Translate Wisely! An Evaluation of Close and Adaptive Translation Procedures in an Experiment Involving Questionnaire Translation

Repke, Lydia; Dorer, Brita

This is an accepted manuscript of an article published by Taylor & Francis in
International Journal of Sociology, Vol. 51, 2021, No. 2, pp. 135-162
available online at: https://doi.org/10.1080/00207659.2020.1856541

Translate Wisely! An Evaluation of Close and Adaptive Translation Procedures in an Experiment Involving Questionnaire Translation

Lydia Repke and Brita Dorer

Department of Survey Design and Methodology, GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany

ABSTRACT

To challenge the commonly made assumption in cross-national survey projects that close translation yields more comparable data than adaptation, we implemented a translation experiment in the CROss-National Online Survey Panel. The English source questionnaire was split into three batches of 20 items each and was translated by three translation teams into Estonian and three teams into Slovene. The teams received specific instructions on how to translate each batch (either closely or adaptively) so that, by design, the teams translated two batches following one approach and one following the other approach. Respondents in the two countries (Estonia and Slovenia) were randomly assigned to three distinct questionnaire versions based on the same source questionnaire, each consisting of translations by all three teams and including close and adaptive translations. We developed an analytical framework to assess the *translation potential* of the source items (i.e., all theoretically possible translations of a specific item) and the actual *translation scores* (i.e., the degree of closeness vs. adaptiveness of a specific translation). We show that some items are more sensitive to the wording (small linguistic changes result in a different response behavior) while others are more robust (the meaning of the concept is retained despite linguistic changes).

KEYWORDS

questionnaire translation; adaptation; ask-the-same-question; closeness; cross-country comparison

Whenever social researchers relying on survey data are interested in making cross-national comparisons, they implicitly assume that the questionnaire versions people respond to in each country are equivalent. This entails that the source questionnaire is translated in such a way that all respondents understand the questions in exactly the same way and that, as a consequence, their answers are comparable despite the survey language. What may sound like a simple task at first glance—i.e., translating an item “correctly”—often leads to discussions in the translation process of large-scale survey projects such as the European Social Survey (ESS) or the International Social Survey Programme (ISSP). Especially so, when an item cannot be translated straightforwardly but instead might need to be adapted to account for cultural or linguistic differences.

The question of how closely an item should be translated or what degree of adaptation is acceptable, or even necessary, to ensure that a specific item is functionally

CONTACT Lydia Repke ✉ Lydia.repke@gesis.org ✉ GESIS – Leibniz Institue for the Social Sciences, P. O. Box 12 21 55, 68072 Mannheim, Germany.

equivalent across countries is not easily answered. Although statistical methods (e.g., multigroup confirmatory factor analysis, item response theory) exist to test whether the assumption of measurement equivalence holds for already collected data, an indispensable intermediate step in designing cross-national surveys is to create good translations before the actual data collection phase. So how can we give translation guidelines that apply to 20, 35, or even more languages bearing in mind that many translation mechanisms are language-pair specific?

For a long time, it has been general practice in cross-national surveys to implement close translations, as they are assumed to lead to comparable data (Harkness 2003). While this may be true for specific items, an adaptive translation approach allowing for “free” translation might be better suited for items that otherwise would suffer from possibly inaccurate or, even worse, wrong translations due to the limited potential to buffer cultural and linguistic differences. In this paper, we challenge the assumption that close translation yields more comparable data than adaptive translation. We do so by analyzing the translation experiment implemented in the CROss-National Online Survey (CRONOS) Panel Wave 5 and, with this, intend to find answers to our research questions: Do the two translation approaches lead to distinct translations? Do these distinct translations lead to different responses in the target population? How can these different answers be explained? Is one of the two approaches better than the other?

In this paper, we present an innovative method to tackle our research questions exploratively. In particular, we assume that each survey item has a target-language specific *translation potential*. The idea is that some survey instruments can only be translated closely, some others only adaptively, and again others both ways. All translations that are hypothetically possible comprise the translation potential of a specific item. In addition to these theoretical considerations, each translated item can be placed somewhere on a closeness-adaptiveness continuum when compared to the source version. So every translation holds a specific (i.e., actually realized) *translation score*. Building our analytical framework on these two concepts (i.e., the theoretical translation potential of the source item in the target language and the empirical translation score of the target item), we are able to show that giving translators instructions on how to translate that do not fit the translation potential (e.g., telling them to translate closely although a close translation makes no sense in the target language) may lead to bad or perhaps even incorrect translations.

In the remainder of this paper, we first review the literature on close and adaptive translation, which is the basis for our translation experiment. Second, we further elaborate on our study by explaining its design, the translation procedure used, and the implementation of the experiment in the CRONOS Panel. Third, we introduce our newly developed method to analyze the data and, fourth, present some general and some item-specific results. Finally, we conclude with a discussion on how this method can help improve translation processes in surveys.

The spectrum of translations: purpose and text type

Questionnaire translation is located at the interface between survey research and translation studies. While interdisciplinary research offers many opportunities, it also brings some challenges in terms of which of the disciplines involved should prevail and how

researchers should deal with different approaches to the same problems. One aspect of translation where both fields of research—survey research and translation studies—have not always been following the same principles is the question of how “close” or “free” a good questionnaire translation should be. Interestingly, not even the understanding of what a “close” translation means is the same in both fields.

In translation studies, the debate about the acceptable degree of closeness or adaptiveness is relevant to any translation. The recommended level of freedom depends on a translation’s purpose and the particular text type. Concerning the *purpose*, one has to differentiate between translations for documentary and publication purposes, or as Nord (1997, 2005) puts it, between documentary and instrumental purposes. For documentary translations—whose purpose is solely to inform the reader of the content of the source text, as he/she may not be fluent enough in the source language to understand the full meaning of the source text—, the accepted level of closeness is relatively high. Here, it is less important whether the translation is expressed in the right style or whether the communication culture of the target language is met (on the importance of communication culture for translation, see, for instance, Reiss 1981). As soon as a translated text needs to be published and needs to function as a standalone text, these two aspects become more important: The readers of the target language need to be able to understand and capture the full meaning of the text at a comparable level of ease as the readers of the source text, without being disturbed by a communication culture that would be unusual in the target language. For achieving a better style and the right communication culture in the target language, the level of free or adaptive translation will be higher than if the translation merely needs to enable the reader to understand “what is in the source text.”

The second important quality of a translation that determines the accepted level of adaptation is *text type*. There are text types for which it is of utmost importance to be close to the source text as deviations from the source text would create significant harm. Examples are legal texts such as laws, technical instructions or manuals, or medication package inserts. These text types must be translated very closely to the source text to ensure that no information is altered or lost. At the same time, the style of the translation is of less importance. In contrast, creative texts, such as poems or public relations (PR) texts, allow more freedom in the translation.

Within this spectrum of text types, *questionnaires* are a very specific form of text, and its particularities must be taken into account in the translation process. In particular, questionnaires for cross-cultural survey projects lie between the two poles discussed above. On the one hand, a questionnaire must work in the target language and culture and, therefore, not be recognized as a translation, but rather be easily understood as a piece of natural language (Harkness, Pennell, and Schoua-Glusberg 2004). On the other hand, questionnaires are measurement instruments that need to function correctly. That means that they need to measure accurately and thus reflect what the source questionnaire expresses, ideally, in a comparable way across all the language versions involved. This concerns especially the answer scales, which in many cases, have to follow measurement purposes based on specific approaches to questionnaire design and not the purely idiomatic and natural use of the target language. In support of this idea, Villar (2009) pointed out how important it is for comparative surveys to translate answer

scales instead of choosing other, more convenient scales that are more common in the target survey culture, as this would modify the measurement instrument itself (on the importance of the text type for translation, see Dorer 2020; Reiss and Vermeer 1984).

What does "close" mean?

Since early ideas of translation (such as in Roman times), the accepted level of freedom versus closeness has been discussed (Chesterman 2016; Munday 2016), and this discussion is known to be a "two thousand year-old chain of theory revolving around the faithful vs. free axis" (Gentzler 2001). Both translation studies and survey research recommend translating as close as possible to the source text, but do they really mean the same thing? In *survey research*, close translation is often understood in the sense of literal or word-by-word translation. Survey translations are indeed often very close to the source text, including its syntax and word order, without considering the meaning this creates in the target language. Harkness and her colleagues (2010), in contrast, describe close translation as maintaining proximity to the source text (also in terms of its structure) while trying to meet the target language's needs for correct syntax and vocabulary. However, it can be observed that close translation often does not meet the pragmatic needs of the target language (Harkness, Villar, et al. 2010).

A guiding principle in practical translation and *translation studies* is the translation of sense and meaning rather than individual words (see, e.g., Munday 2016). In translation studies, this is often referred to as "equivalent" translation. However, the term equivalence is not easy to define, neither in survey research nor in translation studies. Nida (1964) introduced the notion of equivalence to translation studies and differentiated between formal and dynamic equivalence. *Formal equivalence* refers to keeping the same content as in the source text while also keeping as many formal elements as possible. In contrast, *dynamic equivalence* (later also called functional equivalence or referred to as communicative translation by Newmark 1981) puts a stronger focus on the source text's message. This message should be identical to the one expressed in the translation (Munday 2016). In fact, large-scale cross-national survey projects, such as the ESS, currently strive for functional (i.e., dynamic) equivalence in their questionnaire translation (European Social Survey 2018).

Another differentiation that contributes to understanding the close-free/adaptive translation continuum is the role that semantics and pragmatics play in the translation: *Semantics* deals with the meaning of words or expressions, whereas *pragmatics* deals with the message transported by these words or expressions (Chesterman 2016). Reiss and Vermeer (1984) explain this difference with an illustrative example. In German, a sign prohibiting the crossing of rails would read, "Überschreiten der Gleise verboten!" A semantically correct translation of this phrase would be, "It is forbidden to cross the rails!" However, this is not how English native speakers would express this message. The pragmatically correct translation would rather be, "Don't cross the rails!" (Reiss and Vermeer 1984).

Applying this distinction to questionnaire translation means that any translation that focuses on the source text's semantic meaning would be semantically correct. However, as the example shows, the pragmatic meaning of the source text (i.e., its intended message) may not be correctly conveyed in the translation. For a questionnaire, this can be

detrimental since there is no guarantee that the source's stimulus is the same in the target language. A thorough understanding of both the source and target language and a good understanding of both languages' communication habits (i.e., how native speakers in both languages understand the words) is required for this type of translation (Reiss 1981). It is therefore not uncommon that people with less experience or knowledge of translation mechanisms do not trust free translations that strive more for pragmatic rather than semantic meaning, thus limiting the use of such adaptations in the translation process.

Translation strategies in practice

Translating semantics and pragmatics, adaptation, and free translation are techniques that translators learn in their training. Generally, one can observe that translators often become more self-confident and sovereign in applying these techniques, the more experienced they become. In contrast, inexperienced or untrained translators tend to translate literally because they might feel safer or more comfortable to stay close to the source text. A good level of translation training and experience is required to know how to stay close to the source text's meaning despite applying a certain degree of freedom (Arffman 2012). Experienced translators usually recognize easily that, in some cases, a direct translation is not sufficient to transfer the meaning from the source to the target context. Consequently, they would either use a certain degree of freedom or make some adaptations as a strategy to improve comprehensibility in the target language (Bastin 2011; adaptation as a translation strategy is also explained by Vinay and Darbelnet 1995).

Although there is no uniform definition of what is meant by *adaptation* in translation studies, there is consensus that it describes "a set of translative interventions which result in a text that is not generally accepted as a translation but is nevertheless recognized as representing a source text" (Bastin 2011, p. 3), deliberately deviating from a literal rendering of the source text. In the field of survey research, Harkness and her colleagues (2010) define the aim of adapting survey instruments as tailoring the questions to the needs of the respondents while maintaining the measurement characteristics of the source question. Hence, adaptation is necessary in all cases where a mere translation is not sufficient to reproduce the source text to be understood comparably in a target language and culture.

Experienced translators, on the one hand, have learned in their training and their professional lives that free or adaptive translation is often required to make the meaning of a source text understandable in the target language and culture (Harkness, Pennell, and Schoua-Glusberg 2004). Survey researchers, on the other hand, often seem to fear that comparability between translated questionnaire versions is impaired by allowing translations that are perceived as too flexible and could thus jeopardize measurement properties (on a general tendency to translate closely in cross-cultural survey translation, see Kleiner, Pan, and Bouic 2009). Examples are the ESS Round 1 Translation Guidelines and Harkness' (2003) criticism of the way the "ask the same question" (ASQ) approach was implemented. In her eyes, its implementation focused too much on semantically correct and formally equivalent translations and not enough on

pragmatically correct and functionally equivalent translations (see the example of the pets in the following).

In 2003, Harkness described the requirements for cross-cultural survey projects as the quest for close translations using the ASQ approach, aiming to keep the source text's measurement properties the same in all translations. However, the line taken in these surveys was at odds with what she felt was not functionally equivalent. She pointed to an example where the English source text referred to domestic cats and where the translations were expected to refer to "the same entities" (i.e., domestic cats or felines), although dogs or parrots may have been functional equivalents in some target countries (Harkness 2003). While this particular translation approach strives for asking the same question by expressing the same stimulus, it is questionable whether it would fulfill this task in this example. Using the semantically correct translation of domestic cats in a country or culture where other pets, such as dogs or parrots, are more common carries the risk that the stimulus is not the same. If the goal of a question (its stimulus) is to ask about pets, it does not matter which pets are mentioned. Instead, it is the idea of pets that needs to come across. So respondents may get confused by entities that do not have the same function in their culture. In this example, a pragmatically correct translation would more likely express the same stimulus than a translation that sticks to the word's semantic meaning. In this sense, a close translation can sometimes neglect a target language's cultural and structural needs. If cats are not known as pets in a specific country, it would be wrong (culturally and linguistically) to name them in this function because they are not part of the meaning of the word pets in that language.

To overcome this limitation, Harkness (2003) recommended combining close translation and adaptation to achieve translations that are easier understood and functionally better suited for the target population. Nonetheless, although several surveys have adopted a more open attitude toward adaptation and free, pragmatic translation, the general policy of other surveys to date has been to stick close to the source, sometimes even at the level of words. Again, the ESS Round 1 serves as an example. Its translation guidelines included the following (Harkness 2002): "Since the ESS follows an ASQ model, functionally equivalent (but different) components are not envisaged. Translators thus need to be clear, for example, about how close or free translation is required to be as well as how to report back on unavoidable divergences."[1]

Similarly, Arffman (2012) claimed that the instructions given to translators played a major role in the relatively high level of unwanted literal translations detected in the survey instruments used in International Achievement Studies, conducted mainly by the Organization for Economic Cooperation and Development (OECD). Although her observation applied to pedagogical assessment studies from almost a decade ago, the general criticism of certain studies focusing on literal translation (i.e., closeness at the word level) is still up to date. She further argued that literal translation was the default translation strategy, which made it difficult to avoid. Kleiner and his colleagues (2009) also saw the tendency that survey researchers in cross-cultural contexts placed more focus on formal equivalence (in Nida's terms) and seemed to aim for standardization and equivalence of stimulus within the translation of source instruments.[2]

Perhaps, this confusion about semantic equivalence being functional equivalence is due to different understandings of the word close in the disciplines. In survey research,

close translation has often been understood as semantic closeness (e.g., ESS Round 1). From a translation studies perspective, however, functional or dynamic equivalence (i.e., the correct conveyance of the pragmatic meaning or the message of a text) is just as important for staying close to the source and would not be neglected for the sake of semantic meaning (Harkness, Pennell, and Schoua-Glusberg 2004). That is to say that, in translation studies, a certain level of adaptation is part of a well-trained and experienced translator's tool kit (Arffman 2012).

Translating closely versus adaptively

At the crossroad between survey translation and linguistics, Harkness and her colleagues (2004) saw a danger in too close questionnaire translations, focusing on the semantic rather than the source text's pragmatic meaning. They feared that this could cause: (a) a different perception of source text elements by respondents compared to the original intention of the source text; (b) increased respondent burden by stilted and more difficult to understand questionnaire formulation; and (c) unnatural wording. Hence, they recommended interpreting closeness in terms of pragmatic language usage (i.e., translating meaning in context). This way also the fallacy of believing that semantic equivalence would guarantee functional equivalence could be avoided (Harkness 2007).

Even now, we can observe skepticism about adaptation in major cross-national survey projects; yet, adaptations become increasingly more common in the questionnaire translation landscape. They mostly pertain to three major areas (Behr and Shishido 2016): (1) *culture* (e.g., due to factual or normative differences between source and target systems, such as political systems involving a US-American President or a British Prime Minister); (2) *measurement* (e.g., because of differences in familiarity in measurement approaches or measurement formats, such as making answer categories less strong for Japanese contexts because the Japanese public would not easily select an extreme answer category such as "strongly disagree"); and (3) *language* (e.g., to guarantee easy understanding from an idiomatic point of view, linguistic correctness, or correct pragmatic understanding, such as adding both the masculine and feminine gender in languages where this is grammatically required in contrast to languages, such as English, that only use one gender).

One reason for this general skepticism may be related to the fact that a more sovereign attitude toward free or adaptive translation generally comes with increased translator training and experience, still lacking in many questionnaire translation teams. Another reason may be that it is challenging to decide on a uniform translation approach that meets all the different mechanisms, peculiarities, and needs of the various linguistic and cultural systems present in large-scale cross-cultural survey projects.

In summary, in translation studies, deviating from purely literal translation and using adaptations where necessary is a standard method for achieving adequate and functionally equivalent (i.e., "close" or comparable) translations. In contrast, survey research tends to follow the source text as far as possible (also in lexical and structural aspects) since this strategy is assumed to ensure measurement equivalence (Harkness, Pennell, and Schoua-Glusberg 2004). The current study examines the complex mechanisms of adaptations in the field of questionnaire translation to better understand what level of

Table 1. Experimental condition for each translation team and set of items.

	Team 1	Team 2	Team 3
Items 1-20	Close	Adaptation	Adaptation
Items 21-40	Adaptation	Close	Adaptation
Items 41-60	Close	Adaptation	Close

Note. Close = translate as close as sensible. Adaptation = adapt as necessary.

free translation or adaptation should be encouraged or avoided in cross-national survey projects and when which translation strategy might be more appropriate.[3]

Methods

The translation experiment

Experimental design

The translation experiment's goal was to test whether closely translated survey items—compared to adaptively translated items—always result in better questionnaire versions in the target language, that is, questionnaire versions that are highly comparable to the source questionnaire. We decided on a research design that was as close as possible to the typical translation process setting in large-scale survey projects such as the ESS. These surveys usually employ a so-called team or committee approach, which includes the TRA steps of the TRAPD (translation, review, adjudication, pretesting, and documentation) translation process (Dorer 2020; Harkness 2003). To mimic real-life situations in the survey business, we relied on translation teams translating the items (instead of reyling on systematically modified translations with varying degrees of closeness manipulated step by step).

Estonian and Slovene translation teams had to translate 60 survey items from English into their respective languages. For this purpose, we divided the 60 source items into three batches of 20 items each. The translation teams received one batch at a time, along with instructions on how these items should be translated (i.e., either closely or adaptively). The teams had a break of at least one week before receiving the next batch and its instructions.

To minimize team effects (e.g., one team always translating the same way irrespective of the instructions or a team having a particular translation style), we chose a 2x3x3 factorial research design (Crump et al. 2018). That is two translation strategies (i.e., free translation and adaptation), three batches (i.e., each consisting of 20 items), and three translation teams (i.e., teams 1, 2, and 3) per country (see Table 1). This design allowed all teams to use each translation approach at least once while ensuring that every item was translated according to both approaches.

Item selection and translation procedure

When selecting the items for the experiment, we aimed to obtain variance in the three translation versions per language. For this reason, we asked professional Estonian and Slovene speaking survey researchers who had been involved in cross-national survey projects to earmark challenging items in their respective surveys based on their experience in the questionnaire translation process. The items finally selected had all been

proven problematic in the translation process, either with regard to general translation difficulty or specifically with respect to the degree of closeness.

As mentioned above, we chose the team approach for our translations (Harkness 2003). In total, we had six translation teams, three for English-Estonian translations and three for English-Slovene translations. Each team consisted of three people, all of whom (a) were native speakers of the target language living in the respective country (e.g., native Estonian speaker living in Estonia) and (b) had an excellent command of English. Further, two of the three members were translators. The third member was a reviewer-cum-adjudicator, who was an experienced social scientist and a survey researcher with a background in designing and fielding similar questionnaires in the target country.[4]

With each item batch, the translation teams received the English source items in an Excel translation template and written instructions on how to translate these items (with the special note that the translations had to be “fieldable”). Due to lack of resources, there was no personal training, neither in person nor web-based. The translation instructions consisted of five pages briefing the translators on the team approach (identical across all instructions) and explaining the translation strategy they were supposed to apply (varied at least once across the three batches). We never mentioned the names of two competing translation approaches (i.e., close translation and adaptation) explicitly but instead described them so that the teams were not aware they were participating in an experiment. The team members believed they were doing a regular translation job.[5]

In line with the team approach, after two parallel translations, there was a review session (R in the TRAPD process) in which both translators and the reviewer-cum-adjudicator discussed all items one by one. They then jointly agreed on one final translation. To check that each translation team was using the assigned translation method, we conducted follow-up interviews based on a fixed protocol with the six reviewers and some of the translators. We were particularly interested in whether (a) they had recognized the difference between the methods in the instructions for each batch and whether (b) they had applied the assigned method.

Implementation of the experiment

Based on the final translations, we prepared three questionnaire versions. Each questionnaire version included adaptively and closely translated items from all three translation teams (see Table 2), minimizing group effects caused by both the translation teams and the translation methods. We fielded these questionnaire versions in Wave 5 of the CRONOS Panel, a probability-based online panel of the ESS in Estonia, Great Britain, and Slovenia.[6]

Participants. The participants of our study came from the participant pool of ESS Round 8, conducted in 2016. After their face-to-face interview within the ESS context, they were invited to participate also in the CRONOS panel, which consisted of six waves spanning 12 months. A total of 499 respondents in Estonia and 615 respondents in Slovenia participated in our experiment. They were all randomly assigned to the three questionnaire versions in each country: In Estonia, 173 persons filled in questionnaire version 1, 153 persons received version 2, and 173 persons responded to version 3. In Slovenia, 194 respondents answered questionnaire version 1, 217 respondents version 2, and 204 respondents version 3.

Table 2. Composition of questionnaire versions with regard to the translation method and translation team.

	Version 1		Version 2		Version 3	
Items	*Method*	*Team*	*Method*	*Team*	*Method*	*Team*
1-3	Adaptation	3	Adaptation	1	Close	2
4-22	Adaptation	2	Close	3	Close	1
24-41	Close	1	Adaptation	2	Adaptation	3
42-58	Adaptation	3	Adaptation	1	Close	2
59-60	Close	1	Adaptation	2	Adaptation	3

Note. The numbering of items in the questionnaire versions does not correspond to the numbering of items in the batches that were sent out to the translation teams. Numbering in this table simply reflects the order of items in the final questionnaire.

Procedure. The online survey lasted approximately 20 minutes. Participants who did not have Internet access at home received a tablet and an Internet connection for the duration of the entire panel, that is, the welcome survey and all six waves.

Instruments. The questionnaire consisted of items borrowed from the ESS, the ISSP, the European Values Study (EVS), and the Survey of Health, Ageing and Retirement in Europe (SHARE). Topics covered were normative beliefs about female sex roles, procreation, societal attitudes, individual attitudes toward gender roles, life planning, political participation/engagement, political efficacy, national identity, church involvement, beliefs, generalized social trust, social network, and perceived consequences of social policies, among others. Most of these latent constructs were measured with at least three items.

Analytical strategy

Challenges

When analyzing our data, we had to consider two challenges: a methodological and a conceptual one. First, from a statistical point of view, the respondent groups in each condition (i.e., questionnaire version) were too small. This made it impossible to use available methods (such as multigroup confirmatory factor analysis or item response theory) to test for measurement equivalence. Second, although the translation teams had received instructions on how to translate and although we employed manipulation checks, conceptually, we could not be sure that the translations would match the instructed approach. In fact, according to the manipulation checks, one team in each country failed to notice that they had received different instructions for the three batches. This may mean that the teams translated all items in the same way regardless of the instructions. But even if the teams had noticed differences in the instructions, we would not know whether the outcome would correspond to the instructions. That is because they may have—consciously or unconsciously—simply translated as they always do, or because their resulting translation may not be consistent with the intended translation approach despite trying to follow the instructions.

Further, items may vary in their potential for close or adaptive translation. For instance, one could imagine an item that would need to be adapted because a literal translation would not make sense. Hence, asking the translators to stick to a close approach could lead to two scenarios: (a) the item being closely translated (which would be in line with the instructed method but might result in a bad or, even worse, an

incorrect translation) or (b) the item being adaptively translated (which would be linguistically correct but not necessarily in line with the intended method). It follows from this conceptual challenge that we cannot deduce from the mere instructions that the resulting translations were correct (i.e., according to the assigned translation method and, ideally, also linguistically correct).

Development of a new analytical method

In our first attempt to overcome these challenges, we asked two independent translators (i.e., one per country) to assess whether the final translations were close or adaptive regarding the English source text. While they were able to give qualitative explanations on each segment of the items (e.g., sentence, response category), it proved to be very difficult to classify the items as either close or adaptive. The reason for this is that some items contain close and adaptive elements, and to complicate matters further, these elements may also vary in their degree of closeness and adaptiveness. Despite these difficulties, we established from the first feedback of our independent experts that the final translations did not always match the instructed translation approach.

Consequently, we developed a new explorative analytical framework that is associated with greater precision than the “simple” assessment of the entire item as either close or adaptive. The method is based on the assumption that every item (consisting of a question and an answer scale) has the following: a translation potential and as many translation scores as actual translations. The *translation potential* is the theoretical translation space of an item. It is target-language specific and includes all possible translations that a source item could have in the target language. The underlying notion is that items differ in their potential in how they could be translated. For some items, only a close translation may seem sensible. Other items may need an adaptive translation in order to make sense in the target language. Then again, there might be items for which close and adaptive translations may both be possible. Note, the translation potential can be different for the question and the answer scale (on the specific challenge to translate answer scales in a cross-national context, see, for instance, Villar 2009).

The *translation score*, in return, refers to the actual translation realized within the universe of all possible translations. On a continuum ranging from very close to very adaptive, each translation (in relation to its source version) can be placed on this dimension, and a specific value of closeness or adaptiveness can be attributed to its location. In other words, if two people translate the same item, there is a possibility that they produce slightly different translations. These translations may differ in terms of their degree of implemented closeness or adaptiveness. Thus, they can be placed in different parts on the closeness-adaptiveness continuum and therefore receive different translation scores. These scores are translation-specific. As each item in the experiment consisted of a question element (including the question itself, but also its introduction) and an answer scale element (the given answer options), each item had six translation scores (i.e., 3 translation teams x 2 item elements).

In a two-step evaluation process, our experts first evaluated the degree of closeness and adaptiveness for each translation on a 7-point Likert scale ranging from –3 (*overly close*) to 3 (*overly adaptive*). Additionally, our experts had the option to choose wrong translation whenever the translation did not capture the content of the English source

	overly close	close	rather close	somewhat close and somewhat adaptive	rather adaptive	adaptive	overly adaptive	wrong translation
	-3	-2	-1	0	1	2	3	999
Q								
S								

Figure 1. Closeness-adaptiveness scale for the evaluation of the translation scores of the question (= Q) and the answer scale (= S) of an item.

	close	rather close	somewhat close and somewhat adaptive	rather adaptive	adaptive
	-2	-1	0	1	2
Q					
S					

Figure 2. Closeness-adaptiveness scale for the evaluation of the translation potential of the question (= Q) and the answer scale (=S) of an item.

text (see Figure 1). Second, the two experts assessed each question's and each answer scale's translation potential by thinking up all possible translations that would make sense in the target language and culture. On a 5-point Likert scale ranging from –2 (*close*) to 2 (*adaptive*), they marked the entire space that seemed plausible (see Figure 2). After the two steps, we combined the two evaluation types and visualized them.

Our final analytical strategy was a mixed-methods design incorporating data triangulation and three different approaches (Creswell 2009; Hussy, Schreier, and Echterhoff 2013). First, we tested significant mean differences between the three respondent groups within each country for all items. We based this on actual *survey data* and applied the Kruskal-Wallis H test for continuous and ordinal variables (Kruskal and Wallis 1952).[7] In case of significance, we conducted the Dunn test with Bonferroni adjustment to determine which of the three groups was different (Dinno 2015). For binary and nominal variables, we applied Pearson's chi-squared test, followed by Fisher's exact test (Fisher 1922). Second, we compared the mean differences found with the translation potential and the translation scores based on *expert evaluations*. Third, we added the experts' *qualitative explanations* to understand how statistical mean differences can be explained by differences in translation.

Results

Translation instructions and translation outcome

According to the experimental design, all items should have been translated closely and adaptively at least once. For 60 items, each consisting of a question and an answer scale,

we had 360 expert evaluations per language (i.e., 60 items x 2 item elements x 3 translations/translation teams). Surprisingly, the follow-up interviews with the reviewers and translators revealed that, in each country, there was one team that did not realize that they had received different instructions for different batches. This means that they did not understand the main intention of the translation instructions (i.e., the explanation of the method to apply: close translation or adaptation) correctly.

Likewise, when comparing the translation instructions with the expert evaluations, it became apparent that the items had not necessarily been translated the way they were supposed to (see Table 3). On closer examination, we find that instructions to translate closely are more often associated with close translation outcomes (EE: 79.1%; SI: 83.5%) than instructions to translate adaptively are related to adaptive translation outcomes (EE: 6.9%; SI: 16.3%). This may be because, generally, it is cognitively more manageable for translation teams to stay as close as possible to the source text or because they feel more comfortable translating closely. Likewise, close instructions result in adaptive translations in far fewer cases (EE: 7.0%; SI: 14.6%), while adaptive instructions produce close translations in more than 80% of the cases (EE: 84.7%; SI: 81.7%). It is noticeable that only for one item in Slovene do the instructions match the translation outcomes of all three translation teams. Comparing both languages, Estonian translations were more frequently classified as somewhat close and somewhat adaptive (close instructions: 13.9%; adaptive instructions: 8.4%) than Slovene translations (close instructions: 1.9%; adaptive instructions: 1.5%), irrespective of the instructions. While this could be a language-specific observation, it could also be an artifact due to the different coding styles of our experts.

There might be another explanation for the overrepresentation of close translations. Although the items selected for the experiment had already proven problematic in the translation process of other surveys in the past, they may still lack an adaptive translation potential. Table 4 lists the translation potential for each question and each answer scale. Note that about one-third of the questions and 70% (SI) to 75% (EE) of the answer scales had an exclusively close translation potential. In contrast, no item in Estonian and only one question and two answer scales in Slovene had a solely adaptive translation potential. Yet, in more than half of the questions (EE: 33; SI: 35), an adaptive translation would have been at least possible, whereas there were only three answer scales in Estonian and 18 answer scales in Slovene for which an adaptive translation was seen as imaginable (see Table 5).[8]

Table 3. Instructions versus translation scores.

	Estonian			Slovene	
	Close	*Adaptive*		*Close*	*Adaptive*
+	125 (79.1%)	14 (6.9%)	+	132 (83.5%)	33 (16.3%)
O	22 (13.9%)	17 (8.4%)	O	3 (1.9%)	3 (1.5%)
–	11 (7.0%)	171 (84.7%)	–	23 (14.6%)	165 (81.7%)
Ø	0 (0.0%)	0 (0.0%)	Ø	0 (0.0%)	1 (0.5%)
Total	158 (100%)	202 (100%)	**Total**	158 (100%)	202 (100%)

Note. Based on a 7-point Likert scale ranging from -3 (overly close) to 3 (overly adaptive) (see Figure 1), the matches between the translation instructions and the realized translation scores were coded the following: + Instructions and translation scores matched (i.e., instructions were close, and translation scores were either -3, -2, or -1; instructions were adaptive, and translation scores were either 1, 2, or 3). – Instructions and translation scores were opposite (i.e., instructions were close, but translation scores were either 1, 2, or 3; instructions were adaptive, but translation scores were either -3, -2, or -1). o The translation score was the middle category (i.e., 0) despite the instructions. ø Translation was wrong.

Table 4. Translation potential for each question and answer scale.

	Estonian			Slovene	
	Question	*Answer Scale*		*Question*	*Answer Scale*
c	21	45	c	20	42
o	0	0	o	0	0
a	0	0	a	1	2
co	6	12	co	5	0
coa	30	2	coa	18	11
oa	3	1	oa	2	1
ca	0	0	ca	14	4
Total	60	60	**Total**	60	60

Note. Based on a 5-point Likert scale ranging from -2 (*close*) to 2 (*adaptive*) (see Figure 2), the translation potential for each question and answer scale was coded in the following way: *c* The translation potential was either close (-2) or rather close (-1) or both. *o* The translation potential was somewhat close and somewhat adaptive (0). *a* The translation potential was either rather adaptive (1) or adaptive (2) or both. *co* The translation potential included at least one close category and the middle point of the rating scale. *coa* The translation potential included close and adaptive parts of the scale and the middle point. *oa* The translation potential included the middle category and at least one adaptive category.

Table 5. Translation possibilities.

	Estonian		Slovene	
Possible translation	*Question*	*Answer Scale*	*Question*	*Answer Scale*
close	57	59	57	57
adaptive	33	3	35	18

Note. Based on the translation potential, these are the counts of questions and answer scales for which a close translation would have been possible and for which an adaptive translation would have been possible. The unit of reference is 60 for the number of items. As a question and an answer scale can have both a close and adaptive translation potential, they may be counted twice in this overview.

Types of adaptation

When analyzing the translations, we differentiated between four types of adaptation: pragmatic-semantic-lexical, factual-technical, syntactic-grammatical, and (survey) methodological adaptations. First, *pragmatic-semantic-lexical* covers adaptations made because a word or expression is different in the target language (adaptation refers to the meaning of the language). Second, the category *factual-technical* captures translation issues that result from differences between source and target cultures, systems, or structures (e.g., different legal or administrative settings) (adaptation results from differences in reality). Third, *syntactic-grammatical* includes syntactic or grammatical modifications because of purely linguistic reasons (adaptation due to the structure of the language). Fourth, *(survey) methodological* adaptations refer to the changes made due to specific survey habits in the target language/culture, as often found in distinct formulations of answer categories (adaptation due to survey habits of the country, so-called “survey speak”). We developed our adaptation typology using the categorizations of Van de Vijver and Leung (2011) and Behr and Shishdo (2016) as starting points.

Overall, we found more adaptations in Estonian. Nevertheless, the distribution of the adaptations across the first three categories was relatively similar in both languages. Interestingly, 20% of the adaptations in Slovene could be categorized as survey methodological adaptation, whereas only 8.2% of Estonian adaptations fell into this category (see Table 6).

Table 6. Types of adaptation.

	Estonian	Slovene
Pragmatic-semantic-lexical	116	69
(meaning-language)	(67.8%)	(60.0%)
Factual-technical	28	17
(reality)	(16.4%)	(14.7%)
Syntactic-grammatical	13	6
(linguistic structure)	(7.6%)	(5.2%)
(Survey) methodological	14	23
(survey habits)	(8.2%)	(20.0%)
Total	171	115

Note. We only counted distinct adaptations. That is, if several items had, for example, the same answer scale, which was always adapted in the same way, we only counted it once.

Table 7. Mean differences and expert evaluations in Estonian.

Differences in		*Means*	
		yes	no
Expert Evaluations	**yes**	1, 2, 3, 5, 6, 7, 8, 9, 10, 16, 31, 35, 36, 47, 49, 50, 53, 60	4, 11, 12, 14, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 32, 33, 34, 38, 39, 40, 42, 44, 45, 48
	no	46, 54, 55, 56, 57, 58	13, 15, 23, 37, 41, 43, 51, 52, 59

Note. Numbers are item numbers in the order they appear in the questionnaire. Whenever we found statistically significant mean differences between at least two respondent groups for an item in the survey data, we classified it as “differences in means: yes” and “differences in means: no” otherwise. Whenever the Estonian expert coded at least one of the realized translations (i.e., the translation score) as closer or more adaptive than the others on a 7-Likert scale, the item was classified as “differences in expert evaluations: yes” and “differences in expert evaluations: no” otherwise.

General patterns

To generate some patterns across all 60 items in the two languages, we first checked for significant mean differences across the three respondent groups and contrasted these with differences in the expert evaluations (see Table 7 for Estonian and Table 8 for Slovene). Whenever we found mean differences between at least two respondent groups at the significance level of .05 or higher, the item was categorized as “differences in means: yes” and as “differences in means: no” otherwise. Similarly, whenever our experts found at least one of the three translations of the question or the answer scale to be closer or more adaptive than the others on the 7-point Likert scale, the item was categorized as “differences in expert evaluations: yes.” The item was counted as “differences in expert evaluations: no” when all three translations received the same translation score.

In both languages, the expert evaluations revealed that the three translations of most items were different along the closeness-adaptiveness dimension (EE: 45 items; SI: 57 items). Yet, among these items, we only found differences in the means for a much smaller proportion (EE: 18 items; SI: 15 items). While the observed patterns may be due to the small sample size, they could also support the idea that some items are more robust than others. That is, despite the different translations of an item, the meaning and understanding of the concept itself can be preserved. In contrast, other concepts might be more sensitive to the exact wording so that the slightest linguistic change might provoke a

Table 8. Mean differences and expert evaluations in Slovene.

Differences in		*Means*	
		yes	no
Expert Evaluations	**yes**	1, 4, 7, 13, 24, 26, 28, 31, 34, 37, 42, 46, 52, 53, 56	2, 3, 5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 19, 21, 22, 23, 25, 27, 29, 30, 32, 33, 35, 36, 38, 39, 40, 41, 43, 44, 45, 47, 48, 49, 50, 51, 54, 55, 58, 59, 60
	no		18, 20, 57

Note. Numbers are item numbers in the order they appear in the questionnaire. Whenever we found statistically significant mean differences between at least two respondent groups for an item in the survey data, we classified it as "differences in means: yes" and "differences in means: no" otherwise. Whenever the Slovene expert coded at least one of the realized translations (i.e., the translation score) as closer or more adaptive than the others on a 7-Likert scale, the item was classified as "differences in expert evaluations: yes" and "differences in expert evaluations: no" otherwise.

Table 9. Types of items.

Differences in		*Means*	
		yes	no
Expert	**yes**	sensitive items	robust items
Evaluations	**no**	error	same/similar wording

different answer in the respondent. Further, when the experts found no differences in the translations, we would not expect any mean differences unless there was some kind of error (e.g., random error or respondent groups were different with respect to the concept of interest). Table 9 presents these considerations. In total, there are 18 items in Estonian and 15 items in Slovene that fall into the category of sensitive items, while 27 items in Estonian and 42 items in Slovene may be considered robust. Interestingly, there are (almost) no items in the error category (EE: 6; SI: 0), lending support to the error idea. Finally, nine of the 60 Estonian items and three of the Slovene items appeared to have similar wording in the different translations. In fact, the direct comparison of item type by language (see Table 10) shows that only 23 items were classified the same way with respect to the four categories (i.e., sensitive items, robust items, similar/same wording, error), highlighting the importance of the target language in the translation process.

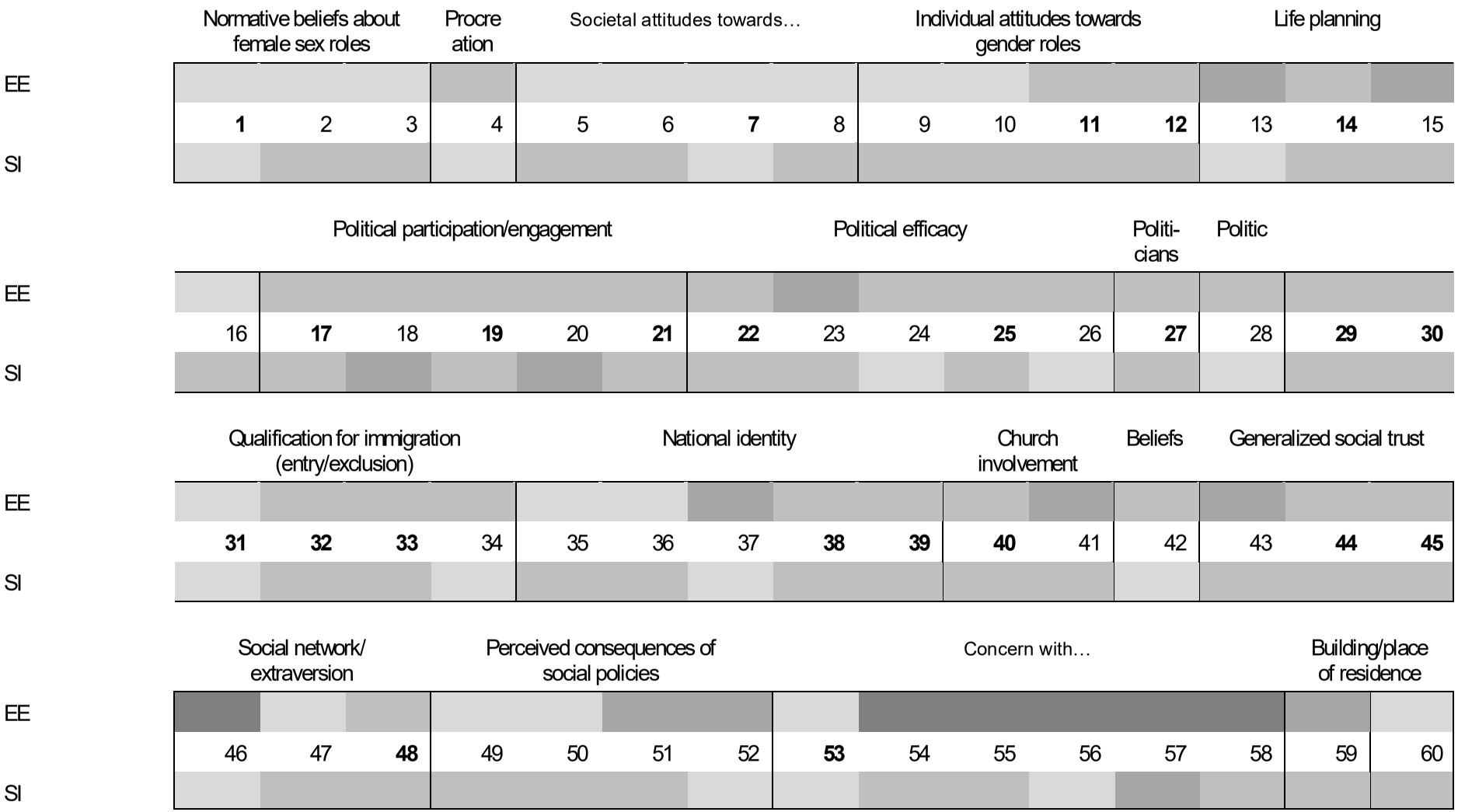
Specific examples: plan the future and ability to take active role in political group

Since the analysis of all 60 items would go beyond the scope of this paper, we have selected two of the sensitive items to illustrate our newly developed analytical strategy, which combines survey data with expert evaluations and qualitative explanations. First, we chose item 13, one of four items measuring the theoretical concept of life planning. The source item stems from ESS Round 3, item D52, and reads:

> *Do you generally plan for your future or do you just take each day as it comes?*
>
> *Please express your opinion on a scale of 0 to 10, where 0 means 'I plan for my future as much as possible' and 10 means 'I just take each day as it comes'.*

Table 10. Item Type Comparison.

Item Type Comparison



Note: The numbers are item numbers. ☐ Sensitive items; ☐ robust items; ☐ same/similar wording; ☐ error. **Bold** – item falls into the same item type in both languages. The lines group the items together, which belong to the same theoretical concept.

0 I plan for my future as much as possible

1

2

3

4

5

6

7

8

9

10 I just take each day as it comes

In Figure 3, we visualized the mean comparisons of the respondent groups (survey data) as well as the translation potential and the translation scores (expert evaluations) of item 13 in Slovene. Respondent groups 2 and 3 (bold-italic) should have received close translations by design and actually did so, as the translation scores show (the translations of both groups were rated as close, with the exception of the question of respondent group 3, which was rated as overly close). Likewise, respondent group 1 (bold) should have received an adaptive translation and did so. This is reflected in the mean differences. Respondent group 1 (M_1 = 3.8) differs significantly from groups 2 and 3 ($M_{2,3}$ = 5.3; p <.0000). What stands out is the translation potential, which was previously evaluated as adaptive but not as close; yet, the expert had rated translation versions 2 and 3 as close. Even more interesting is that the translation of the question received in group 3 was rated as overly close. All taken together, this may point to the direction that close translation instructions may lead to unnatural translations when the theoretical translation space does not include a close translation potential.

Looking at the response patterns of the individual respondent groups (see Figure 4), the most striking difference is the frequency with which group 1 chose the extreme categories “0 – I plan my future as much as possible” (more than twice as often as the other two groups) and “10 – I just take each day as it comes” (between four to five times less often than the other groups). According to the qualitative explanations, “take each day as it comes” was translated as “to live for each day” in group 1, which is a very common Slovene idiom. In contrast, the other groups received rather literal translations, which seemed to be unusual ways of expressing the concept, again supporting the idea that close translation instructions may yield functionally nonequivalent translations when the translation potential is adaptive.

Second, we selected item 24, which is one of five items measuring the latent concept of political efficacy in ESS Round 8 (item B3):

How able do you think you are to take an active role in a group involved with political issues?

1 - Not at all able

2 - A little able

		close				adaptive		wrong translation
	-3	-2	-1	0	1	2	3	999
Q	*3*	***2***				**1**		
S		***2 3***				**1**		

1 vs. 2***
1 vs. 3***

Note. The scale ranges from -3 (*overly close*) to 3 (*overly adaptive*). Numbers in the rows of Q (= question) and S (= answer scale) represent the three translation versions/respondent groups. Their placement on the scale reflects the specific translation score. The numbers below are significant mean differences between respondent groups.
Translation potential. ***Bold-italic*** – translation instructions were close (2, 3). **Bold** – translation instructions were adaptive (1).
***$p < .001$.

Figure 3. Expert evaluations and mean differences for “Do you generally plan for your future or do you just take each day as it comes?” (Slovene).

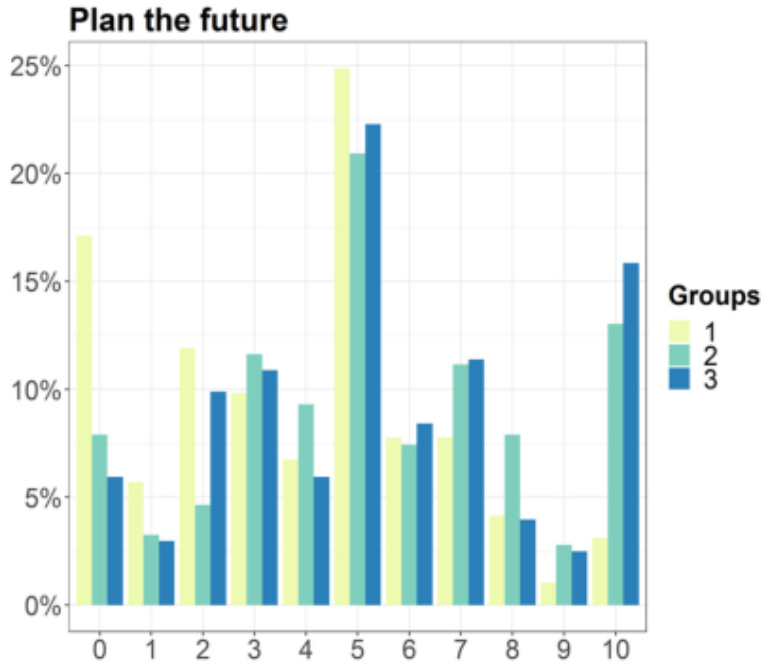


Figure 4. Response pattern for “Do you generally plan for your future or do you just take each day as it comes?” (Slovene).
Note. Answer scale ranges from 0 (*I plan for my future as much as possible*) to 10 (*I just take each day as it comes*). Groups are respondent groups. Group 1 received an adaptive translation, groups 2 and 3 close translations.

3 - Quite able

4 - Very able

5 - A great deal

Figure 5 summarizes the mean comparisons of the respondent groups, the translation potential, and the translation scores of item 24 in Slovene. Although respondent groups

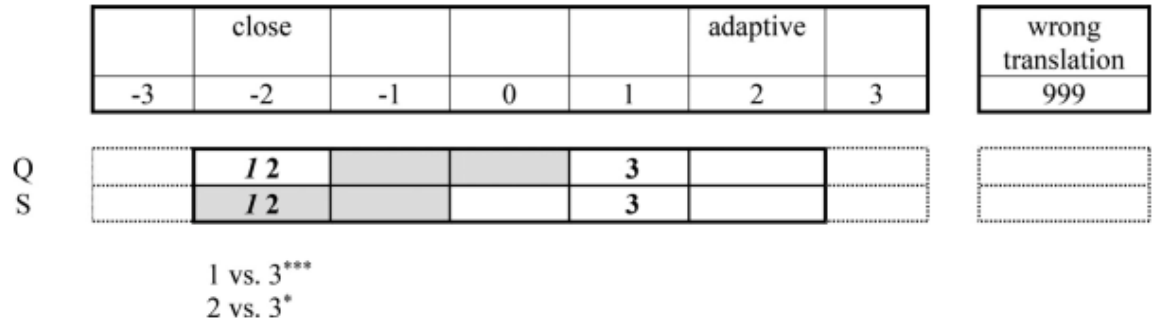


Figure 5. Expert evaluations and mean differences for "How able do you think you are to take an active role in a group involved with political issues?" (Slovene).
Note. The scale ranges from -3 (*overly close*) to 3 (*overly adaptive*). Numbers in the rows of Q (= question) and S (= answer scale) represent the three translation versions/respondent groups. Their placement on the scale reflects the specific translation score. The numbers below are significant mean differences between respondent groups. Gray Box – translation potential. ***Bold-italic*** – translation instructions were close (1). **Bold** – translation instructions were adaptive (2, 3).
*$p < .05$. ***$p < .001$.

2 and 3 (bold) were supposed to receive adaptive translations, only group 3 did. Respondent groups 1 (bold-italic) and 2 (bold) both got close translations. The fact that groups 1 and 2 received similar translations (although this was not intended by design) is also reflected in the mean differences which we found between groups 3 (M_3 = 1.99) and 1 (M_1 = 2.32; p = .0004), and groups 3 and 2 (M_2 = 2.19; p = .0262), but not between groups 1 and 2 (which both had close translation versions). The expert evaluated the translation potential of this item on the close side of the scale (-2 to 0).

Moreover, the qualitative explanations reveal two major differences in the translation version respondent group 3 received: (1) "group involved with political issues" was translated as "politically active group" and (2) "able" was translated as "competent." Both versions differ in content from the original source item. The second adaptation is particularly striking, as the adjective "able" reflects internal and external circumstances or constraints of participating, whereas "competent" covers only internal capacities. Based on this distinction, we expected that the mean of group 3 (the "competent" group) would be lower than the mean of the other two groups (the "able" groups). That is because "able" consists of two components (i.e., internal and external), while "competent" consists of only one (i.e., internal). As such, people who do not feel competent but would, nonetheless, have the external possibilities (i.e., are still able) to take an active role in such a group might respond more positively and less negatively on the scale. Indeed, Figure 6 mirrors this expectation. The respondents in group 3 chose the middle category (quite able/competent) less often and the extreme category on the negative side of the scale (not at all able/competent) more often than the other groups. However, we cannot say with certainty whether this effect is due to the first (i.e., "politically active group") or the second (i.e., "competent") translation difference.

Discussion

In this experimental study, we investigated two translation approaches: close translation and adaptation. In doing so, we challenged the assumption often made in cross-

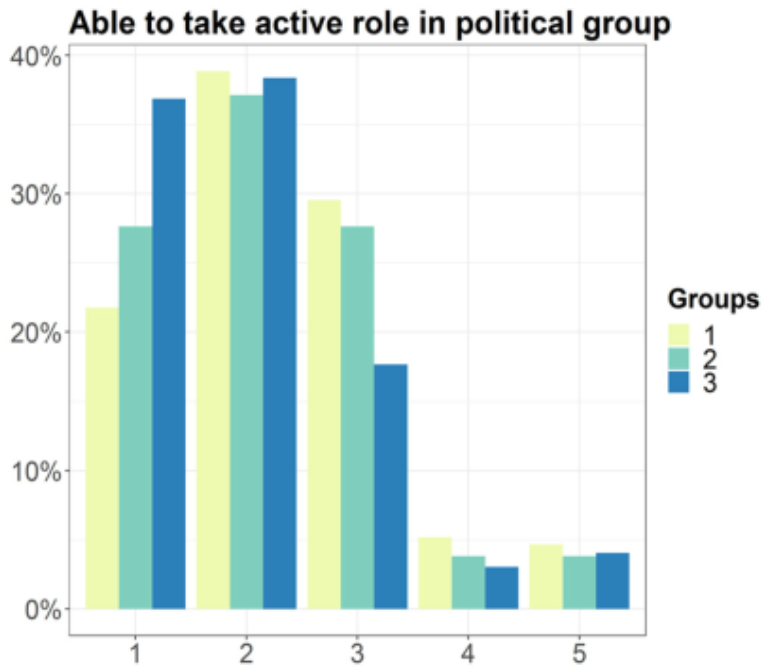


Figure 6. Response pattern for “How able do you think you are to take an active role in a group involved with political issues?” (Slovene).*Note*. Answer scale ranges from 1 (*not at all able*) to 5 (*a great deal*). Groups are respondent groups. According to expert evaluations, groups 1 and 2 received close translations, group 3 an adaptive translation.

national survey projects that close translation provides more comparable data than adaptation. In particular, we investigated whether the instructions given to translators to follow one of the two translation approaches would lead to different translation versions and whether these translation versions would be associated with changes in the target population’s response behavior. We studied 60 items frequently used in multinational surveys such as the ESS, the ISSP, the EVS, and SHARE. Three teams translated these items from English into Estonian and another three teams from English into Slovene. Each translation team consisted of three experienced professionals. The translation experiment was implemented in the fifth wave of CRONOS, a probability-based online panel whose participants had previously been interviewed in ESS Round 8.

We developed a new analytical framework that allowed us to evaluate a translation based on (a) the theoretical translation potential of its source version for the specific target language and (b) its specific translation score on a closeness-adaptiveness continuum. Further, we related this to the response behavior of the participants. Our analysis suggests that some items might be more sensitive to the wording (i.e., small linguistic changes lead to a different response behavior in the respondent), while other items might be more robust (i.e., meaning and understanding of a concept remain the same despite linguistic changes). Furthermore, we showed that items vary in their potential for close and adaptive translation. Consequently, instructing translators to use a specific approach that does not fit the linguistic or cultural aspects of the target language or population could lead to translations that are too close or too adaptive, or in the worst case, even wrong. All of which is to be avoided as a translation should sound natural (Harkness, Pennell, and Schoua-Glusberg 2004).

Moreover, follow-up interviews with members of the translation teams revealed that four out of six teams did not realize that they had received different translation instructions throughout the experiment. From this observation, our take-home message is that written translation guidelines may not be sufficient for instructing translation teams, not even if all team members are familiar with questionnaire design and translation matters. Additional training methods seem beneficial, such as practical training or spot-checks, to see whether the instructions were understood correctly before the actual translation process. In-person meetings involving all translators or the senior person responsible for each language may be necessary since direct contact seems to be more rewarding for training new collaborators. The ESS, for example, has been organizing such translation meetings with adjudicators since ESS Round 8. These meetings bring together the ESS translation team, representatives of the developers of the source questionnaire, and senior representatives of all the languages participating in a particular survey round. An ideal training scenario for translation teams in cross-national survey projects probably consists of a mixture of written and electronic training material combined with live training and the possibility to record, save, and make these trainings available to all those requiring more detailed information or those who need to get back to the material. Note that live or face-to-face training does not necessarily need to happen in person but could be held via online tutorials or webinars with subsequent question-and-answer sessions. In addition, this could be recorded and put on the project's website for later reference and transparency in the translation process.

The generalizability of our analysis is subject to certain limitations. First, we only had one independent expert in each language who coded the translation potential and the translation scores along the closeness-adaptiveness dimension. It would have been better to have had at least three coders to see if differences between the two languages were not due to our experts' distinct coding styles but due to the target languages themselves. Second, once a translation contained several adaptations, we could no longer disentangle which adaptation led to what exact change in the response behavior. Third, the small sample size did not allow us to statistically check for measurement equivalence, which would have been an excellent addition to our analysis. Finally, the translation teams did not always notice that they had received different translation instructions, which inflated the complexity of our study. As foreshadowed above, appropriate training of the translation teams and checking their correct understanding rather than relying solely on written instructions could have mitigated this effect.

While we chose an experimental setting that was as close as possible to the natural environment of the translation processes typically found in cross-national survey projects, future research could choose a more controlled setting. For instance, it would be conceivable that translation teams were not involved at all, but that instead, a close translation was taken as a starting point and then adapted gradually. The resulting translation versions could then be tested systematically. Moreover, one could think of a split-ballot multitrait-multimethod (MTMM) design, in which random sub-samples of the sample population received more than one translation version, thus opening up the possibility of also obtaining information on the reliability and validity of each item version (Saris, Satorra, and Coenders 2004). In any case, larger sample sizes are required for statistical analysis. Future research could also examine other combinations of source

and target languages to see if our findings can be generalized to other populations. If, for example, certain items appear to be robust or sensitive for certain language groups, this may give valuable guidance for translation teams. It could indeed contribute to a more efficient allocation of time and resources, as translators could focus on sensitive items and invest in robust items only what is necessary.

Despite all these limitations and directions for future studies, our research provides first experimental evidence that, in questionnaire translation, close translation does not always yield better translations than adaptation. In fact, under certain circumstances, an adaptive translation would be preferable to close translation if the latter led to unnatural or incorrect translations or made it more challenging to understand them. This finding empirically supports the recommendations of Harkness (2003, 2007) and her colleagues (Harkness, Pennell, and Schoua-Glusberg 2004) to combine both approaches to obtain more appropriate translations that take into account the cultural and linguistic characteristics of the target population.

Moreover, our research makes a valuable contribution to translation studies as we developed a novel analytical framework for assessing and rating the level of close versus adaptive translation against a source text by using a mixed-methods approach. Human expert translators measured the level of closeness and adaptiveness of the translated text (i.e., the translation scores) on a continuum ranging from overly close to overly adaptive. They also quantified the translation potential of the source text for the specific target language. We then cross-checked these measures against the qualitative assessments of the experts.

We started with a sequential mixed-methods design by first gathering qualitative data from two expert translators and then asking them to rate the source and target texts quantitatively. In the end, however, our analysis followed a concurrent mixed-methods approach, in which the qualitative explanations added complementary explanations to the quantitative ratings (cf. Creswell 2009). The combination of both information types allowed us to better understand the mechanisms of close and adaptive translation in the context of survey translation. This approach could be transferred to other text types, different translation settings, or used in translation training. Besides, we linked these expert evaluations to actual survey data. This data triangulation added to the robustness of our findings. However, this method may not be feasible for many other text types that cannot be immediately linked to statistical data. Nevertheless, this extended approach could enrich the research landscape of translation studies, especially in cross-cultural survey research.

Furthermore, the application of measuring the level of closeness in the context of questionnaire translation is, to our knowledge, innovative. So far, in translation studies, the level of closeness has mainly been measured to assess the quality of machine translation (MT), where the level of closeness is used as a benchmark for translation performance (for an example of measuring closeness in translation for assessing MT performance, see Biçici, Groves, and van Genabith 2013). In contrast, in our study, we used the measurement of closeness to evaluate human translations.

Returning to the topic of setting translation guidelines for cross-cultural survey projects, we recommend the following: In general, we need more flexibility and acceptance of a higher degree of adaptive or free translation. However, continued efforts are still

required to work toward a better understanding of where adaptation would be useful. Cross-cultural survey projects could, for instance, include an additional step in the pretesting stage before the actual translation process starts, in which human experts determine the *translation potential* of the prefinal source text for a variety of language families. Next, the source text should be translated according to the state-of-the-art method (i.e., the team approach). Then, these translations should be pretested, and finally, the pretest data should be triangulated with the *translation scores* gained by human experts. Where this data triangulation shows that too close or too adaptive translations hamper the quality of the data, social scientists could develop more informed guidelines concerning the acceptable or targeted level of closeness in the final translation stage.

Since research on closeness and adaptation in survey translation is still relatively new and highly dependent on the language pair, the application of our analytical approach has many potentials. In particular, our procedure would be useful to gain more systematic knowledge in various languages, language groups, and cultures and, as such, could contribute to the creation of a rich database across survey projects. Such a database, if properly maintained, could incorporate more and more empirical data over time, that is, data on which items can be considered robust and which ones need to be considered sensitive to different translation approaches. Plus, such a database would make it easier for those setting up the translation processes in similar survey projects to make informed decisions about the level of closeness or the degree of acceptable adaptation in their questionnaire translations.

In closing, the findings from our translation experiment address questions relevant to the translation of questionnaires and could, therefore, be informative, especially for those involved in the design of cross-national surveys, for translators, but also more generally for those who wish to make cross-national comparisons using survey data. Despite our mixed-methods approach, we recommend to those who perform cross-country comparisons to test for measurement equivalence before drawing possibly erroneous conclusions (e.g., Byrne and van De Vijver 2010; Kankaraš and Moors 2010; Reeskens and Hooghe 2007; for a systematic review on invariance testing in cross-cultural research, see Boer, Hanke, and He 2018). But whenever this is not possible, we would, at least, like to draw attention to the importance of translations as a potential adjusting screw that influences the comparability of survey data.

Notes

1. Please note that in the meantime, the ESS explicitly pursues a policy of encouraging functionally equivalent translations.
2. Kleiner, Pan, and Bouic (2009) differentiate between equivalence of *stimulus* and equivalence of *effect* (i.e., formal and dynamic equivalence in Nida's terms).
3. Note that in this paper, the terms *free translation* and *adaptation* are used as quasi-synonyms. We thereby follow the logic that free translations apply different adaptation types to make sure that the target audience more readily understands the translated text and its message (i.e., intended meaning) in contrast to staying close to the source text on a semantic level. This also corresponds to a translation approach that focuses more on the target text (i.e., the text's receptor) than on the source text (on the relationship between the terms adaptation and free translation, see, e.g., Gambier 1992).

4. For more detailed information on the team composition as well as the organization of the review sessions, see Dorer (2019b).
5. For the exact translation instructions, see Dorer and Villar (2017).
6. One item on income had to be implemented in the 6th wave of the CRONOS Panel due to timing reasons of the waves and the formulation of the item itself. For more information on the CRONOS Panel, see European Social Survey (n.d.).
7. The data is openly available at https://www.europeansocialsurvey.org/download.html?file=CRONOS_Wave5_e01_1&y=2016.
8. Note that these numbers are not corrected for the fact that some of the answer scales are repeated, that is, some of the items share the same answer scale.

Acknowledgments

We are grateful to the anonymous reviewers for the useful comments they provided on an earlier version of this article. We are particularly thankful to Ana Villar, who was deeply involved in setting up the research design, and Natalja Menold, who gave valuable feedback on the experimental design. We are also indebted to the Estonian and Slovene national ESS teams for supporting us by carrying out the translation review sessions as well as the ESS Methods Advisory Board and the ESS Translation Expert Panel. We also thank various international colleagues for helping us to recruit qualified translation team members, the translation team members themselves, as well as our translation experts Epp Aareleid and Maruša Telban, who assisted us in making sense of the different translations.

Disclosure statement

The authors declare no potential conflict of interest with respect to the research, authorship, and/or publication of this article. Some of the ideas included in the present article are updated discussions of issues also presented in Dorer (2019a, 2019b) and Repke et al. (2019).

Funding

This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 654221.

Notes on contributors

Brita Dorer is a senior researcher at GESIS – Leibniz Institute for the Social Sciences (Mannheim, Germany) specialized in the translation of questionnaires for cross-cultural surveys. She holds a Ph.D. in translation studies from the Faculty of Translation Studies, Linguistics and Cultural Studies (FTSK), Johannes Gutenberg University Mainz-Germersheim, is a university-trained translator for English, French and Italian (German mother tongue), and holds a degree from the University of Freiburg (Germany) in interfacultative studies on France. She has been heading the translation team of the European Social Survey (ESS) since 2009 and acted as coeditor of the 2018 Wiley monograph on Advances in Comparative Survey Methods in the 3mc context. She has been teaching practical as well as methodological translation classes at the translation and applied language faculties of the universities of Strasbourg, Mainz-Germersheim and Geneva, and worked more than ten years as a full-time translator for English, French, and Italian. Her current scientific interests include the evaluation and quality enhancement of questionnaire translation, translatability and intercultural portability of source questionnaires, close versus adaptive translation, machine translation, and translation process research.

Lydia Repke is a postdoctoral researcher in the Department of Survey Design and Methodology at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany. She works there on the comparability of measurement instruments, equivalence testing, and questionnaire design. Further, her research interests include acculturation processes and multicultural identity, cultural and linguistic effects, and social network analysis. In particular, she integrates social-personality psychology and social network approaches to tackle the dynamic interplay of individual-level psychological variables (e.g., personality, cultural identification) and meso-level social-network measures (i.e., content-based and structure-related). Before joining GESIS, she worked at the Research and Expertise Center for Survey Methodology (RECSM) in Barcelona, Spain. Lydia Repke obtained her Ph.D. in Cultural Psychology at the Department of Political and Social Sciences at Universitat Pompeu Fabra (UPF) in Barcelona, Spain. Throughout her doctoral phase and her general university studies, she was funded by the Studienstiftung des deutschen Volkes (German Academic Scholarship Foundation). In 2019, UPF awarded Lydia Repke's Ph.D. thesis with the Special Doctorate Award. In 2020, she joined the Academy of Sciences and Literature | Mainz.

ORCID

Lydia Repke http://orcid.org/0000-0002-7907-4648

References

Arffman, I. 2012. "Unwanted Literal Translation: An Underdiscussed Problem in International Achievement Studies." *Education Research International* 2012:1–13. doi: 10.1155/2012/503824.

Bastin, G. L. 2011. "Adaptation." Pp. 3–6 in *Routledge Encyclopedia of Translation Studies*. London: Routledge.

Behr, D., and K. Shishido. 2016. "The Translation of Measurement Instruments for Cross-Cultural Surveys." in *The SAGE Handbook of Survey Methodology*, edited by C. Wolf, D. Joye, T. W. Smith, and Y. Fu. London: Sage.

Biçici, E., D. Groves, and J. van Genabith, 2013. "Predicting Sentence Translation Quality Using Extrinsic and Language Independent Features." *Machine Translation* 27 (3-4):171–92. doi: 10.1007/s10590-013-9138-4.

Boer, D., K. Hanke, and J. He, 2018. "On Detecting Systematic Measurement Error in Cross-Cultural Research: A Review and Critical Reflection on Equivalence and Invariance Tests." *Journal of Cross-Cultural Psychology* 49 (5):713–34. doi: 10.1177/0022022117749042.

Byrne, B. M., and F. J. R. van De Vijver, 2010. "Testing for Measurement and Structural Equivalence in Large-Scale Cross-Cultural Studies: Addressing the Issue of Nonequivalence." *International Journal of Testing* 10 (2):107–32. doi: 10.1080/15305051003637306.

Chesterman, A. 2016. *Memes of Translation: The Spread of Ideas in Translation Theory* (Vol. 123). Amsterdam: John Benjamins Publishing Company. 10.1075/btl.123

Creswell, J. W. 2009. *Research Design. Qualitative, Quantitative, and Mixed Methods Approaches* (3rd ed.). Thousand Oaks: Sage.

Crump, M. J. C., P. C. Price, R. Jhangiani, I.-C A. Chiang, and D. C. Leighton. 2018. *Research Methods for Psychology*. Brooklyn, NY: Brooklyn College Edition.

Dinno, A. 2015. "Nonparametric Pairwise Multiple Comparisons in Independent Groups Using Dunn's Test." *The Stata Journal: Promoting Communications on Statistics and Stata* 15 (1): 292–300. doi: 10.1177/1536867X1501500117.

Dorer, B. 2019a. *Outline of best practices for implementation of questionnaire translation approaches based on empirical findings, Deliverable 3.4 of the SERISS project funded under the European Union's Horizon 2020 research and innovation programme GA No: 654221.*

Dorer, B. 2019b. *Report on the translation process used in testing close vs. adaptive questionnaire translation approaches. Deliverable 3.2 of the SERISS Project Funded under the European*

Union's Horizon 2020 Research and Innovation Programme GA No: 654221. www.seriss.eu/resources/deliverables

Dorer, B. 2020. *Advance Translation as a Means of Improving Source Questionnaire Translatability? Findings from a Think-Aloud Study for French and German*. Berlin: Frank & Timme.

Dorer, B., and A. Villar. 2017. Standards for the implementation of the two survey translation approaches: the 'stay close to the source' approach & the adaptive approach. *Deliverable 3.1 of the SERISS Project Funded under the European Union's Horizon 2020 Research and Innovation Programme GA No: 654221*. www.seriss.eu/resources/deliverables

European Social Survey. (n.d.). *CROss-National Online Survey (CRONOS) panel*. Retrieved April 25, 2020, from http://www.europeansocialsurvey.org/methodology/methodological_research/modes_of_data_collection/cronos.html

European Social Survey. 2018. *ESS Round 9 Translation Guidelines*. London: ESS ERIC Headquarters.

Fisher, R. A. 1922. "On the Interpretation of v2 from Contingency Tables, and the Calculation of P." *Journal of the Royal Statistical Society* 85 (1):87–94. doi: 10.2307/2340521.

Gambier, Y. 1992. "Adaptation : une Ambiguïté à Interroger." *Meta: Journal Des Traducteurs* 37 (3):421–5. doi: 10.7202/002802ar.

Gentzler, E. 2001. *Contemporary Translation Theories*. Clevedon: Multilingual Matters.

Harkness, J. A. 2002. *An Outline of ESS Translation Strategies and Procedures*. P. 19. London: European Social Survey.

Harkness, J. A. 2003. "Questionnaire Translation." in *Cross-Cultural Survey Methods*, edited by J. A. Harkness, F. J. R. Van de Vijver, and P. P. Mohler. Hoboken: John Wiley & Sons.

Harkness, J. A. 2007. "Improving the Comparability of Translations." in *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by R. Jowell, C. Roberts, R. Fitzgerald, and G. Eva. London: Sage.

Harkness, J. A., B.-E. Pennell, and A. Schoua-Glusberg. 2004. "Survey Questionnaire Translation and Assessment." Pp. 453–73 in *Methods for Testing and Evaluating Survey Questionnaires*, edited by S. Presser, J. Rothgeb, M. Couper, J. Martin, & E. Singer. Hoboken: John Wiley and Sons.

Harkness, J. A., A. Villar, and B. Edwards. 2010. "Translation, adaptation, and design." Pp. 115–40 in *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. P. M. Pennell, B.-E. Pennell, and T. W. Smith. Hoboken: John Wiley & Sons.

Hussy, W., M. Schreier, and G. Echterhoff. 2013. *Forschungsmethoden in Psychologie und Sozialwissenschaften für Bachelor* (2nd ed.). Berlin: Springer. doi: 10.1007/978-3-642-34362-9.

Kankaraš, M., and G. Moors, 2010. "Researching Measurement Equivalence in Cross-Cultural Studies." *Psihologija* 43 (2):121–36. doi: 10.2298/PSI1002121K.

Kleiner, B., Y. Pan, and J. Bouic, 2009. "The Impact of Instructions on Survey Translation: An Experimental Study." *Survey Research Methods* 3 (3):113–22.

Kruskal, W. H., and W. A. Wallis, 1952. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association* 47 (260):583–621. doi: 10.1080/01621459.1952.10483441.

Munday, J. 2016. *Introducing Translation Studies: Theories and Applications* (4th ed.). Abingdon: Routledge.

Newmark, P. 1981. *Approaches to Translation*. Oxford: Pergamon Press. Republished (2001) by Shanghai Foreign Language Education Press.

Nida, E. A. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. Leiden: Brill Archive.

Nord, C. 1997. *Translating as a Purposeful Activity: Functionalist Approaches Explained*. Manchester: St. Jerome.

Nord, C. 2005. *Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis* (No. 94). Amsterdam: Rodopi.

Reeskens, T., and M. Hooghe, 2007. “Cross-Cultural Measurement Equivalence of Generalized Trust. Evidence from the European Social Survey (2002 and 2004).” *Social Indicators Research* 85 (3):515–32. doi: 10.1007/s11205-007-9100-z.

Reiss, K. 1981. “Type, Kind and Individuality of Text: Decision Making in Translation.” *Poetics Today* 2 (4):121–31. doi: 10.2307/1772491.

Reiss, K., and H. J. Vermeer. 1984. *Grundlegung Einer Allgemeinen Translationstheorie*. Tübingen: Max Niemeyer.

Repke, L., B. Dorer, Y. Petttinicchi, and E. Sommer. 2019. Report on Empirical Findings from Applying both Translation Methods. *Deliverable 3.3 of the SERISS Project Funded under the European Union’s Horizon 2020 Research and Innovation Programme GA No: 654221.* www.seriss.eu/resources/deliverables

Saris, W. E., A. Satorra, and G. Coenders, 2004. “A New Approach to Evaluating the Quality of Measurement Instruments: The Split-Ballot MTMM Design.” *Sociological Methodology* 34 (1): 311–47. doi: 10.1111/j.0081-1750.2004.00155.x.

Van de Vijver, F. J. R., and K. Leung. 2011. “Equivalence and Bias: A Review of Concepts, Models, and Data Analytic Procedures.” Pp. 17–45 in *Cross-Cultural Research Methods in Psychology*, edited by D. Matsumoto and F. J. R. Van de Vijver. New York: Cambridge University Press.

Villar, A. 2009. Agreement answer scale design for multilingual surveys: Effects of translation-related changes in verbal labels on response styles and response distributions. In *Survey Research and Methodology program (SRAM)-Dissertations & Theses*. 3. http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1002&context=sramdiss

Vinay, J.-P., and J. Darbelnet. 1995. *Comparative Stylistics of French and English: A Methodology for Translation*. Amsterdam: John Benjamins Publishing Company.