International Journal of Nursing Studies xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

International Journal of Nursing Studies

journal homepage: www.elsevier.com/ijns

A systematic survey instrument translation process for multi-country, comparative health workforce studies

Allison Squires [a,*], Linda H. Aiken [b], Koen van den Heede [c], Walter Sermeus [d], Luk Bruyneel [c], Rikard Lindqvist [e], Lisette Schoonoven [f], Ingeborg Stromseng [g], Reinhard Busse [h], Tomas Brozstek [i], Anneli Ensio [j], Mayte Moreno-Casbas [k], Anne Marie Rafferty [l], Maria Schubert [m], Dimitris Zikos [n]

[a] New York University College of Nursing -Global Health Division, New York, NY 10003, USA
[b] Center for Health Outcomes and Policy Research, University of Pennsylvania, USA
[c] Katholieke Universiteit Leuven, Belgium
[d] Healthcare Management & Nursing Science, Katholieke Universiteit Leuven, Belgium
[e] Department of Learning, Informatics, Management and Ethics, Karolinska Institutet, Sweden
[f] IQ Healthcare, The Netherlands
[g] The Norwegian Knowledge Centre for Health Services, Oslo, Norway
[h] Department of Health Care Management, WHO Collaborating Centre for Health Systems, Research & Management, University of Technology Berlin, Berlin, Germany
[i] Department of Internal Medicine and Community Nursing, Institute of Nursing and Midwifery, Faculty of Health Sciences, Jagellonian University Medical College, Krakow, Poland
[j] Department of Health and Social Management, Shiftec Research Unit, University of Eastern Finland, Kuopio Campus, Kuopio, Finland
[k] Investén-ISCIII, Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación, Madrid, Spain
[l] School of Nursing & Midwifery, Division of Health and Social Care Research, Kings College – London, London, England, United Kingdom
[m] Institute of Nursing Science, Faculty of Medicine, University of Basel, Basel, Switzerland
[n] National & Kapodistran University of Athens, Athens, Greece

ARTICLE INFO

Article history:
Received 26 May 2011
Received in revised form 4 February 2012
Accepted 9 February 2012

Keywords:
Translation
Languages
Cross-cultural research
Health services research
Nurses
Nursing
Europe
Instrument validation
Content validity indexing

ABSTRACT

Background: As health services research (HSR) expands across the globe, researchers will adopt health services and health worker evaluation instruments developed in one country for use in another. This paper explores the cross-cultural methodological challenges involved in translating HSR in the language and context of different health systems.
Objectives: To describe the pre-data collection systematic translation process used in a twelve country, eleven language nursing workforce survey.
Design and settings: We illustrate the potential advantages of Content Validity Indexing (CVI) techniques to validate a nursing workforce survey developed for RN4CAST, a twelve country (Belgium, England, Finland, Germany, Greece, Ireland, Netherlands, Norway, Poland, Spain, Sweden, and Switzerland), eleven language (with modifications for regional dialects, including Dutch, English, Finnish, French, German, Greek, Italian, Norwegian, Polish, Spanish, and Swedish), comparative nursing workforce study in Europe.
Participants: Expert review panels comprised of practicing nurses from twelve European countries who evaluated cross-cultural relevance, including translation, of a nursing workforce survey instrument developed by experts in the field.
Methods: The method described in this paper used Content Validity Indexing (CVI) techniques with chance correction and provides researchers with a systematic approach

* Corresponding author. Tel.: +1 212 992 7074.
E-mail addresses: aps6@nyu.edu (A. Squires), laiken@nursing.upenn.edu (L.H. Aiken), koen.vandenheede@med.kuleuven.be (K. van den Heede), walter.sermeus@med.kuleuven.ac.be (W. Sermeus), rikard.lindqvist@ki.se (R. Lindqvist), anneli.ensio@uef.fi (A. Ensio).

0020-7489/$ – see front matter
doi:10.1016/j.ijnurstu.2012.02.015

for standardizing language translation processes while simultaneously evaluating the cross-cultural applicability of a survey instrument in the new context.
Results: The cross-cultural evaluation process produced CVI scores for the instrument ranging from .61 to .95. The process successfully identified potentially problematic survey items and errors with translation.
Conclusions: The translation approach described here may help researchers reduce threats to data validity and improve instrument reliability in multinational health services research studies involving comparisons across health systems and language translation.

What is already known about the topic?

- Forward and back translation alone is insufficient to produce reliable and valid translation of instruments.
- Methods used by researchers for cross-cultural adaptation of survey instruments developed in one country for use in another are inconsistent.
- Surveys describing illnesses or conditions translate more easily across cultures than the language of health system administrative hierarchies.

What this paper adds

- Describes a systematic process for conducting cross-cultural instrument translation and evaluation for health services oriented research.
- Quantifies where problematic areas may occur with adapting instruments for use in other contexts.
- Illustrates the importance of a highly structured approach to evaluating an instrument before data collection occurs.

In any research where multiple cultures are involved, unique systematic measurement biases may occur that affect the final results of the study (Gjersing et al., 2010; Thrasher et al., 2011; Van de Vijver and Leung, 1997). Researchers in health and social science disciplines have mainly focused on cross-cultural approaches for translating instruments used to measure patient psychology or symptomatology. In contrast, the translation of health services research (HSR) language is relatively new. Most HSR studies that involve instrument translation usually stop after the forward and back translation processes, even if multiple translators are used (Perneger et al., 1999). Therefore, methods for tackling the translation of structural and institutional differences that shape HSR remain inconsistent and largely untested, even though this is a significant research challenge for multi-country comparative studies of health systems and their workforces.

The RN4CAST study, funded by the Seventh Framework of the European Commission, aims to capture the state of the European nursing workforce through a twelve country comparative study across differently organized and financed health systems, different sociopolitical histories, and nursing professions in varying stages of professional development (Sermeus et al., 2011). The study required translating the survey instrument into eleven languages from its original American English version and ensuring its relevance to the nursing practice and healthcare contexts of twelve countries. This article describes the methodological approach employed by RN4CAST for the pre-data collection translation and evaluation of a health services research (HSR) survey instrument prior to conducting post-data collection psychometric evaluations. The study aimed to standardize and systematize the translation process across countries and develop a quantifiable measure of the cross-cultural applicability of an established instrument. Health services researchers who seek to conduct similar studies with healthcare professionals may find the approach useful for minimizing the systematic measurement biases resulting from language translation that often occur in any kind of cross cultural research.

1. Background

Only a few researchers have attempted to translate the administrative and role-based hierarchies found in organizations that comprise health systems or organizations (Choi et al., 2009; Gibson et al., 2003; Hyrkäs et al., 2003). When it comes to the language of health and health services delivery, subtle differences in the conceptual meaning of words can often create completely different survey question structures and alter language use (Mason, 2005; Peña, 2007; Ramirez et al., 2006; Temple, 2005). For example, the role of "director" may mean the person in charge of an entire hospital in one country whereas the same word could reference a middle manager in another. A survey question where the word "director" was literally translated without factoring in the meaning of that word in the context will change the content, construct, and concept the question tries to measure.

Despite multiple publications with a variety of translation guidelines (Beaton et al., 2000; Guillemin et al., 1993; Herdman et al., 1997), researchers generally tend to focus on the technical aspects of language translation and use only forward and back translation. Brislin's (1970) decentering method is perhaps the best known translation method. It emphasizes the semantics and technical aspects of translation during the forward and back translation process. Yet its limited focus on the technical aspects of the translation process and the absence of expert feedback creates some limitations for a full cross-cultural evaluation (Jones et al., 2001). Limiting the translation to simple forward and back translation techniques also generates several methodological problems (Maneersriwongul and Dixon, 2004; Tran, 2009). Maneersriwongul and Dixon (2004) statistically analyzed multiple studies that employed only forward and back translation approaches

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

Table 1
Definitions of cross-cultural validity in instrument translation and where they were addressed in the study's process.

Criteria	Definition	Process
Content equivalence	The content of each item of the instrument is relevant to the phenomena of each culture being studied.	• Researcher expert panel review • Practicing nurses expert panel review through the content validity indexing scoring process
Semantic equivalence	The meaning of each item is the same in each culture after translation into the language and idiom (written or oral) of each culture.	• Translation guide developed prior to translation to clarify the meaning of words found in the survey instrument • Separate, experienced translators conducting forward and back translation • Additional confirmation of translations through qualitative comments by practicing nurses expert panel
Technical equivalence	The method of assessment is comparable in each culture with respect to the data that it yields.	• Translation guide developed prior to translation to clarify the meaning of words found in the survey instrument • Separate, experienced translators conducting forward and back translation • Translation evaluation score
Criterion equivalence	The interpretation of the measurement of the variable remains the same when compared with the norm for each culture studied.	• Researcher expert panel review
Conceptual equivalence	The instrument is measuring the same theoretical construct in each culture.	• Translation guide development • Quality of translators forward and back translation • Dual scoring process with content validity indexing and translation evaluation scores

Adapted from Flaherty et al. (1988), p. 258.

and found multiple problems with the quality of translation that affected study results.

In the case of HSR, because the field began largely in English speaking countries, most instruments are products of their healthcare systems or the status of the profession when it was developed. They were not developed to be used outside their own borders or in other languages, even though it is a common (yet flawed) assumption in cross-cultural research that measurement is automatically equivalent across groups (Thrasher et al., 2011). Therefore, in studies where instruments are translated for use in different health system contexts, the research team should assess not only the technical and semantic equivalence of the survey questions, but also the cultural relevance of the questions included in these instruments before data collection occurs (Erkut, 2010; Mason, 2005; Temple, 2005). Without sufficient pre-data collection evaluation of the relevance and cultural equivalence of survey questions, factor analyses post-data collection could be flawed and less rigorous (Maneersriwongul and Dixon, 2004).

To address these collective issues, Flaherty and colleagues' (1988) helped refine cross-cultural adaptation methods further when they synthesized nearly two decades of multi-disciplinary research to create five levels of cross-cultural equivalence: content, semantic, technical, criterion, and conceptual. Table 1 provides the definitions for each criterion and how the RN4CAST team attempted to address each one with their approach. The strength of Flaherty et al's synthesis is its ability to capture the *emic* – constructs and concepts specific to a culture – and *etic* –concepts and constructs universally understood cross-culturally – aspects of the cultural differences between countries. Their criteria can improve an instrument's sensitivity to the place and culture where researchers may use it. The method outlined by Flaherty et al. (1988) has a limitation in that it lacks a systematic quantification method to illustrate where potential problems with previously validated instruments might arise (Mallinckrodt and Wang, 2004).

More broadly, however, there is little consensus among cross-cultural researchers in terms of the "best" method to use when translating survey instruments (Beaton et al., 2000; Guillemin et al., 1993; Reichenheim and Moraes, 2007). Researchers agree the most rigorous studies involve some combination of forward and back translation, a translation process that manages the emic and etic aspects of translation, and some kind of expert panel review (Cha et al., 2007; Im et al., 2004; Sidani et al., 2010). Pilot testing and qualitative feedback from group interviews also add additional rigor when resources allow (Harkness et al., 2003; Herdman et al., 1998; Tran, 2009). Where the literature diverges methodologically is in the timing of an expert panel review; the number of experts involved; the use of rating scales to evaluate an item's contextual relevance, a clear definition of conceptual equivalence, and how they report about the technical aspects of the translation (Hilton and Skrutkowski, 2002; Hyrkäs et al., 2003; Jones et al., 2001). Researchers also vary about when to initiate statistical analyses like principal component and confirmatory factor analyses (Johnson, 2006; Wang et al., 2006; Weeks et al., 2007). Another deterrent to pilot testing after translation is that it is often cost prohibitive. Consequently, researchers often present results without considering how translation processes may have affected their results (Erkut, 2010; Temple, 2005). In conclusion, the RN4CAST translation team found no consistent standards for a single, systematic method for translating HSR instruments that could adequately manage the challenge of simultaneously translating the languages of organizations, administrative hierarchies, and professions.

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

2. Methods

The countries and languages participating in RN4CAST include Belgium (Dutch, French, and German), England (British English), Germany (German), Finland (Finnish), Greece (Greek), Ireland (Irish-English), the Netherlands (Dutch), Norway (Norwegian), Poland (Polish), Spain (Spanish), Sweden (Swedish), and Switzerland (Swiss French, Swiss German, Swiss Italian). The overall methodology included the finalization of a core battery of instruments, forward/backward translation, and three levels of expert feedback with one group quantifying their feedback through Content Validity Indexing techniques. The comprehensive and systematic approach was used across all participating partners in the study with the aim of establishing a pre-data collection process that could improve the chances that the instrument would consistently and accurately measure the same concepts and constructs across countries during actual data collection. Parts of the methods described in this paper were pilot tested with a related instrument that required translation from English to Mandarin Chinese (Liu et al., 2011). To establish a baseline CVI based on the original instrument, raters from the US (for American English) were also included.

2.1. Instrument finalization

The instrument finalization process for RN4CAST involved multiple steps (Fig. 1), many of which are standard parts of instrument development and cross-cultural adaptation processes. Instrument development began many years ago by the Center for Health Outcomes and Policy Research, University of Pennsylvania and was used for many US-based studies focused on patient outcomes research sensitive to nursing care processes (Aiken et al., 2002, 2003). European researchers piloted tested it in several countries there (Bruyneel et al., 2009; Rafferty et al., 2007).

During two plenary meetings with expert health services researchers and statisticians present from each country, the first level of expert panel review occurred with a goal of creating an instrument that could produce comparable results across countries. The first level review by expert researchers helped to increase the likelihood that the final questions comprising the research instrument reflected the use of the strongest evidence available as a base. After extensive discussions among the researchers, the coordinating research teams organized the nurse survey instrument into four sections, some coming from established survey instruments. Section A evaluates the nursing work environment, as measured by the Practice Environment Scale from revised Nurse Work Index [PES-NWI-R] (Lake, 2002; Lake and Friese, 2006; Warshawsky and Havens, 2011), and gauges nurse burnout through the Maslach Burnout Inventory (MBI) for healthcare professionals (Maslach et al., 2001). Section A also includes other questions that measure job characteristics of nurses. Both instruments have been extensively tested and validated in the literature. Section B asks nurses questions about quality and safety of nursing services delivery in their organizations. Section C asks nurses to evaluate the nature and division of their work on the most recent shift that they worked. The final section (D) asks demographic questions, including different educational levels and the country of origin of nurses participating in the study. The latter was important to assess since nurse migration is a significant workforce issue in Europe.

2.2. Translation guide development

After finalization of the instrument, the project's translation manager (AS) reviewed all questions involved in the survey. She identified potentially problematic questions, phrases, or terms that might be difficult to translate conceptually from English to the target language. Five experienced, graduate educated, English-only speaking, US nurses with recent clinical experience recruited by the project manager reviewed the potential problem terms and phrases and provided equivalent terms or phrases through free-listing techniques. The project manager synthesized their feedback into one document. This group served as the second expert panel review in the process. The result was a uniform translation guide that each country's Forward translators used for their translation processes. It helped address the issues with interpreting the American English slang of the MBI and the professional language of American nursing. The translation guide was also a way to compensate for expected variation in translator skills and experience.

2.3. Forward–backward translation

Ideally, professionally certified translators with medical language translation experience were sought to conduct the translation, but this level of professional was not necessarily available in every country nor within the country's budget. At a minimum, translators needed at least a bachelor's degree and have five years of translation experience. A graduate degree with health translation experience was the preferred minimum qualification set. It is important to note that the German team did not participate in the forward and back translation process of the NWI-R and the MBI because those parts of the instrument had already been translated in German and used in a previous study (Aiken et al., 2001). Belgium also used the German translation so those results apply to that country. The Swiss-German translation was modified from the original German translation. In addition, Spain and Norway only participated in the relevance assessment process. Those three countries, in one way, ended up serving as a "control" measure as the team tested this process. Two other countries incorporated additional team evaluations of the translation into the process in order to provide the most accurate translation prior to expert evaluation. Adding this approach was left up to individual countries and teams incorporated it as resources allowed. Once the translations had finished, the translation project manager completed the final

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

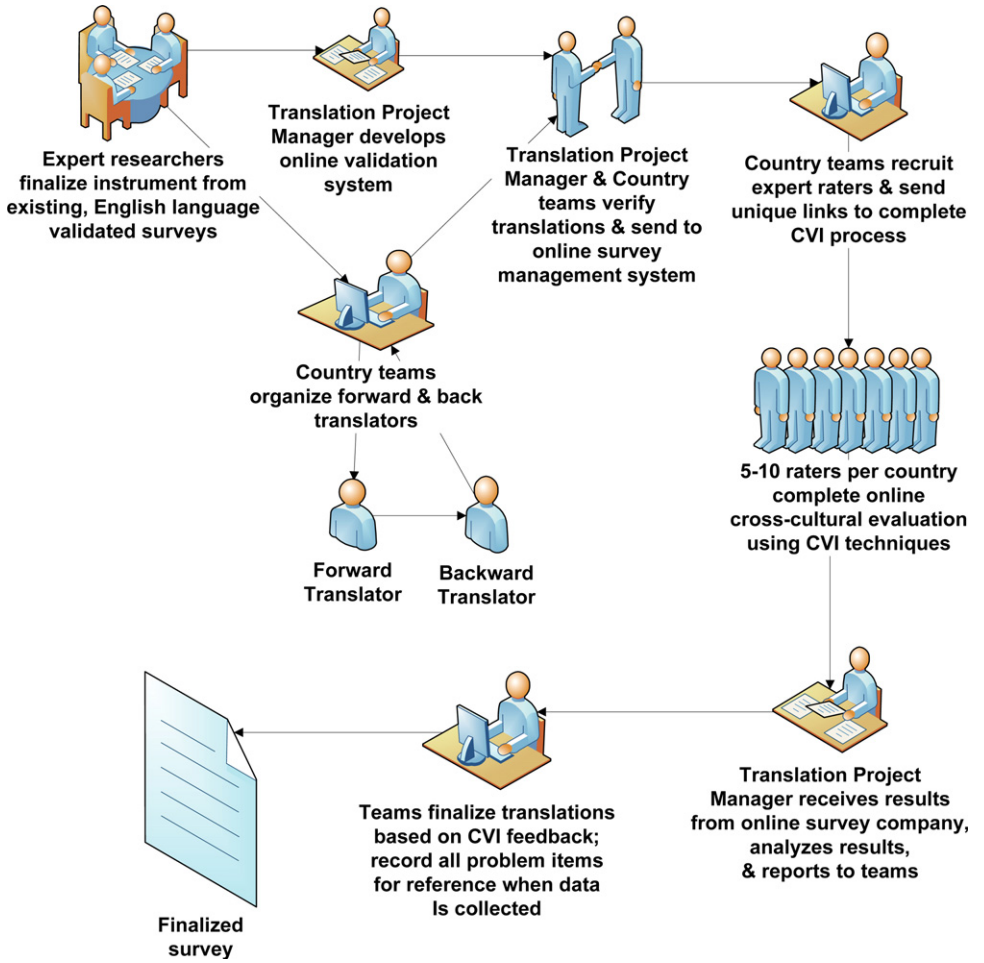

Fig. 1. Illustration of the systematic translation and evaluation process prior to data collection.

review of the translation. She made recommendations to each country's team for modifications. The team fine tuned the translation, based on the feedback, and prepared it for online evaluation by expert raters.

2.4. Expert panel review using content validity indexing

CVI is a technique traditionally used to evaluate and judge the relevance of potential survey questions for instrument creation. Commonly, it involves the scored feedback from five to ten "expert" raters who evaluate if the question is appropriate and relevant to the study's population, if the format of the question is appropriate, and offer suggestions for improvement (Polit and Beck, 2006). "Experts" are usually researchers with methodological and content expertise (Polit and Beck, 2006). A recent analysis showed that the "expert raters" method used with CVI processes consistently predicts potentially problematic survey items (Olson, 2010). The CVI scores indicate the degree of agreement between the raters (Polit and Beck, 2006). For this study, we theorized that the CVI process could provide researchers using established instruments in survey research with a way to evaluate the cross-cultural relevance and accuracy of translation of an item and overall instrument.

For the RN4CAST process, the first level of expert review by the research teams covered the methodological and content expertise, yet many researchers on the team did not have recent hospital nursing experience. Therefore, in order to capture the perspective of currently practicing hospital nurses from each country, the RN4CAST teams opted to use practicing nurses as experts because their expertise lies in current health system operations and realities of nursing practice. They would be the best judges of the relevance of question content and construct to nursing practice in their countries and clinical settings. Furthermore, the additional expert panel review would add rigor to the instrument's pre-data collection evaluation process.

Each country's team was asked to invite 10–13 raters, defined as bilingual (English and native language), practicing hospital nurses with working experience in their country's healthcare system within the last five years. Familiarity with research processes was another desirable (but not required) characteristic of an expert rater. The teams did not evaluate language proficiency of

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

Table 2
Example of the online CVI relevance evaluation tool sent to expert raters with two sample questions from the NWI-R and the Belgian Dutch translation.

Question/translation	Relevance score[a]				Translation equivalent?		Comments
	1	2	3	4	Y	N	
Adequate support services allow me to spend time with my patients.							
Adequate ondersteunende diensten stellen mij in staat tijd door te.							
Physicians recognize nurses' contributions to patient care.							
Artsen erkennen de bijdrage van verpleegkundigen aan de patiëntenzorg.							

[a] 1 = not relevant, 2 = somewhat relevant, 3 = very relevant, 4 = highly relevant.

the raters because it was assumed that self-selection based on their English-language proficiency would mediate the team's lack of resources to measure that individually with each rater.

With expert raters assembled, through an online survey evaluation system with unique links generated for each rater, the expert panels were asked to evaluate the relevance of the questions to nursing practice in their home country. The standard four-point CVI rating scale was used to evaluate the items for their content, construct, conceptual relevance (1 = not relevant, 2 = somewhat relevant, 3 = very relevant, 4 = highly relevant) (Polit and Beck, 2006). A second question was used to evaluate if items were semantically and technically equivalent through a simple "Yes" or "No" answer. The second question was created specifically for this study and is not a part of traditional CVI rating processes. Table 2 provides an example of how the relevance scoring was presented to raters. Of note, Spain opted to use only expert panel review of the semantic and technical aspects of their Spanish translation, so their raters did not participate in the evaluation of the translation. Overall, the entire process was managed online through email communications, video or audio conversations, and an online data collection tool.

2.5. Data analysis

CVI calculations occur at both the item (I-CVI) and scale (S-CVI) levels (Polit and Beck, 2006). An I-CVI score represents the number of raters scoring an item with a 3 or 4 (very or highly relevant) divided by all participating experts. The TI-CVI score, developed for this study and based on the CVI approach, is the number of raters positively indicating the translation was semantically and technically equivalent, also divided by all participating experts. For this study, the TI-CVI and I-CVI are calculated separately but using the same formulas. That approach attempts to prevent one set of scores or the other from skewing the results since it is possible to create a semantically and technically equivalent translation but still have a question that is evaluated as not relevant to the culture or context. The S-CVI and TS-CVI score are the averages of all the raters' scores for a country.

Polit et al. (2007) developed a formula that integrates an I-CVI score into a modified kappa statistic calculation that corrects for chance. The modified kappa evaluation criteria are: Fair 0.40–0.59; Good 0.60–0.73; and Excellent $\geq$0.74. That formula addresses a weakness of the CVI calculation in that it does not, on its own, correct for chance agreement among the raters. Chance agreement among the raters could decrease the reliability of the CVI results. Based on this methodological concern, for this study, we used their formula calculating a modified kappa statistic as the final evaluation criteria.

Initial discussions with the first expert panel review alerted the translation team that some items might emerge as "potentially problematic" because of how raters might evaluate the question in their home country and because of health system administrative hierarchy differences. Consequently, the team needed to set minimum standards for what would constitute a potentially problematic item (PPI). A PPI could occur at the item level, a group of them specific to a country, and also for the entire study. Early testing of the I-CVI calculations (chance corrected and not) supported Polit et al.'s (2007) finding that any non-chance corrected item level CVI score over 0.78 would be acceptable. The team, however, did find that at times non-chance corrected I-CVI scores between 0.63 and 0.77, when converted to modified kappa scores, fell into modified kappa evaluation category of "good". Therefore, first we conceptually defined a PPI as a survey question that received a chance-corrected I-CVI or TI-CVI score below 0.59 (Cicchetti and Sparrow, 1981; Fleiss, 1981; Polit et al., 2007).

For item level evaluations, a translation score below 0.73 required review and if necessary, would need correcting prior to data collection. That was then coordinated between the country's team and the translation group. If an item received a poor relevance score (0.59 or lower) from the raters, first it was compared with its translation score to determine if that was the cause. Then the item was automatically flagged for observation for atypical responses by survey participants during final survey analysis. Item level relevance scores between 0.60 and 0.73 would be marked for observation during data analysis but were less concerning since they fell into the "good" category. Then adapting the criteria to the country level, the team concluded a country would need to conduct a pilot study of the instrument if both relevance and translation scores for more than 50% of the items did not receive good ratings. Across the entire study, it was agreed

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

Table 3
CVI with chance correction results.

Country name (language)	Scale CVI[d]	Range of I-CVI scores[d]	Translation scale CVI[d]	Range of TI-CVI scores[d]
Belgium (Dutch)	0.84	0.40–1.0	0.95	0.34–1.0
Belgium (French)	0.69	0.15–1.0	0.89	0.15–1.0
Finland (Finnish)	0.79	0.25–1.0	0.91	0.34–1.0
Germany (German)[c]	0.95	0.42–1.0	0.79	0.05–0.76
Greece (Greek)	0.90	0.41–1.0	0.99	0.66–1.0
Netherlands (Dutch)	0.84	0.30–1.0	0.88	0.20–1.0
Norway (Norwegian)[b]	0.75	0.63–1.0	0.94	0.27–1.0
Poland (Polish)	0.61	0.21–1.0	0.91	0.66–1.0
Spain (Spanish)[a]	0.81	0.60–1.0	N/A	N/A
Sweden (Swedish)	0.91	0.34–1.0	0.93	0.41–1.0
Switzerland (French)	0.70	0.20–1.0	0.84	0.20–1.0
Switzerland (German)	0.81	0.26–1.0	0.89	0.20–1.0
Switzerland (Italian)	0.95	0.61–1.0	0.93	0.48–1.0
United States (American English)	0.70	0.21–1.0	N/A	N/A

[a] NWI-R scores only with no translation evaluation.
[b] Survey did not include the MBI-HSS.
[c] Used an existing translation, did not conduct a new one.
[d] All scores are the result of the CVI integrated modified kappa calculation formulas developed by Polit et al. (2007) which integrates I-CVI scores into a formula that adjusts the I-CVI relevance rating for chance agreement. The modified kappa rating scale comes from Cicchetti and Sparrow (1981) and Fleiss (1981) with 0.39 or lower as poor; 0.40–0.59 or lower as fair; 0.60–0.73 good; ≥0.74 excellent.

that if more than 50% of participating countries' raters scored the relevance of the question poorly, it would be considered a PPI for the entire survey. The analytic teams would then use those items in specific analyses post-survey that would determine if they had (1) affected survey responses and (2) had a predictive effect on subject responses. This allowed the team to maintain a single format for the survey but account for culturally sensitive response variations.

Once the raters completed their evaluations of the survey instrument, data was managed by country and also merged into a single dataset. Formulas were then programmed into an MS excel spreadsheet to conduct the calculations.

2.6. Ethical issues

The project has been granted financial support from the European Commission. Depending on national legislation, the study protocol was approved by either central ethical committees (e.g. nation or university) or local ethical committees (e.g. hospitals) (Sermeus et al., 2011).

3. Results

A total of 117 nurses out of 131 invitees (89%) served as raters for the evaluation process. With one exception (Germany with only five raters participating), each country had a minimum of 7 raters and a maximum of 11. Complete scale-level results of the CVI with chance correction evaluation processes are found in Table 3. The US S-CVI score was .70, which established the baseline criteria for comparisons with other countries. For all other countries, the initial S-CVI country average ranged from .61 (Poland) to .95 (Swiss Italian), with an overall S-CVI average of .78. TS-CVI scores ranged from .79 (Germany) to .99 (Greece). A low S-CVI score did not necessarily mean the TS-CVI score was also low. For the ranges of the scores, it is notable that Germany was the only country without a 1.0 rating for the translation. This was attributable to the effect a single rater had on the evaluation process and the fact that there were only 5 raters for that country.

Further review of the scores suggested that there were significant differences between the initial S-CVI and the translation S-CVI scores. With the country as the unit of analysis, the team performed a paired *t*-test to assess for differences between S-CVI and TS-CVI. The overall difference between the average S-CVI scores (0.81) and TS-CVI scores (0.90) was −.09 ($p = 0.0229$, 95% CI = −.16 to −.01). These discrepancies in the scoring of the items pushed the team to investigate each section of the instrument in detail to see if there were differences between sections that might have affected the overall results and explain the significant differences between the two scores.

For the entire survey, a total of 35 out of 140 items fell into the category of a PPI. When the team examined the problematic questions, usually an item with a low rating had a problem with translation that the team addressed and corrected as appropriate to the country. Upon further examination of the results, unique findings emerged in each section. First, the source of the significant differences between the scale and translation evaluations came from a single section of the instrument: The MBI-HSS. The majority of PPIs for every country came from that section of the instrument. Seven out of 22 items in the MBI received low relevance scores for 9 out of 11 countries. Most of those items had the most American English slang in them. Interestingly, however, questions about workplace violence and abuse found in that section also received low scores by the majority of raters. The MBI's ratings had the effect of reducing the overall S-CVI and TS-CVI of the instrument in each country. The TS-CVI for the MBI had lower scores across all countries, most likely due to the challenges of translating the American English slang found in many of the MBI's items. When the MBI rating scores were

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

removed from the calculations, the numbers improved for the S-CVI and TS-CVI, often increasing scale level scores for relevance and translation by a full tenth of a point.

Meanwhile, in the NWI-R, only one question received a problematic rating and that question addressed the presence of a Chief Nursing Officer in the organization, a role not present in many hospitals in some of the participating countries. Given the differences between administrative systems, hospital and nursing hierarchies across countries, this was not surprising. Patient outcome measures also proved problematic for many raters as many considered the common nursing-sensitive outcome measures like intravenous site infection, pressure ulcer prevalence, and like items as not relevant questions to ask nurses in a survey.

A striking cultural difference from US survey practices emerged in the results as most of the participating countries found the demographic questions as irrelevant for a nurse's survey, even basic ones about level of education. Some of these countries had less well-developed nursing research infrastructure, even though their medical research infrastructure might be more developed, which is one possible explanation for the scoring. Another explanation might be cultural differences related to concepts of privacy. Some groups might perceive the demographic questions as too revealing of personal information or easily trackable and would therefore have rated the questions as less relevant.

Finally, another interesting finding from this study was the scores between languages were more similar than scores within countries that had more than one official language. A country with multiple languages spoken was more likely to have differences between the languages reflected in S-CVI scores, as Table 3 illustrates with Belgium and Switzerland. Sample size limitations did not allow for a statistical examination of the differences, but they are apparent and noteworthy.

4. Discussion

Overall, the pre-data collection approach developed for the RN4CAST study appears to be able to identify where problems with the survey might occur and if the problem is specifically related to relevance or translation. Important for this study, it helped to assess if questions were relevant to nursing practice in the country before data collection.

The findings produced many salient points for discussion. To begin, the significant differences between the initial S-CVI and TS-CVI scores demonstrate the importance of conducting a holistic evaluation of the cross-cultural relevance of an instrument that includes a separate evaluation for each of the five areas described by Flaherty et al. (1988). These findings show that one could have a good translation, but the relevance of the questions might not score well in a different context and vice versa. The finding also highlights the potential problem a poor translation could produce for a research study where no pre-data collection evaluation occurs. Ideally, a single score that could be compared against the final results of the study using the instrument would be ideal but was out of the scope of this current project.

The MBI is one area where the effects of cross-cultural relevance appeared prominently. The MBI's evaluation by the raters decreased the overall S-CVI of the instrument and illustrates a potential consequence of merging several instruments into a single one, even if they have been previously validated through other methods. Normally, the CVI process helps researchers to filter survey questions during the initial instrument development stage. Therefore, when integrating established research instruments into a larger one for a cross-national study, researchers may need to closely examine scores for the established instruments and anticipate them having an effect on overall relevance scoring.

Several cases from the participating countries also had specific methodological issues that are important to highlight. Germany, to start, had significant difficulties recruiting bilingual nurses to serve as expert raters. German nursing, compared to the other countries participating in this study, is far less professionalized and "academic nursing" – which often requires knowing a second language in Europe – is a new, twenty-first century concept there. Consequently, it proved more difficult to recruit bilingual experts and that is one reason why it had only five raters. In that case, researchers may have to use either a blended group of bilingual, non-healthcare raters or two sets of raters with varying linguistic skills. That approach could potentially mediate the lack of bilingual providers, but would require further testing. Furthermore, with only five raters, when one of the raters scored nearly all of the translations poorly, it skewed the results. While some of the criticisms were accurate and resulted in modified survey items, others were fairly subjective according to the German team and did not require modifying the translation. Therefore, whether or not five raters truly is enough when translation is involved, especially if a country had at least one person who consistently scored translations lower than the rest of the group, needs further exploration. It is important to note that when at least seven raters were involved, the effect of a single rater's evaluations diminished.

Poland also presents an interesting case because the S-CVI scores between the initial and the translation ones had the largest gap of any country. The borderline relevance rating of the initial S-CVI for Poland of .61 caused concern for the Translation Project Manager and the Polish team, who did decide to undertake a pilot study before proceeding with their larger data collection processes. The pilot study proved a useful exercise that resulted in additional instrument modifications that enhanced the overall rigor of the Polish version of the instrument prior to implementation in the RN4CAST study. Because this method provided a quantifiable measure of the relevance of the instrument, the team was able to efficiently allocate resources to a pilot study for that country because the preliminary evaluation appeared to require it. In an era of increasingly constrained resources, the example of the Polish team

Please cite this article in press as: Squires, A., et al., A systematic survey instrument translation process for multi-country, comparative health workforce studies. Int. J. Nurs. Stud. (2012), doi:10.1016/j.ijnurstu.2012.02.015

and the overall study results may help researchers decide if and when to initiate a pilot test of a translated, existing instrument prior to starting the larger study.

Another striking aspect of this study's was how even within the same country, when multiple languages were present, scores varied significantly. As Table 3 illustrates, Switzerland is the prime example for this study because its scores varied by as much as a tenth of a point between the three official languages in the country. In contrast, scores between Swiss French and Belgian French were more similar than those of their respective countrymen. Belgian Dutch and Netherlands Dutch received similar scores. Yet even though some languages produced similar scores, further qualitative examination of response patterns among the raters in each language showed distinct scoring patterns even within the same language. The scoring patterns were unique to where the language was spoken and reflected health system differences between the countries. These findings should emphasize the importance of evaluating an instrument's in the context of each country, even if one already exists in the language.

Finally, even though the baseline scale-level US relevance score was "good" based on the kappa evaluation criteria, the results suggests that over time, the relevance of some survey items may diminish as health system improvements occur, especially around nurse's work environments. Consequently, many of the questions in the survey, like those found in the MBI, might not have been perceived as relevant to those raters especially if they worked in hospitals with highly supportive work environments for nurses. A question about burnout, therefore, might not be perceived as relevant because it would not be prevalent in the nurses' workplace. That case illustrates the importance of rater identity and the potential consequences of health system evolution.

Several other findings cause us to recommend several things when selecting raters. First, we recommend that researchers have at least seven raters when instrument translation is involved in their study in order to compensate for how a single rater can skew translation results. We also recommend, as do Polit et al. (2007), having well defined criteria for selecting the expert raters. Those provided by Grant and Davis (1997) are useful for shaping selection criteria. When translation occurs and if resources allow, having a way to gauge language proficiency of the raters would also be useful.

Finally, as with any study, this one had several limitations. First, for the evaluation process, an inconsistent number of raters between countries may have affected the results along with how countries implemented the overall process. Furthermore, inasmuch the CVI method is normally used during the instrument development process, this approach may still have some methodological issues to resolve since it was developed specifically for this study. Finally, since the teams were unable to test language proficiency among the practicing nurse expert raters and assumed self-selection would automatically eliminate those with lower language proficiency, there could have been significant variability in language levels between the raters from a country that would have affected the evaluation processes.

5. Conclusion

The instrument verification method based on the CVI-with-chance-correction grounded approach provided valuable insights into the overall relevance of the instrument in different contexts and a quantifiable measurement of those features. It resolved several methodological issues that emerged in the literature related to quantifying the different aspects of translation and improving the speed at which translation validation methods could occur. For the RN4CAST research team, it offered a way to standardize a complex process across many countries and cultures. It also helped to identify and anticipate where problematic or outlying results might emerge during data analysis.

A key lesson from this study is that health system administrative hierarchies and even the language of professions require careful translation in order to improve the likelihood that the language used to describe them is conceptually equivalent across countries. Those languages present specific translation challenges that are different from language for behavior and symptoms that more easily translate. Another lesson from this study, which will be important for replicating this method in low resource settings, is how the quality of professional research infrastructure may affect responses. The relevance score differences between countries on specific parts of the survey illustrate how questions that may seem important and relevant to expert researchers may not contain the same relevance to healthcare workers if no culture of research or quality improvement exists within the country.

Other health services researchers who seek to design methodologically rigorous multi-country studies can benefit from the methods and approaches described in this article. As HSR and related research expands and grows in many countries with varying levels of health system resources, ensuring that instruments designed in one setting translate well into another and account for health system differences will be an important step in any comparative country study.

Acknowledgements

The authors would like to thank all members of RN4CAST country teams for their contributions to the study.

Conflict of interest statement

None declared.

Funding

This research is funded by the European Union's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 223468 (W. Sermeus, PI) and the National Institute of Nursing Research, National Institutes of Health (P30NR05043 L. Aiken, PI). The Norwegian Nurses Association funded the Norwegian part of this study. The Swedish Association of Health Professionals,

Committee for Health and Caring Sciences (CfV) and Strategic Research Program in Care Sciences (SFO-V) at Karolinska Institutet provided additional funding for the Swedish study.

Ethical approval

KU Leuven and IRB per country involved in the study.

References

Aiken, L.H., Clarke, S.P., Sloane, D.M., Sochalski, J.A., Busse, R., Clarke, H., Giovannetti, P., Hunt, J., Rafferty, A.M., Shamian, J., 2001. Nurses' reports on hospital care in five countries. Health Affairs 20, 43–53.

Aiken, L.H., Clarke, S.P., Sloane, D.M., Sochalski, J., Silber, J.H., 2003. Hospital nurse staffing and patient mortality, nurse burnout and job dissatisfaction. Journal of the American Medical Association 288 (16), 1987–1993.

Aiken, L.H., Clarke, S.P., Sloane, D.M., 2002. Hospital staffing, organizational support, and quality of care: cross-national findings. International Journal for Quality in Health Care 14 (1), 5–13.

Beaton, D.E., Bombardier, C., Guillemin, F., Ferraz, M.B., 2000. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 25 (24), 3186–3191.

Brislin, R.W., 1970. Back-translation for cross-cultural research. Journal of Cross Cultural Psychology 1 (3), 187–196.

Bruyneel, L., Van den Heede, K., Diya, L., Sermeus, W., Aiken, L.H., 2009. Predictive validity of the international hospital outcomes study questionnaire battery: an RN4CAST pilot study. Journal of Nursing Scholarship 41 (2), 202–210.

Cha, E., Kim, K.H., Erlen, J.A., 2007. Translation of scales in cross-cultural research: issues and techniques. Journal of Advanced Nursing 58 (4), 386–395.

Choi, B., Bjorner, J.B., Ostergre, P., Clays, E., Houtman, I., Punnett, L., Rosengren, A., De Bacquer, D., Ferrario, M., Bilau, M., Karasek, R., 2009. Cross-language differential item functioning of the job content questionnaire among European countries: the JACE study. International Journal of Behavioral Medicine 16, 136–147.

Cicchetti, D.V., Sparrow, S., 1981. Developing criteria for establishing interrater reliability of specific items: application to assessment of adaptive behavior. American Journal of Mental Deficiency 86, 127–137.

Erkut, S., 2010. Developing multiple language versions of instruments for intercultural research. Child Development Perspectives 4 (1), 19–24.

Flaherty, J.A., Gaviria, F.M., Pathak, D., Mitchell, T., Wintrob, R., Richman, J.A., Birz, S., 1988. Developing instruments for cross-cultural psychiatric research. Journal of Nervous & Mental Disease 176 (5), 257–263.

Fleiss, J., 1981. Statistical Methods for Rates and Proportions, 2nd ed. John Wiley, New York.

Gibson, C.B., Zellmer-Bruhn, M.E., Schwab, D.P., 2003. Team effectiveness in multinational organizations: evaluation across contexts. Group & Organization Management 28 (4.), 444–474.

Gjersing, L., Caplehorn, J.R., Clausen, T., 2010. Cross-cultural adaptation of research instruments: language, setting, time and statistical considerations. BMC Medical Research Methodology 10, 13.

Grant, J.S., Davis, L.L., 1997. Selection and use of content experts for instrument development. Research in Nursing and Health 20, 269–274.

Guillemin, F., Bombardier, C., Beaton, D., 1993. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. Journal of Clinical Epidemiology 46 (12), 1417–1432.

Harkness, J.A., Van de Vijver, F.J.R., Mohler, P.P., 2003. Cross-cultural Survey Methods. J.C. Wiley & Sons, Hoboken, NJ.

Herdman, M., Fox-Rushby, J., Badia, X., 1997. 'Equivalence' and the translation and adaptation of health-related quality of life questionnaires. Quality of Life Research 6 (3), 237–247.

Herdman, M., Fox-Rushby, J., Badia, X., 1998. A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. Quality of Life Research 7, 323–335.

Hilton, A., Skrutkowski, M., 2002. Translating instruments into other languages: development and testing processes. Cancer Nursing 25 (1), 1–7.

Hyrkäs, K., Appelqvist-Schmidlechner, K., Oksa, L., 2003. Validating an instrument for clinical supervision using an expert panel. International Journal of Nursing Studies 40, 619–625.

Im, E., Page, R., Lin, L., Tsai, H., Cheng, C., 2004. Rigor in cross-cultural nursing research. International Journal of Nursing Studies 41, 891–899.

Johnson, T.P., 2006. Methods and frameworks for crosscultural measurement. Medical Care 44 (11, S3), S17–S20.

Jones, P.S., Lee, J.W., Phillips, L.R., Zhang, X.E., Jaceido, K.B., 2001. An adaptation of Brislin's translation model for cross-cultural research. Nursing Research 505, 300–304.

Lake, E.T., 2002. Development of the practice environment scale of the Nursing Work Index. Research in Nursing & Health 25, 176–188.

Lake, E.T., Friese, C.R., 2006. Variations in nursing practice environments: relation to staffing and hospital characteristics. Nursing Research 55, 1–9.

Mallinckrodt, B., Wang, C., 2004. Quantitative methods for verifying semantic equivalence of translated research instruments: a Chinese version of the experiences in close relationships scale. Journal of Counseling Psychology 51 (3), 368–379.

Maneersriwongul, W., Dixon, J.K., 2004. Instrument translation process: a methods review. Journal of Advanced Nursing 48 (2), 175–186.

Maslach, C., Schaufeli, W.B., Leiter, M.P., 2001. Job burnout. Annual Review of Psychology 52, 397–422.

Mason, T.C., 2005. Cross-cultural instrument translation: assessment, translation, and statistical applications. American Annals of the Deaf 150 (1), 67–72.

Liu, K., Squires, A., Ming, Y.L., 2011. A pilot study of a systematic method for translating patient satisfaction questionnaires. Journal of Advanced Nursing 67 (5), 1012–1021.

Olson, K., 2010. An examination of questionnaire evaluation by expert reviewers. Field Methods 22 (4), 295–318.

Peña, E.D., 2007. Lost in translation: methodological considerations in cross-cultural research. Child Development 78 (4), 1255–1264.

Perneger, T.V., Leplege, A., Etter, J.F., 1999. Cross-cultural adaptation of a psychometric instrument: two methods compared. Journal of Clinical Epidemiology 52 (11), 1037–1046.

Polit, D.F., Beck, C.T., 2006. The content validity index: are you sure you know what's being reported? Critique and recommendations. Research in Nursing and Health 29, 489–497.

Polit, D.F., Beck, C.T., Owen, S.V., 2007. Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. Research in Nursing & Health 30, 459–467.

Rafferty, A.M., Clarke, S.P., Coles, J., Ball, J., James, P., McKee, M., et al., 2007. Outcomes of variation in hospital nurse staffing in English hospitals: cross-sectional analysis of survey data and discharge records. International Journal of Nursing Studies 44 (2), 175–182.

Ramirez, M., Teresi, J.A., Holmes, D., Gurland, B., Lantigua, R., 2006. Differential item functioning and the mini-mental state examination: overview, sample, and issues of translation. Medical Care 44 (11, S3) S-95-S106.

Reichenheim, M.E., Moraes, C.L., 2007. Operationalizing the cross-cultural adaptation of epidemiological measurement instruments. Revista de Saude Publica 41 (4), 665–673.

Sermeus, W., Aiken, L.H., Van den Heede, K., Rafferty, A.M., Griffiths, P., Moreno-Casbas, M.T., Busse, R., Lindqvist, R., Scott, A.P., Bruyneel, L., Brzostek, T., Kinnunen, J., Schubert, M., Schoonhoven, L., Zikos, D., 2011. Nurse forecasting in Europe (RN4CAST): rationale, design, and methodology. BMC Nursing 10 (6), doi:10.1186/1472-6955-10-6.

Sidani, S., Guruge, S., Miranda, J., Ford-Gilboe, M., Varcoe, C., 2010. Cultural adaptation and translation measures: an integrated method. Research in Nursing & Health 33, 133–143.

Temple, B., 2005. Nice and tidy: translation and representation. Sociological Research 10 (2) online, http://www.socresonline.org.uk/10/2/temple.html.

Thrasher, J.F., Quah, A.C.K., Dominick, G., Borland, R., Driezen, P., Awang, R., Omar, M., Hosking, W., Sirirassamee, B., Boado, M., 2011. Using cognitive interviewing and behavioral coding to determine measurement equivalence across linguistic and cultural groups: an example from the international tobacco control policy evaluation project. Field Methods, doi:10.1177/1525822X1141876 first published online August 25, http://fmx.sagepub.com/content/early/2011/05/18/1525822X11418176.full.pdf+html.

Tran, T.V., 2009. Developing Cross-cultural Measurement. Oxford University Press, London.

Van de Vijver, F.J.R., Leung, K., 1997. Methods and Data Analysis for Cross-cultural Research. Sage, London.

Wang, W., Lee, H., Fetzer, S.J., 2006. Challenges and strategies of instrument translation. Western Journal of Nursing Research 28 (3), 310–321.

Warshawsky, N.E., Havens, D.S., 2011. Global use of the practice environment scale of the nursing work index. Nursing Research 60 (1), 17–31.

Weeks, A., Swerissen, H., Belfrage, J., 2007. Issues, challenges, and solutions in translating study instruments. Evaluation Review 31 (2), 153–165.