks of gestation

Table 2
Random-sample of four French items generated using the infection-during-pregnancy cognitive model

11. Une femme de 24 ans dans son 2e trimestre de grossesse présente des signes et symptômes cliniques et radiologiques d'infection urinaire. Lequel des antibiotiques suivants serait le plus indiqué ?
    1. Pénicilline.
    2. Macrolide.
    3. Furantoïne.
    4. Murantoïne.
    5. Céphalosporine.
24. Supposons qu'une femme enceinte est hospitalisée parce qu'elle présente des signes de cellulite. Elle est dans le 1er trimestre de sa grossesse. Lequel des antibiotiques suivants serait le plus indiqué ?
    1. Sulfa.
    2. Macrolide.
    3. Pénicilline.
    4. Furantoïne.
    5. Céphalosporine.
34. Une femme de 24 ans dans son 3e trimestre de grossesse présente des signes et symptômes cliniques et radiologiques d'infection urinaire. Lequel des antibiotiques suivants serait le plus indiqué ?
    1. Sulfa.
    2. Furantoïne.
    3. Macrolide.
    4. Pénicilline.
    5. Céphalosporine.
40. Supposons qu'une femme enceinte est hospitalisée parce qu'elle présente des signes de pneumonie du lobe inférieur gauche. Elle est dans le 2e trimestre de sa grossesse. Lequel des antibiotiques suivants serait le plus indiqué ?
    1. Sulfa.
    2. Furantoïne.
    3. Pénicilline.
    4. Macrolide.
    5. Céphalosporine.

NB These items have been changed as per author's instructions to use the TRIMESTER parameter rather than weeks of gestation

Conclusions and Directions for Future Research

Testing agencies now require large numbers of high-quality items in multiple languages that are produced in a cost-effective and timely manner. One way to address this challenge is by increasing the number of content specialists and test translators who are assigned the task of developing multilingual items. But this approach is expensive, tedious, and slow. An alternative approach that may help overcome some of the drawbacks of a more traditional approach to item development and translation is to combine content specialization and translation expertise with computer technology to systemically produce multilingual test items using AIG. AIG is the process of using item models to generate test items with the aid of computer technology. We described a three-step AIG approach where test development specialists first identify the content that will be used for item generation, then create item models to identify the content in the assessment task that must be manipulated using a template-based approach and, finally, manipulate the elements in item models using computer algorithms. Language can be added as a layer in the *n*-layer item model to permit multilingual AIG. Hence, multilingual AIG can be considered a specific application of *n*-layer item modeling. In this study we introduced a new method for multilingual AIG by adapting concepts from classical machine translation that can be used to facilitate the translation process during the item modeling step. An example of multilingual AIG was also presented using therapeutic treatments for infection during pregnancy where we generated 720 items across two language groups.

Multilingual AIG and Item Model Banking

In their chapter on "Technology and Testing" in the 4th Edition of the handbook *Educational Measurement*, Drasgow, Luecht, and Bennett (2006, p. 471) proclaim:

> This chapter describes our vision a 21st-century testing program that capitalizes on modern technology and takes advantage of recent innovations in testing. Using an analogy from engineering, we envision a modern testing program as an integrated system of systems. Thus, there is an item generation system, an item pretesting system, and examinee registration system, and so forth. This chapter discusses each system and illustrates how technology can enhance and facilitate the core processes of each system.

This bold view of educational measurement asserts that integrated technology-enhanced systems will direct all testing processes in the future. This perspective captures a growing consensus among researchers and practitioners that a well-defined science of educational measurement will emerge to guide our development, administration, scoring, and reporting practices. It also provides a more complete picture of how the component parts function together to govern this new testing system. We presented a concrete example of how an "item generation system" could operate by describing a three-step AIG process and then illustrating how this process can be used to generate hundreds of medical test items. We develop this analogy further for multilingual test development by describing how an "item generation system" could also contain an explicit "item translation subsystem" using the components and procedures presented in our multilingual AIG method. This system would not only permit the development and generation of items in multiple languages but also allow developers to create large multilingual item banks. The unit of analysis, however, would no longer be the test item. Rather, the unit of analysis for a multilingual AIG item bank would become the item model. Hence, one more component of this new technology-enhanced testing system would be an item model bank.

Current practices in test development are focused on the item as the unit of analysis, meaning that each item is individually written, reviewed, revised, edited, and banked. If a developer wants to have 360 items in a test bank, then 360 items must be written, reviewed, revised, edited, and banked. An item bank serves as an electronic repository for maintaining and managing information on these 360 items. Because the item is the unit of analysis, maintenance of the bank also focuses on item-level information. For instance, the format of the item must be coded. Item formats and item types can include multiple choice, numeric response, written response, linked items, and passage-based items. The content for each item must be coded. Content fields include learning outcomes, blueprint categories, item identification number, item format, date, source of item, and copyright information. The developer attributes must be coded. These attributes include year the item was written, item writer name, item writer demographics, editor information, development status, and review status. The statistical characteristics for the item must also be coded. Statistical characteristics can include word count, readability, classical item statistics, item response theory parameters, distracter func-

tioning analyses, item history, differential item functioning flags, and history of item use. In short, large amounts of detailed information must be collected to maintain and manage a bank because the item is the unit of analysis. The requirements and work needed to address the bank specifications is merely expanded when items are developed in multiple languages.

To address the problem of collecting and maintaining large amounts of item-level information, we propose that item model banks be created. AIG is a scalable process because it treats the model as the unit of analysis where one model yields many items. Hence, the unit of analysis moves from the item to the model, meaning that each model is individually written, reviewed, revised, edited, and banked (but not necessarily the items generated from the model). If, for instance, a developer wants 720 items, then only one bilingual item model is required, as we illustrated in our study. An item model bank serves as an electronic repository for maintaining and managing information on each item model. The item model bank will contain information on every model, but not necessarily on every item. As a result, the amount of information collected to maintain and manage the model bank may still be considerable but, compared to the item bank, it would be substantially reduced. AIG, therefore, leads to more cost-effective test development because the model is continually re-used to yield many test items, compared with developing each item for a test from scratch. Costly errors commonly found in test item translation (e.g., omissions or additions of words, phrases, or expressions as well as spelling, punctuation, capitalization, item structure, typeface, and formatting problems) are minimized because a generalized method for multilingual AIG exists. The monitoring and maintenance of this information is also streamlined by focusing on the attributes of the model. Thus, an item model bank would be another important requirement for an "item generation system" of the future.

REFERENCES

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., & 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. doi: 10.1016/S0169-7161(06)26023-1

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

Gierl, M. J., & Lai, H. (2013). Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32,* 36-50. doi: 10.1111/emip.12018/abstract

Gierl, M. J., Lai, H., Fung, K., & Zheng, B. (in press). Using technology-enhanced processes to generate items in multiple languages. In F. Drasgow (Ed.), *Technology in testing: Measurement issues.* New York, NY: Routledge.

Gierl, M. J., & Lai, H. (2012, March). *Using automatic item generation to create items for medical licensure exams.* In K. Becker (Chair), Beyond essay scoring: Test development through natural language processing. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC. doi: 10.1111/j.1365-2923.2012.04289.x/abstract

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

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 from http://www.jtla.org.

Haladyna, T. (2013). Automatic item generation: A historical perspective. In M. J. Gierl & T. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 13-25). New York, NY: Routledge.

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

Isabelle, P., & Foster, G. (2006). Machine learning: Overview. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (2nd ed., pp 404-422). Oxford, UK: Elsevier.

Jurafsky, D., & Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (2nd ed.). Upper Saddle River, NJ: Pearson

Koehn, P. (2010). *Statistical machine translation.* Cambridge, UK: Cambridge University Press. doi: 10.1017/CBO9780511815829

Senellart, J., Dienes, P., & Várdi, T. (2001). New generation SYSTRAN translation system. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=9A507FD444755EA3B27B2075AD123D68?doi=10.1.1.68.568&rep=rep1&type=pdf

van der Linden, W. J. (2005). *Linear models for optimal test design.* New York, NY: Springer.

Vauquois, B. (1968). A survey of formal grammars and algorithms for recognition and transformation in machine translation. In *IFIP Congress-68* (pp. 254-260). Edinburgh, UK. Reprinted in C. Boitet (Ed.), *Berbard Vauquois et la TAO: Vingt-cinq ans de traduction automatique – Analectes* (pp. 201-213). Grenoble, France: Association Champollion.