Evaluating the Quality of Multilingual Items Generated Using Automatic Processes: Preliminary Results from a Reliability Study

Karen Fung

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

Dr. Hollis Lai
Undergraduate Medical Education
University of Alberta

Paper Presented at the Annual Meeting of the National Council on Measurement in Education

San Francisco, CA

April 28, 2013
Introduction

The traditional approach to test item translation is an effortful, time-consuming process conducted by bilingual or multilingual content specialists. One important problem that arises when content specialists perform translations is the introduction of subjectivity into the process. For instance, Hambleton (1993) reported that when a content specialist is told that another specialist will be back translating his or her work, the content specialist selects words that are more likely to be translated into the original item. Of course, translations can also be evaluated using other methods such as administering the original and the translated test items to a group of specialists who are fluent in both languages. However other issues may arise such as the possibility that the translators are stronger in one language compared to another (Hambleton, Merenda, & Spielberger, 2005).

As the implementation of technology in test development grows, the emergence of automatic item generation (AIG; Gierl & Haladyna, 2013) with the use of item models (Gierl, Alves, & Zhou, 2008) is becoming more prominent. The translation of item models may make it possible for large numbers of items in two or more languages to be generated simultaneously. To-date, the quality of such translation has yet to be evaluated. Hence, the purpose of our study is to begin this evaluative process by determining the similarity of the translated items when generative models are used to create items in multiple languages. We will also attempt to take advantage of the available technology for translation by using the online engine Google Translate.
Automatic Item Generation

With the growing demand for computer-based testing and computer adaptive testing, a large number of items are now required to permit flexible test administration while maintaining adequate item security. When more items are needed for this item banking model, the traditional approach of item development becomes effortful and time consuming and it requires extensive financial resources. Traditionally, items are hand crafted by content specialists one-by-one to ensure certain content areas with appropriate skill levels are covered. But this process becomes expensive as more items are needed. The use of Automatic Item Generation (AIG) saves time and money by combining content expertise with computer technology to automatically create new items.

Process of AIG

The development of multiple-choice items in the medical context using AIG was recently described by Gierl, Lai, and Turner (2012) using a three-step process. In the first step, content specialists create a graphical representation known as a cognitive model for AIG in order to identify the thinking processes required to solve medical problems. In this stage, important key information that would help with solving the problem is identified and this information is used in the next stage AIG for item model building. In the second step, content specialists create item models. An item model represents the features and the structures for the items to be generated (Bejar, 2002; Drasgow, Luecht, & Bennett, 2006; Lai, Alves, Zhou, & Gierl, 2009). It is created based on the parent item (the original item created by content expert), and the modification of the parent item is based on several components: the stem, options, and auxiliary information. The stem is the question statement that examinees will be answering; the options (in a multiple-choice scenario) would be the choices given for examinees to choose their response for that item;
the *auxiliary information* are illustrations of figures, tables, and graphs that would appear on the stem, and/or in the options (Gierl et al., 2008; Lai et al., 2009). Alterations of these three components using *elements* yields new test items. Elements are the variables that can be manipulated to create different instances of the stems and the options. The use of *constraints* control the elements found in the final items so the newly created items are permissible for assessment. In the third step, all of the components (i.e., elements, stems, auxiliary information, options and constraints) are put together and iterations are conducted to create different combinations of elements. These iterations are conducted using an AIG computer program such as the Item GeneratOR (IGOR; Gierl et al., 2008).

**Multilingualism in AIG**

In our previous research we were able to demonstrate the use of AIG to create items in different contexts such as Mathematics, Science, and Medicine. Most recently we were also able to use AIG to create items in multiple languages such as Spanish and Chinese (Gierl, Fung, Lai, & Zhang, 2013). To generate items in multiple languages, a type of item modeling known as the *n-layer item model* is required. There are two types of models: 1-layer and *n-layer item model*. One-layer item models is commonly used when only one layer of elements is being controlled using constraint; it is used when only a small number of elements are being modified in the generative process. By way of contrast, *n-layer modeling* extends from the 1-layer model into multiple layers (two or more) (Gierl, Lai, & Breithaupt, 2012). It allows for greater expansion of an item model, providing greater generative power, as an element may include another element embedded within as a value. In the case of multilingual AIG, translation will be done at the item model level, and the modifying elements of a second or third language will be embedded into the language variable. When embedded elements are used in an item model, as more iteration are
required, the generation time may increase. To increase the efficiency of the generation process, a method known as *linked elements* can be used.

Linked elements use a linear approach to group related variables together (in this case, language), in which values in each element will be matched in a parallel manner (Gierl, Fung, Lai, & Zhang, 2013). Using the logic of linked elements saves processing time and constraint programming time when elements will be automatically linked by the computer program. With the use of n-layer modeling and linked elements we were able to successfully generate items in multiple languages simultaneously. However, given the capability of generating items automatically in multiple languages, the quality of these items generated using a translated item model needs to be determined. As a first of many steps required in the quality-control check, this study evaluates the consistency of item model translations by comparing human translation versus an online translator, and content expert versus non-content expert translation on the similarity of the generated items. To do so, the *cosine similarity index* (CSI) will be used.

**Cosine Similarity Index – A Measure of Item Similarity**

The cosine similarity index (CSI) is a measure that can be used to determine intra-model differences in terms of unique word occurrence. It measures similarity between two lines of texts represented in vectors, by taking into account the occurrence of unique words in each pair of comparison (Gierl et al., 2012). The cosine would then be calculated to measure the difference between the two vectors of unique words in a multidimensional space. An average cosine obtained from all the pairs provides a single index of reference. The CSI can be represented as

\[ \cos(\theta) = \frac{A \cdot B}{||A|| ||B||}, \]
where A and B are items of word occurrences being expressed in a binary vectors (Gierl et al., 2012). CSI ranges from 0 (no overlapping unique words between two vectors) to 1 (all unique words between the vectors are the same). Currently, there is no rule on what numeric point a CSI should be considered as highly similar or very dissimilar. In this study, the CSI will be used our first approach to measure the consistency of item model translation by both human (i.e., expert and non-expert) and machine translator.

**Purpose of Study**

The purpose of this study is to demonstrate the use of AIG and online translators to determine their efficiency and effectiveness when item models have been translated. This study is designed to answer three questions: (1) When items of multiple languages are generated by n-layer modeling, do the generated items match the original items? (2) How consistent is an online translator such as Google Translate in performing translation when compared to human translator? (3) Are items generated from content expert translation more similar than items generated from non-content expert translation? If so, is there a significant difference between the two matches? Results of this study will provide the initial evidence about the degree of similarity between traditional approaches of translating individual items to the new machine-based methods for item models translation.

**Methodology**

**Item Model Translation.** To being, an item model developed from a medical licensure exam item was selected for translation. The item model was in the content area of surgery using the specific topic of hernia repair. Two approaches were used to translate the item model from English to Traditional Chinese: non-content expert and content expert. A machine translator was also used to perform translation on the generated items for making comparisons between the
content expert and non-content expert, and between the human and the machine translator to determine the consistency of these translations. The non-content expert is the primary researcher who is proficient in both English and Chinese, but who has no background in medicine studies or surgery. The non-content specialist was used as the first condition in this study to determine if the content in a medical item model could be translated without content-knowledge expertise. The content expert is an experienced surgeon who is proficiency in both English and Chinese. The content specialist was used as the second condition in this study to determine if the content in a medical item model requires additional content-based knowledge and expertise in order to produce a reliable translation. The machine translator chosen was Google Translate developed by Google Inc. It is a popular translator that provides translation online for free. It is capable of translating phrases back and forth in 65 languages. Google Translate is a very powerful translator where translation is performed with the scanning of millions of human translated documents from its database to best match the context of the phrases.

**Non-expert translation.** The study began with the forward translation of the hernia item model using a three-steps procedure. First, varying elements and major terminology were translated at the word/terminology level. For example, in the item model, gender was a varying element. Therefore, “male” and “female” are two of the words to be translated first as “男” and “女”, respectively. Second, sentences in the stem were translated to match the translated elements. Third, sentences were combined to ensure the stem as a whole made sense and flowed well together.

**Content expert translation.** To facilitate the process of translation with the content expert, the item model translated by the non-expert, along with the original English item model was shown to the content expert for preliminary comments. A comparable three-step process was
conducted in order to perform the translation in a systematic manner. First, varying elements including medical terminology where verified by the content expert and adjustment were made. Second, verification and modification of translations were done at the sentence level with the varying elements added. The third step involves combining all sentences together and verifying the stem as a whole. The task of the content expert was to ensure that (a) medical terminology were properly translated; (b) the translation is suitable for the medical exam context; and (c) sentences and phrase were grammatically or structurally correct. The goal was to ensure the item model would produce Chinese items that are functional in the medical examination context.

Procedure. Upon completion of the item model translations, the Chinese model was then combined with the English model so items in both languages would be generated simultaneously. The programming of item model was performed using linked elements, in which stem, elements, and constraints were grouped into a category based on language. The AIG program IGOR was used to generate the English and Chinese items.

Machine translation. In order to determine the consistency of translation by each method, Google Translate was used to perform forward translation and back translation on the items generated. Back translations were done for items that have been generated from the two item models translated by the content expert and the non-content expert. This process of back translation is necessary because in order to compare similarity of translated items to their original English items using CSI, the Chinese items had to be translated back into English. Using a machine translator would provide a convenient and efficiency approach when hundreds or thousands of items are generated with AIG. In addition, it may remove the subjectivity factor commonly found in back translation with content specialists (Hambleton, 1993). For the purpose of comparing the efficiency and consistency of machine translator versus human translator,
forward translation was also performed by Google Translate, from translating the generated English items one at a time to Chinese.

Since the CSI measures the similarity between two lines of texts represented in vectors by taking into account the occurrence of unique words in each pair of comparison, in this case the comparisons were made between the original item and the back translated item. Each two-phrase comparison will produce a CSI and an average will provide an overall similarity index for each translation method. A test of statistical significance was then conducted to determine if the average CSI in each method of translation is statistically different from another. Table 1 below outlined the list of comparisons performed using t-test (Factor 1 versus Factor 2). Illustrations of the study design are also included in the Appendix.

Table 1

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Factor 1</th>
<th>Factor 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Non-expert forward translate of item model; Google back translate</td>
<td>Google forward and back translate individual items</td>
</tr>
<tr>
<td>2</td>
<td>Content expert forward translate of item model, Google back translate</td>
<td>Google forward and back translate individual items</td>
</tr>
<tr>
<td>3</td>
<td>Non-expert forward translate of item model; Google back translate</td>
<td>Content expert forward translate of item model; Google back translate</td>
</tr>
</tbody>
</table>

Results

A total of 360 items (i.e., 180 items in each language) were generated by each translation approach - content expert and non-content expert. Google Translate was then used to back translate each individual Chinese item to English (see Appendix figure a and b). Google
Translate was also used to perform both forward and back translation using the generated English items from the original Hernia model (see Appendix figure c). A sample of translation results using different approaches is shown in Table 2. The back translation from Google Translate appeared to be similar in terms of wordings but different in terms of structures between the translation using content expert translated items and non-content expert translated items. To further explore into the consistency for each approach, 100 pairs of English and Chinese items were randomly chosen for comparisons using CSI.

Non-Expert Translation versus Google Translate Translation

In the case of translation of item model done by non-expert, the results provided two mean CSIs: 0.47 for item model translations by non-expert, which indicate a moderate similarity between the items generated from the non-expert translation and the original English items, and 0.71 for the individually translated items by Google Translate which indicate a higher similarity between the translated items to the original items. A t-test was conducted and the resulting two versions of the test items were statistically different ($t = -20.1162, p < 0.001$). This result indicates that there is a significant difference between the use of Google Translator to translate items individually and the use of item model by non-expert. Google provides a more consistent translation as the CSI was 0.71 indicating a high similarity among the translated test items to the original English items.

Content Expert Translation versus Google Translate Translation

When the content expert performed the translation of item models, the items generated produced a moderate mean CSI of 0.48 in terms of similarity to the original English generated items. T-test was conducted to compare with the mean CSI from Google’s forward and back translation (mean = 0.71) and revealed a statistically significant difference ($t = -19.2103, p <$
Translation in AIG 11

.01). Similar to the results with non-expert, Google Translate provided a more consistent translation than the content expert’s translation.

Non-Content Expert versus Content Expert Translation

With the mean CSI of 0.47 for non-expert translation and the mean CSI of 0.48 for the expert translation, a t-test was conducted to determine if any significant difference exist between these two item sets. The result revealed no significant differences between the two means. This finding indicates that there is no significant difference on their degree of similarities between the human translators with the original items. Although it would be expected that the expert translation is at a higher quality than the non-expert translation, t-test on the CSI reveals that the items are similar across the two groups. At this stage of our research on reliability of translation, content specialist do not seem to provide a more consistency translation to the original items in comparison to the translation done by the non-expert. Without a method of evaluating quality of translation, no validity conclusion can be made on whether content expert can be omitted in the process of item model translation. Hence, further research on this topic is required.

Conclusions and Discussion

In this study, the translation of a medical item model was performed by a humans and machines. The idea of using machine translator was introduced to perform item-by-item translation in comparison to item model translation. By using the approach of item model forward translation using content specialist, non-expert, and the novel method of translating individual items using machine translator, the results from the present study demonstrate that: (1) as items of multiple languages are being generated using n-layer modeling, the generated items tend to match the original items moderately (with mean CSI ranging from 0.47 to 0.71), which suggest that translation using item models do provide some similarity; (2) Google Translate is
very consistent in its translation performance in comparison to human translation, and (3) the differences between the expert and non-expert are small which may mean that content expertise is not required for creating the linked elements in multilingual AIG. This argument was supported with the higher mean CSI of about 0.71 when Google Translate performed the forward and back translation of individual items. Hambleton (1993) claimed that bilingual translators should be experts in the content area. Therefore it would be expected that the content expert would provide a better translation and possibly more consistent translation than the non-expert. However, at this stage of our research, the calculation of CSI only allow us to conclude that there was no significant difference between the consistency of translation between content expert and non-expert.

It should be note that consistency does not represent the quality of translation using an item model. This study serves as a first step in exploring the process of translation using item models, and to evaluate the products being generated. This study is significant in providing an alternative and, possibly, more efficient method for test translation with the use of item models. This study also demonstrated the effectiveness of AIG and the use of item models, which could benefit further in the field of Psychometrics, test development, and test/item translations.

**Limitations and Directions for Future Research**

There are important limitations in this study. First, one cannot be sure if Google Translate is the best available machine translator for back translating the items created by content expert and non-expert. Therefore, any discrepancies due to translator problems or due to differences in human translator were not identified. Future studies may compare performance of different translation engines. Future studies may also have content experts back translate the items instead
of having machine translator to do so for further examination on the performance of human versus machine translators.

Second, in this study Google Translate was used to translate items but not the item model. In future studies, it would be interesting to see how Google Translate would perform in comparison to human translators when the translations were done at the item model level.

Third, Google Translate was not able to pick up the difference between content expert and non-expert translation, as both obtained similar mean CSIs. Although translation of content expert seem to slightly outperform the non-expert in terms of consistency, this difference was not statistically significant in our study. These insignificant differences may be due to the structure of the item model, in which translation requiring no knowledge of the subject matter, such as patient’s background information and context, could be done in an easier manner by a non-expert. However, the slightly higher CSI found in content expert suggested that such difference may be found in the translation of medical terminology, which cannot be done perfectly by the non-expert. Until a study designed to evaluate the validity and the quality of the translated item model is conducted, we are not able to make validity conclusion using the results from the current study.
REFERENCES


Table 2

Sample of Generated Items using Different Translation Approach

<table>
<thead>
<tr>
<th>Original English Item</th>
<th>Items Generated from Non-Expert Translation</th>
<th>Google Translate from Non-Expert Translation</th>
<th>Items Generated from Content Expert Translation</th>
<th>Google Translate from Expert Translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 21. A 25-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 vitals. What is the best next step?</td>
<td>33. 一名 25 歲的女病人在左側腹股溝出現一團組織。病徵已延續了幾個月。經檢查後，那組織是凸出而不痛，化驗結果顯示命脈正常。那最佳的下一步是應該怎樣?</td>
<td>A 25-year-old female patient in the left groin, a mass organization. The symptoms had lasted for a few months. After inspection, that is a bulge without pain, and test results show that the lifeblood of normal. Under the best one should be what?</td>
<td>33. 一名 25 歲的女患者在左側腹股溝出現一個包塊。徵狀已持續了幾個月。經檢查後，那包塊是突出而不疼痛，化驗結果顯示生命體徵正常。下一部最佳處治是那一個?</td>
<td>A 25-year-old female patient in a mass in the left groin. The symptoms have lasted for several months. After inspection, that the mass is highlighted without pain. Laboratory results showed normal vital signs. Best Treatment of the next one is that one?</td>
</tr>
<tr>
<td>2 25. Patient presents with a mass in the left groin from a few months ago. The patient is a 25-year-old woman. Upon further examination, the patient had normal vitals and the mass is protruding but with no pain. What is the best next step?</td>
<td>34. 一名病人的左側腹股溝從幾個月前出現一團組織。病人性別女，25 歲。經長細檢查後，病人命脈正常，而那組織是凸出而不痛。那最佳的下一步是應該怎樣?</td>
<td>The left groin of a patient's emergence from a few months ago a group of organizations. Patient's offerings do not female, 25 years old. After slenderness check the lifeblood of the normal, the organization, and that organization is a bulge without pain. Under the best one should be what?</td>
<td>34. 一名患者的左側腹股溝從幾個月前出現一個包塊。患者性別女，25 歲。經身體檢查後，患者生命體徵正常，而那包塊是突出而不疼痛。下一部最佳處治是那一個?</td>
<td>A mass in the left groin of a patient from a few months ago. Patients offerings do not female, 25 years old. After physical examination, the patient's vital signs are normal, and that mass is prominent without pain. Best Treatment of the next one is that one?</td>
</tr>
<tr>
<td>3 29. A woman was admitted with pain in the left groin from a few months ago. There is protruding but with no pain in the left groin and the patient had normal vitals. What is the best next step?</td>
<td>35. 一名女子因幾個月前左側腹股溝出現痛楚。在左側腹股溝上感覺到凸出而不痛，而病人的命脈正常。那最佳的下一步是應該怎樣?</td>
<td>A woman a few months ago left groin pain. Feel the bulge without pain in the left groin, the lifeblood of the patient's normal. Under the best one should be what?</td>
<td>35. 一名女子因幾個月前左側腹股溝出現疼痛而入院。在左側腹股溝上感覺到突出而達不疼痛，而患者的命脈體徵正常。下一部最佳處治是那一個?</td>
<td>A woman was hospitalized a few months ago left groin pain. Highlight without pain felt in the left groin, the patient's vital signs were normal. Best Treatment of the next one is that one?</td>
</tr>
<tr>
<td>4 6. Patient presents with a mass and intense pain in the right groin from a few hours ago. The</td>
<td>77. 一名病人的右側腹股溝從幾個小時前出現一團強烈痛楚的</td>
<td>A patient's right groin from a mass organization with a strong pain in a few hours ago. Patient</td>
<td>77. 一名患者的右側腹股溝從幾個小時前出現一個有強烈痛感的包塊。</td>
<td>Right groin of a patient from a few hours ago, a strong pain mass. Patients offerings do male,</td>
</tr>
<tr>
<td>5</td>
<td>10. Patient complaints of a mass and intense pain in the right groin which has been a problem since a few hours ago. With normal vitals and tenderness in the area, the patient is otherwise normal. What is the best next step?</td>
<td>78. 一名病人正憂慮幾個小時前在右側腹股溝出現痛楚的組織。病人生命體徵正常，加上受影響範圍軟的，病人大概一切正常。那最佳的下一部是應該怎樣?</td>
<td>25 years old. After physical examination, the patient's vital signs are normal, and that mass is not rigid. Best Treatment of the next one is that one?</td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>6</td>
<td>14. A man was admitted with pain in the right groin from a few hours ago. There is tenderness in the right groin and the patient had normal vitals. What is the best next step?</td>
<td>79. 一名男子因幾小時前右側腹股溝出現痛楚。在右側腹股溝上感覺到軟的，而病人的命脈正常。那最佳的下一部是應該怎樣?</td>
<td>A man was hospitalized right groin pain a few hours ago. Rigid feel right groin, the patient's vital signs were normal. Best Treatment of the next one is that one?</td>
<td></td>
</tr>
</tbody>
</table>
Appendix

Illustration of Study Design

a) Non-expert translation versus Google Translate translation

b) Content expert translation versus Google Translate translation
c) Non-content expert versus content expert

Diagram:

1. English Item Model → Chinese Item Model
2. IGOR then Google
3. Compare with
   - English Items Generated from Original Model
   - T-Test
   - CSI

Variations:

- Content Expert
- Non-Expert

Diagram:

1. English Item Model → Chinese Item Model
2. IGOR then Google
3. Compare with
   - English Items Generated from Original Model
   - CSI