Institute for Credentialing Excellence

October 14, 2020

Presenter:
Joy Matthews-Lopez, PhD
President, JML Measurement and Testing Services

Link to Interview slide deck

Q: I’m new to this space, so can you tell me what the difference is between test translation, localization and adaptation?

My pleasure. I am happy to help.

Perfect. So Joy, we'll go back to the slide deck you were talking about, about the example that we wanted to actually highlight as part of the test adaptation.

Do you mind if I share my screen?

No, please do.

This is a PowerPoint deck I created about a year ago when I was preparing to present an ICE webinar on this same topic. I believe that date was in October 2019. Can you see my screen?

Yes, we can.

Great. As you can see, the title of this webinar was exactly what you are asking about, an introduction to test translation, localization, and adaptation. I am going to expand my screen so that it will be a little bit easier for you to see.

Keywords

In this slide, I was playing with a mind map of the key words that come into play when adapting an assessment tool. When we talk about test adaptation, these are many of the key words we associate with translation, adaptation, and localization: equivalence, aspects of reliability, validity, fairness, and globalization. The learning objectives of that webinar were not set up to address the specific questions you are asking today, but they do include the differences between test translation, localization, and adaptation. When our call is over, I will send you the webinar’s slide deck so you and your team have it. I believe that ICE also has it somewhere, but this way you will have it at your fingertips. The main example that I mentioned to you earlier, Brenda, is illustrated in this PowerPoint deck and is what I am going to page through for you today. As you can see on this slide, I provided a brief description of translation versus adaptation versus localization and highlighted what the differences are. In the webinar, I started with an example. I have learned from experience from speaking on this topic at NCME, ICE and ATP that it usually helps if we just start with an example, and this is my favorite one to share. This example comes from a battery of career planning tests that were designed to be administered to 10th grade students in the United States and then eventually in Mexico.

An Example

The purpose of the battery was to assist students with vocational planning, to help them think through academic and non-academic options. The business person who owned this battery of tests or owned the company that owned this battery had a business partner in Monterrey, Mexico. As that partnership grew, and they wanted to adapt this English-based battery of tests to work for the equivalent of 10th grade students in Monterrey, Mexico, and eventually to this grade level of students throughout Mexico and possibly into other areas within Latin America. The target population for this work was 10th grade high school students, or the equivalent of 10th grade students.

In the example I am sharing, the examinees were presented with the following list of five words:

juice, breakfast, grapes, jelly, and vines. Response options included four images, which I'll show you in just a second. The students were instructed to choose the image that best depicts the given set of words in some kind of descending, logical order. Again, the words used were breakfast, grapes, juice, jelly, and vines. And these are the four diagrams that the students were provided; and they needed to choose one. The key (correct answer) is Option A. From the vines we get grapes and from grapes we get juice and jelly, which are then consumed at breakfast.

P-Values in Testing

Approximately 600 students responded to the English version of this item. A classical index of difficulty, the p-value, was computed and was approximately 0.79, which means that about 80% of the students who saw this question in English answered it correctly. The point-biserial discrimination index was computed to be 0.21, which suggests that some of the psychometric assumptions we are making are reasonable. Those assumptions include the idea that higher-performing or higher-ability students tend to get the question right and that lower-ability or lower performing students tend to get the question wrong.

Both difficulty and discrimination indices were within expected ranges, so there was no evidence to suggest that there was anything wrong with this question. We used a forward and back translation methodology to translate this (and other) items from English into Spanish. Juice translated into jugo, and jugo back-translated into juice. Breakfast translated into desayuno, jelly into jalea, grapes into uvas, and vines into vid, and all back-translations matched their respective source (English) counterparts. There were no apparent problems with the translated words. This slide contains the translated version of the item. As with the source item, the key is in position A, and all options are in the same order as are the word flows within each option.

There was nothing different between the English and Spanish version of this item other than use of translated words, however, when we tested the Spanish version of this question, it flagged as having poor item statistics. I should note that this was one of many questions that were being piloted. We were piloting many items across the different tests within the battery. But the items statistics here really jumped at me. The item statistics are provided in bold, red font. P-values should have been similar between English and Spanish, a value around 0.8. Note, however, that on the Spanish version of the item, the p-value is 0.15. Not only is this value very different from 0.79, but it is extremely low for a p-value. Because the item only has four options, the probability of randomly selecting the correct answer should be 0.25. Why would this item perform below chance?

What went wrong?

What could have caused the p-value to be 0.15? In addition to the p-value being drastically different between languages, the item's point-biserial correlation, which can be interpreted as a classical measure of discrimination, was negative. Negative discrimination indices always get my attention as they typically indicate a key error (or something egregiously wrong with the item). As we checked the key, we saw that it was correct. Option A was the indicated key in both the English and Spanish versions of the item. The next line of inquiry was to re-check the translation. As we can see, the translation was fine.

As is typical in the test development process, I convened a committee of certified bilingual subject matter experts (SMEs) to review all poorly performing items, which included this one. I met the SMEs at an in-person meeting in Monterrey, Mexico. As I mentioned before, there were other questions that had been flagged for review, but this question in particular was bothering me. I shared the item statistics with the team of SMEs and was very humbled with what happened next. Let me pause and explain that I speak some Spanish, enough to be comfortable with such simple, direct translations. What I missed was the cultural component. Again, this was a "duh" moment for me. I had lived for short periods in Guatemala and Spain and had made numerous visits to other Spanish-speaking countries and territories. I should have known what the SMEs pointed out: In general, people from the target population do not traditionally drink grape juice for breakfast. Instead, they drink orange juice.

As a result of our "ah-ha" moment, a suggestion was made to change the context of the question. Instead of centering the question on grapes, we shifted the focus to oranges. Instead of using vines and grapes, we used orange trees and oranges. The word "juice" was fine, but we replaced "jelly" with the Spanish word for marmalade to improve cultural fit. Juice (jugo) and breakfast (desayuno) remained unchanged. The adapted item flowed the same way as the translated item: from the orange tree we get oranges, and from oranges we get juice and marmalade, and then we consume them at breakfast. This is my favorite example of an adapted item. It is also a reminder as to the critical role that culture plays in the adaptation process and why bi-cultural, bilingual SMEs should be on test development committees of multilingual programs. The mistake was mine. Though I had used bilingual, bi-cultural SMEs to review the flagged items, I had failed to use equally qualified people to review the adapted instruments (exam forms) prior to administering the exam. I also learned how important pilot testing is when adapting instruments. The professional translators I used were not bi-cultural, and therefore were unable to warn about the cultural issue that tripped up this item; grape juice did not make sense to the target population.

Statistics

As we showed earlier, nothing was wrong with the translation. It was perfect. But when working in the testing industry, translation alone may not be sufficient because we are not just trying to translate the questions, we are trying to translate the *spirit* of the questions. We aim for functional equivalence. In psychometrics, we seek to preserve the target construct(s), that is, we want to preserve the thing that we are trying to assess knowledge of or mastery over. From the above example, it is clear that sometimes a strict translation will not suffice. This item provides an example where adaptation was absolutely called for. The fruit used in the example was independent of the target construct (being able to order a set of words in a logically descending order). Once we changed from grapes to orange, the adapted item assessed the same construct as the English-based item and did so at the same relative level of difficulty (and discrimination).

The Importance of Localization

I believe the remainder of the webinar slide deck answers many, if not, most of your questions. What I have listed here are some of the important decisions and considerations to think through and questions to ask: Does the situation call for translation? Adaptation? Or will localization be the best path forward? In terms of localization, let us say that I have a test in English that is used across the United States, and I want to use the same exam form in England. Will translation be needed? No. Will adaptation be needed? Maybe, depending on the items. Will localization be needed? Probably yes. Of course, this will depend on the specific test, the target population, and the intended use of resulting scores. Displays of dates and times may need to be rewritten, as may certain measurements (weights) and the spelling of some words. Regarding dates, consider the date February 4th (2/4/2021). Might this be confused with April 2nd (4/2/2021)? Localization may or may not involve translation and it may or may not require adaptation, but it certainly requires fitting the exam to the target population.

Important decisions to Make

When asked a question, psychometricians often respond “It depends.” This is because context almost always matters. The following is a list of often asked questions to think through before beginning an adaptation initiative. Note that several of the questions have already been mentioned in this document, but for the sake of convenience, they are included here, too.

What is the purpose of the work? How will the target instrument be used? Will scores from source and target instruments be compared or used interchangeably? Will the same cut score be applied to source and target forms? Will the same amount of seat time be allotted for source and target exams? Will new items be written in English and then translated/adapted into the target language(s)? If so, will bilingual SMEs be included as item writers (and reviewers)? Will bilingual SMEs serve on other test development committees, such as JTA or standard setting activities? Will bilingual SMEs also be bi-cultural? Which translation model will be used? Will translators also be SMEs? Do constructs from the source product exist in the target population? Are item types on the source exam equally familiar to the target population as they are with the source population? Are resources available to support piloting the target form? What are the target languages and/or cultures? Do the same security risks exist in the source and target settings? There are some target populations where we have learned from experience to be more careful with test security. We are always vigilant, but from lessons learned, now we know to be obsessive. Also, logistical challenges need to be considered, such as delivery options. Will the exam be delivered via paper and pencil vs. computer? If paper-based, then will printed materials need to be shipped internationally? How will testing across time zones be handled? Answers to these (and other) questions will drive decisions about translation, adaptation, vs. localization, and the appropriate test development and translation models, if needed.

The Use of Committees in Test Development

Let us follow up on something we just discussed: SMEs and test development activities. As you know, in test development, we have JTA panelists, item writers/reviewers, standard setting panelists, and SMEs on other standing committees. How many subject matter experts should be used for the various test development activities? What proportion of those should be from the target language/culture versus the source language/culture. Each of these questions need to be answered.

Another important consideration is program documentation. Be sure to thoroughly document who the translation provider is, and if it's going to be done in-house, then be able to provide documentation and evidence of the necessary credentials, experience, and expertise needed to be considered qualified translators. I urge caution on using “convenience translators” such as certificants that happen to speak the target language(s). Remember our earlier example: just because I was able to speak Spanish and had lived in a couple of the target countries, I still overlooked the significance difference between grapes vs. oranges. Using services of bilingual, bi-cultural subject matter experts to review constructs, create a working glossary prior to the translation process, and to review and approve translated materials is strongly recommended (see International Test Commission Guidelines (v 2.4) for specific recommendations), but not in lieu of using professional translation services.

Also, the opinion, as good as it may be, of one person does not generalize well enough to base translation and/or adaptation decisions. This is why we have committees of 10+ people doing certain test development tasks. Clearly, the translation services provider plays a vital role in the adaptation process. However, the secret to the sauce is the combined use of translators and bilingual, bi-cultural SMEs and knowing what to translate and who to review and approve it once in place.

Other Issues to Consider

In addition to individual items or intact exam forms, will all banked items be translated/adapted? Will language services only apply to items that make it onto a source exam form? Will a glossary or list of key words be prepared for the language service provider? Will forms be created in a source language and then translated into the target language(s)? Which version of the form will serve as the base form for standard setting and equating? How will the review of flagged items be handled and/or replaced? If poorly performing items are not going to be replaced due to bad statistics or otherwise poor performance, then how will adjustments to scoring be handled (and reported)?

So far, we have focused our conversation on written tests. It is important to also ask what approaches will be viable for performance-based assessment? Can or should testing environment be adapted or localized? Will there be implications to the test specifications? For example, consider people who work in water safety. Are the same tools of the trade used in different countries? Do such tools have the same names as called out in the JTA, blueprint, test specification document, or sanctioned reference materials? Will test materials for EMTs differ across settings, tools such portable defibrillators or types of IVs or intubation tubes? Beyond translation, will different equipment be considered standard in different settings? We need to reconcile such differences well in advance of the translation process, along with other things may need to be reviewed by bilingual SMEs, such as the exam blueprint and the constructs of interest. These are very big issues that must be addressed.

Reconciling the Test with the Blueprint

I have worked projects where the source blueprint did not fit with the target population in terms of language and/or culture. Obviously, this was a serious, foundational issue because of the potential impact on the validity argument. As mentioned earlier, there is a concept called de-centering, which is where the test specifications need to be adjusted to make sure that the blueprint and test conditions are equivalent across all versions of the assessment product, including exam forms in the source and the different target languages and cultures. If scores are to be compared, then it is critical to make sure that the same constructs are being assessed. Unfortunately, there are times when the construct from the source language/culture just doesn't exist in the target language/culture. I will give you an example of this in just a minute. Another point I want to make is about the need to translate and/or adapt scoring rubrics and support materials, such as the candidate handbook, website, policies and procedures that are candidate-facing, proctor instructions, and score reports.

Also, it is possible that scores will be reported for the source and target independently. If this is the case, be prepared to document that policy and provide a rationale for it, especially if the program is seeking some type of accreditation. Always be prepared to provide documentation and corresponding rationales.

International Test Commission and Other Resources

And let me give a shout out to the International Test Commission. They have fantastic guidelines in place that are very user-friendly and easy to operationalize. Other valuable sources of guidance include the NCCA Standards, the ICE Handbook (3rd Ed.), and the AERA, APA, & NCME Standards for Educational and Psychological Testing (2016). Each of these resources should be reviewed by any organization that is serious about moving forward with test translation and/or adaptation.

Use of Bilingual Subject Matter Experts (SMEs)

So again, how many bilingual SMEs are needed throughout the test development process? Allow me to share another example? About a year ago I was consulting for a company that wanted to translate/adapt a non-cognitive assessment from English into six other languages. The original instrument was designed to help workers use a self-assessment instrument to better understand their perceptions and attitudes in the workplace. Resulting scores were used to identify areas of potential personal growth that could result in increased happiness and satisfaction with their work.

This company put a lot of effort into their product; the assessment tool was designed to help their workers. The original tool was developed in English and was validated on English-speaking subjects within the United States. Prior to making the decision to offer the tool in other languages, the English-based version had been used within the United States for many years and had undergone extensive research. Wanting to offer the tool to peer institutions in the international market, the company decided to adapt the tool from English into an equivalent product for use in Spanish, French, Russian, German, Portuguese, Italian, and Mandarin Chinese. I was hired to adapt the tool from English into Simplified Chinese and to use that model and work to guide the adaptation process for the other target languages.

An interesting issue we encountered during the adaptation from English into Mandarin had to do with the word “co-worker.” The subject matter experts chosen for this work were highly qualified in terms of content; each held an advanced degree in psychology (everyone held at least a Masters degree, though most held a PhD), and they were fully bilingual and bi-cultural. In addition, they had work experience in both China and the United States, so they had an understanding of the workplace in both cultures. The issue they encountered centered on the word “co-worker.” In English (and a US setting) the word coworkers may refer someone that works at the same company, someone we may encounter in the hall or breakroom at work. The word coworker does not necessarily convey a type of work relationship, such as someone I report to or someone that reports to me. Someone can be my coworker without actually working (directly) with me. For example, a coworker could be a person I work with, including my supervisor or someone in my same peer group or someone that reports to me. And though even if someone reports to me and I am their superior, we are still coworkers. Something similar can be said about the relationship between me and someone I report to. Herein was the issue we encountered; the SMEs strongly advised to use three different words to translate “coworker”, one for each type of work relationship (someone’s subordinate, someone’s superior, or someone’s equal (aka, colleague)). There was no such single, generic word equivalent to coworker (the way that it was being used in English) that we could use on this version of the test.

Interestingly enough, when we got into other languages, we saw the same problem when trying to translate “coworkers” into German and Russian. Unfortunately for our work, this word appeared in probably 70% of the items on the source form because the focus of the assessment had to do with working with one’s colleagues, aka, coworkers. Having so many of the questions contain the word coworker posed a serious challenge. As with Simplified Chinese, both German and Russian considered the hierarchy or direction of the reporting when choosing an equivalent word to coworker, as the type of relationship between the employees changed the meaning of the word. We had to adapt accordingly.

Psychometric Input

From a psychometric perspective, it makes sense to start an adaptation project by reviewing the job task analysis and associated blueprint and test specification document. In the project that I just mentioned, a lot of time was spent reviewing the blueprint and the specifications in English to make sure that all the key words and concepts actually existed in the target language. If the key words and concepts don't exist in the target language, then even the best translation possible will fail. Think back to the example of grapes versus oranges. Translation was not the problem. If there is no equivalent to key words and/or concepts from the source language and/or culture into the target language and/or culture, then it is very likely that the initiative will have a problem with construct validity. So where should the initial focus be, from a psychometric perspective? The job/task analysis, exam blueprint, and test specifications will be at the heart of it. Next in line will be the exam development process.

Use of Source Language

There are going to be issues to address and questions to ask when planning an adaptation project. For example, should item statistics be carried in both source and target languages? In the examples I have given today, the source language has been English, but that is definitely not always the case. At the risk of seeming English-centric, let's say that my source language is English, and I want to adapt into German, but eventually I also want to adapt into French and Russian. What is the best way to proceed? Along this line of questioning, which set of data should item analyses be conducted, and should resulting information be preserved (and reported)? Should the focus be on item stats from the source exam? Perhaps the answers to these questions will change from one company to the next, depending on varying test development policies, but this is an area where psychometric direction may help. Keep in mind that if the program is accredited, then great care should be taken to provide rationales for key decisions and policies/procedures; be prepared to justify everything.

Equivalence Between Forms

Typically, the heart of adapting from one language and/or culture to another involves establishing equivalence between the exam forms (content) and resulting scores. In many cases, we also need to establish equivalence between the source and target constructs as well.

There are established methods for investigating and collecting data on equivalence. We can build a body of evidence that includes a detailed look into the underlying constructs of the source (and target) exams. Included in this evidence should be documentation and justification of any translation models used and the composition of all key test development panels, such as those involved in the program's job/task analysis, item writing, item reviewing, and standard setting workshops. ITC Guidelines, as well as NCCA Standards and AERA/APA/NCME Standards point to the importance of using qualified and representative panels for test development work. This point cannot be over emphasized when testing in multiple languages.

There are times when resulting scores need to be placed on the same scale, are subject to the same cut score, and have the same interpretive value so that they can be compared. If the point is to compare scores, then this objective needs to be front and center during the test development and adaptation processes. I have worked on programs where the end goal was not about comparing scores. Instead, the goal was to create a stand-alone instrument based on a source exam that could work independently in a target setting. In the example I am thinking of, there was a need for an instrument to work in Saudi Arabia but to align with the JTA of a US-based program. The point was not to compare resulting scores to examinees who tested on the US-version of the form; that was not the goal. Instead, a parallel assessment tool was needed and by using the source exam as a base, the client did not need to start from scratch. It was appealing to this client to use an instrument that had already been vetted and was well-respected. In this case, we did not need to establish equivalence in the usual way(s).

Technical Issues and Response Time

There will always be technical issues that psychometricians need to address. For example, the psychometric model of choice will need to be determined, as will test assembly, delivery and proctoring procedures, targets, and models. Performance goals will also need to be determined for examinees, items, and exam forms, such as person fit, item fit, and form level metrics.

When possible, I prefer for my clients to administer/deliver their examinations via computer. Doing so not only allows for different types of items to be used, but computer-delivered tests (CBTs) yield response time. This type of data is extremely useful for a variety of reasons. First and foremost, response time data can be used to assess whether an assessment is speeded. This is particularly important to check when testing across multiple languages because speededness can impact the underlying validity argument. My dissertation addressed this particular area of test adaptation (speededness). I was in a very fortunate position where I had enough funding from the sponsoring organization to allow me to investigate this topic. I was curious about response time across languages and wanted to know if speededness was impacting pass rates on the exam that I was working with. Since then, I have worked with numerous multilingual programs that had significant differences in pass rates between language versions, so now I am always sure to analyze response time and check for possible impact to seat time. It is expected that pass rates between translated/adapted instruments will be similar, but when they are notably different, then we need to find out why and address it (if possible). If source and target populations are assumed (and have been shown) to be equivalent, then why would there be a significant difference in pass rates? One plausible explanation may be academic preparedness; other plausible explanations include unfamiliarity with certain item types, lack of or differential experience with technology, cultural aspects of test taking, and/or differences in practice settings. It is critically important to investigate, identify, and document known or suspected factors that may impact scores, in addition to actions taken to address and mitigate such factors. When pass rates differ between source and target exams, the obvious places to look for problems include the quality of the translation(s) and equivalence of the administration, which includes response time metrics.

My Dissertation

The research I conducted for my dissertation was limited to analyzing an adaptation from English into Spanish. Early data analysis results indicated a significant difference in pass rates. Given that eligibility requirements were the same for members of the source and target groups, it surprised me to see a large difference in pass rates. After analyzing response time, I realized that the Spanish version of the test was speeded. Upon reflection, this made sense given the higher word count and different (perhaps more complex) grammatical structure in Spanish. There were more words in the adapted instrument), so the reading load was heavier, so it made sense that more time may be needed to read, process, and then responds to each item. Depending on the languages involved, it may take more time to consume and process the input and then in turn provide output (answers to questions). This being said, it is also possible that less time may be needed, so be sure to analyze response times.

As a result of this research, I was able to provide quantitative data to show that the test of interest was more speeded in Spanish than in English. This is an area of research that I hope will continue. For certification programs, I strongly recommend bringing in a psychometrician for help and guidance.

Test Specifications

In terms of test specifications, I would like to call your attention to some important questions to ask and answers/issues that should be documented:

- What is the purpose of the assessment?
- Why does the assessment tool need to be adapted?
- Who are the source and target populations?
- What languages will the assessment tool be offer in?
- What evidence of construct equivalence already exists?
- Will the same blueprint (tasks and weights) work for the target population(s)?
- Are there any known differences in terms of test environments, delivery mode, item formats, familiarity with technology, or anything else that may negatively impact equivalence?
- How will the source and target instruments be linked? What will happen if an item flags in the source or target language (for poor performance)? How will score equivalence be established if an item needs to be removed from one version of the exam?
- What type of equating model will be used (if needed) to link scores on source and target exams? Will the same cut score be applied to bother versions?

- Will the same item types be used? We know that unfamiliarity with item types can impact scores. Even how items render on the screen can make a difference. I believe this was one of the taskforce’s questions. I am but one person, and my choices may not generalize well to everyone, but in my experience, it has been very helpful to present items in both the source and target language at the same time (as opposed to toggling back and forth between them). I believe this is especially helpful within the US (Spanish) testing population because although Spanish may be someone’s first language, it is likely that formal education was received in English.

Defining the purpose of the test

I cannot over emphasize the importance of proactively defining and documenting the purpose of the adapted exam. Will scores on the source and target exams be compared or will each be designed to fit their respective audiences? For example, in an example we discussed earlier, an adapted exam was to be used exclusively in Saudi Arabia. Scores on the adapted instrument were intended to stand on their own, so although the original blueprint (in English) was used to assemble exam forms in Arabic, the two exams were otherwise independent. In such a case, there was no need to display items in any language other than the target language. Keep in mind the issue of reading loads and seat time. When items are simultaneously displayed in two languages, or when candidates need to toggle back and forth between source and target renderings, then more time may be needed to maintain an equivalent test session. If time and financial resources permit, then consider piloting adapted instruments before determining seat time.

De-centering

I do not recall where I first saw this word (de-centering), but I use it often when talking about adaptation. In photography, decentering refers to moving or tilting a lens from its principal axis. In the context of educational measurement and adaptation, it refers to adjusting an exam’s blueprint or specifications to improve fit in the target setting. In other words, we “decenter” these tools so that they are not centered in the source environment.

Applying Standards

Many NCCA Standards are relevant when adapting instruments. Standard 13 addresses subject matter expert (SME) panels, how well those panels generalize to the target population, SME qualifications, and responsibilities entrusted to SMEs. This Standard requires that JTA panels, content development panels, and standard setting panels have adequate representation from the target population. When adapting assessment instruments, it is important to have bilingual, bicultural SMEs on these key panels.

Standard 14 relates to JTAs, so again, having representation on the panel as well as appropriate representation in the validation survey will provide valuable insight.

Standard 15 speaks to exam specifications. This document will describe the target populations, the exam blueprint (and a decentered blueprint, if appropriate), item types, delivery models, proctoring models, and seat time, among other things.

Standard 16 is particularly relevant when adapting exam forms. Essential Element B specifically speaks to translation processes.

Standard 17 discusses standard setting and establishing performance standards. Clearly this Standard is relevant to programs that use adapted exams since important decisions will need to be made (and documented) regarding how the cut scores for the source and target exams were established.

Standard 18 also very relevant for adapted instruments as it addresses standardization, issues of equivalence, and exam security.

Standard 19 addresses scoring and score reporting. Will the source and target instruments be linked together? Will scores be reported on the same scale? Will resulting scores be comparable? This Standard speaks to the scoring model(s) that will be used for the source and target exams.

Standard 20 addresses technical issues such as reliability (internal consistency at the form-level as well as decision consistency), and calls for evidence of reasonable item-, person-, and form-level performance metrics. It is here that quantitative evidence of equivalence between source and target instruments (and/or populations) is called for.

Standard 21 addresses equating and how scores from subsequent exam forms will be interpretable. Essential Element 21C specifically calls for evidence of score equivalence when dealing with adapted or translated examinations.

This is not an exhaustive list of NCCA Standards to consider when translating and/or adapting exam forms, but it is a good subset to consider.

Use of an item banking tool and edits

An operational issue that needs to be considered when translating or adapting items is the content management system (aka, item banking tool). This tool must be able to support multiple language versions of the same item. Typically, the item banking tool for translated/adapted items will need to support characters specific to the source and target languages, as well as keyword searches and dictionaries. In addition, the system will need to support multiple sets of item statistics and provide a link to toggle between language versions of items. Ideally, the system should have the capability to provide support for candidate comments in the source and target languages. Without such features, it would be nearly impossible to know if an item has a typo (in any of the involved languages). Other essential functionality includes being able to capture and track item history. This can get tricky when dealing with multiple versions of the item, so it helps to be able to track item history on all the different versions and to track edits of the different versions.

I have one more thought about the use of special characters in the item banking tool. Sometimes small things can make a big difference. For example, consider accent marks or umlauts. Without these being placed correctly (and rendered correctly on screen), the result may be a totally different word than what was intended.

Establishing Equivalence

How is equivalence established? And what does “equivalence” entail? Establishing content and construct equivalence is critically important when resulting scores are to be compared. In my experience, programs new to translation/adaptation tend to focus on language. Clearly, having a defensible translation model and excellent translation services are necessary for any multi-language program. But the conversation must continue beyond translation and include construct equivalence and cultural acumen. Remember the example about grapes and oranges? Even though the translation was correct and an argument could be made in support of content and perhaps construct equivalence, it was a cultural consideration that initially caused that item to fail. Content matters, constructs matter, and culture matters.

Going back to my dissertation, one of the example items consisted exclusively of images. No text was involved, just pictures. The item consisted of five adjacent images and the instructions were to select three images that formed a theme and indicate what the theme was. At the time, I expect this set of items to be super easy to work with (no words to translate), but I was wrong. Several of the items flagged for being exceptionally hard. In my defense, the work that I was doing was well before my dissertation, so I was still learning about adaptation. In this item, the images were of a pine tree (a standard evergreen), a worn-looking log cabin (the thatched roof had a hole in it), a pair of overalls (work clothing, somewhat worn), a flagpole, and a (stereotypical) image of a grandma. The keyed response included the images of the log cabin (“old” cabin), overalls (worn and tattered), and the grandma; the theme was “old things.” Clearly this item should have been sent reviewed for sensitivity and bias, but that’s a different story. The item statistics were fine when included on the source exam (English-speaking 10th grade students) but they were terrible from the target form (10th grade students in Mexico). When this item was presented to a panel of subject matter experts in Mexico, the looks I got spoke a thousand words: how could we have possibly considered a grandma as an “old thing”? I remember this example when talking to others about equivalence. The conversation cannot be limited to content; constructs, culture, and context matter.

Test statistics and source languages

Depending on the program’s design, I may or may not carry items statistics in both the source and target languages. Depending on the project, I usually assemble forms in the source language and then translate/adapt them via a professional translation service provider. I then use bilingual, bicultural SMEs to review all items prior to assembling exam forms. I tend to overbuild in terms of content (items), so that there are more items than needed in case there are translation issues. Sometime, items simply won't translate well. Any pair of items may be enemies in one language yet not enemies in another language. In such a case, the source form may need to be changed. Be prepared to allow additional time and bandwidth to the project timeline for form assembly and review. This process is not trivial. It requires focused effort.

Some Advice

One approach to establishing content and construct equivalence is to use statistical methods and models. If an exam has been piloted and there are sufficient data to support use of item response theory (IRT), then test information and characteristic curves can be produced. Comparing such curves between source and target items (and exam forms) and looking at fit statistics can be very useful when screening for differences. Combining information from IRT analyses with classical analyses can yield stable and useful statistics.

Some useful questions to ask include the following: Are the pass rates similar between source and target groups? Are the cut scores comparable? How are scores being equated? Is equating limited to forms in the source language? Are source and target population sizes similar in size? Are there any cultural differences between source and target settings regarding definitions of cheating and attitudes about test security? There are some cultures where cheating is not considered a bad thing. Instead, it may be acceptable behavior and framed in a positive light, such as being clever or resourceful.

Other questions may include the following: Is support provided to candidates in their own language? Are support materials, such as the candidate handbook, registration materials, and program website offered in the target language? Has the program's budget been adapted along with the assessment tool in order to support new staff, or more robust content management tool, or the inclusion of extra subject matter experts?

In terms of advice, don't make assumptions, stay closely aligned to standards, guidelines, and trusted resources (ITC Guidelines, NCCA Standards, ICE Handbook), and secure the necessary resources (staff and budget) to do things well. Use professional services to guide and support test development, program maintenance, and requisite documentation.

Q: This has been such a wealth of information. I really appreciate that. Everything you've given is absolutely amazing. And I really appreciate your time. One of the things that I'm curious about, that hasn't been addressed, because you were very focused on like, if you're going to do the test, this is how this is like the processes and stuff like that. My question to you is our conversation has been reflective of taking an English test and moving into a localization language. Now that we're becoming much more international, have you seen the actual opposite happen, localization test having to be translated into English and is that process going to be about the same?

A: You talking about localization, not translation, right? Localization makes sure that units of measurement and timestamps and date stamps are correct for the local setting. Just to be clear, your question pertains to a test that was developed in another language and/or setting, say in Saudi Arabia for example, and now an organization wants to use it in an English-speaking market. Is this corect?

Throughout this presentation, I kept referring to the source language as being English, which is definitely not the default. It is very possible that an instrument that was developed outside the US in a language other than English needs to be adapted to function inside the US and/or for an English-speaking audience. For example, we may want to compare student performance on a mathematics test between a source and a target audience. I have not worked on this type of project (where the source form was not in English), but yes, I believe the process would be the same as discussed in this interview. To-date, all of my source forms have been in English and the challenge has been to make them function in a non-English language and/or non-US culture. I have worked in the following target languages: Spanish, Portuguese, Italian, French, German, Mandarin, and Russian. I have never done it the other way around, but it is reasonable to assume that the same strategies would be appropriate. It should not matter what the source languages and/or cultures are in terms of processes. The same safeguards and Standards will apply; the same questions should be asked.

Q: In the process you talk through, is it more the process that is more important than the source versus target language.

A: Yes, I believe so. Again, I have not worked in that direction, but yes, the process should be the same.

Q: Yeah. I just have a one follow-up question. Adaptation is so complicated, it will involve a lot of resources to make sure not only the content is equivalent, but also construct if colorants have we kind of, based on your consulting work and experience in this area, we saw to keep the same source language, but just doing some accommodation instead of translate or accommodate

A: Do you mean instead of adapting? I think the answer to this question will depend on how high the stakes of the test are. If the stakes are low, such as a self-assessment with low-stakes consequences, such as information vocational planning (as opposed to medical licensing), then the focus may be more on culture and the mapping of test outcomes with different professions within the target population. However, if we are talking about being licensed in country X and the content domain is the same in the source and target populations (think nursing or certain types of emergency responders, such as paramedics), then I would think that the skillset needed in the US would likely be similar to the skillset needed in Italy. This is where the test development team would do their research. I cannot be so bold as to say the jobs are identical, but perhaps a large portion of the content overlaps. To find out, the test developers would begin with a strong group of bilingual, bi-cultural SMEs, the exam blueprint and test specs, and take things from there (including looking at all keywords and concepts).

You mentioned that this sounds like a lot of work. Developing quality exams, regardless of language and/or culture, requires a lot of work. The good news is that when adapting an existing exam, most of the infrastructure is already in place. Just as when we adjust designs to handle accommodations, whether for large print or the need for readers, we often can adapt for language and/or culture. From a technical perspective, I would still do everything needed to be able to compare scores (assuming that is the design). I'm not sure that I answered your question. Yes, adaption requires a lot of work, but as long as a strong program is already in place and the objective is to expand that successful program into a new market, then adaptation should not add an undue burden of work. As with any business expansion initiative, go into it prepared and with eyes wide open. The mistake I have seen made is moving ahead, taking a leap, before looking. Do due diligence and be realistically prepared.

I hope I have covered all the key points. My advice to any organization that is thinking to go down this path, is to design the business and test development plans realistically and put the necessary resources in place. There will be some extra work. Be realistic. Be prepared. Lean heavily on the Standards.

Close of Interview

Of course, of course. Thank you for reaching out, and thanks for this opportunity to talk with you about this area of test development. I love the subject. It is my favorite area of psychometrics to play in. I find it to be interesting and fun. Given my passion for psychometrics and my love of being engaged with other cultures, the area of adaptation fits very well with me.

It is always enjoyable to speak on this subject and to share my experiences and lessons learned. Thank you for the opportunity to share with you today.