Assessing the use of back translation: the shortcomings of back translation as a quality testing method

Behr, Dorothée

Postprint / Postprint
Zeitschriftenartikel / journal article

This is an accepted manuscript of an article published by Taylor & Francis in
International Journal of Social Research Methodology, Vol. 20, 2017, No. 6, pp. 573-584
available online at: https://doi.org/10.1080/13645579.2016.1252188
Page numbers have been adjusted to the publishers version, whereby this postprint is fully quotable.

Assessing the use of back translation: the shortcomings of back translation as a quality testing method

Dorothée Behr

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

Abstract

Back translation – the 're-translation' of a translated questionnaire back into the original language and the subsequent comparison of the original version and the back translation – is a controversial assessment method for questionnaire translations. Recently, cross-cultural survey methodologists have followed the call for more empirical research on this method. This article adds to the small body of research by drawing on the back translation documentation from the 2012 European Quality of Life Survey (EQLS). First, results from the official EQLS back translation step are contrasted with results when additional persons perform the same comparison step between back translation and original. We note inconsistency in the detection of presumed flaws. Second, the back translation outcome is contrasted with additional native speaker checks of the actual translation. While back translation can uncover problems, it causes quite a number of false alarms, and even more importantly, many problems remain hidden.

Keywords: Questionnaire translation; survey translation; back translation; assessment method; translation quality

1. Introduction

This article provides insights into the extent to which back translation, a quality assessment method for questionnaire translation, may or may not contribute to translation quality. Evidence is urgently needed in the research community with regard to this method. There is hardly any method in questionnaire translation that is as controversially discussed as back translation. In its simplest form, back translation involves the translation of a questionnaire back into the original language and the subsequent comparison of the two original-language versions. The goal is to identify discrepancies between these two versions that might be due to errors in the actual translation, that is, the translation from the source into the target language. Back translation is strongly associated with the cross-cultural psychologist Brislin (1970), who brought this method to a heightened attention. However, already back in the 70s, Brislin noted that one should not rely solely on the back translation method. Since the 1970s, a large number of translation guidelines have been drafted, most notably across the health disciplines, that list back translation as one of the recommended assessment methods (many of these guidelines are nicely summarized in Acquadro, Conway, Hareendran, & Aaronson, 2008; see also Beaton, Bombardier, Guillemin, & Bosi Ferraz, 2000; Maneesriwongul & Dixon, 2004; Wild et al., 2005). Furthermore, many individual studies in the social sciences, cross-cultural psychology, and health research rely on back translation as a test for translation quality. The various guidelines and studies testify not only to the fact that back translation has become an integral part of translation assessment in many fields but also that there is no back translation method as such. To the contrary, there is quite some variation in how

Contact Dorothée Behr: dorothee.behr@gesis.org

back translation is implemented in actual practice. That is, it may be implemented at different stages in the overall translation process; with different numbers of back translators; with back translators of different backgrounds and skillsets; with evaluators – i.e. those who compare the back translation to the original version – of different backgrounds and skillsets; and with different instructions (if any) supplied to back translators and evaluators. Different implementations of the method are likely to lead to different results as to what back translation may contribute to overall translation quality. Regardless of its actual implementation, the following arguments in favor and against back translation are put forward in the research literature.

2. Arguments pro and contra back translation

The procedure has originally been conceived for situations where a researcher does not speak the target language but wants to make sure that respondents are indeed asked the right questions (Harkness, 2003). In situations such as these, back translations are meant to provide direct control over the quality of the translation (Brislin, 1986). The content-related benefit of the method is often worded quite vaguely. It is argued that differences or discrepancies that are detected between the original version and the back translation may point to problems in the actual translation. The actual translation would then need to be re-assessed. Beaton et al. (2000, p. 3188) restrict these detection possibilities to ‘gross inconsistencies or conceptual errors’; others do not further specify the kind of discrepancies that may get detected.

Arguments against back translation or at least method limitations are more diverse (e.g. Brislin, 1970; Harkness, 2003; Leplège & Verdier, 1995; Swaine-Verdier, Doward, Hagell, Thorsen, & McKenna, 2004). It is argued that the general notion of a forward translation being flawed and a back translation being flawless is inappropriate: Discrepancies when comparing original and back translation can be due to errors in the actual translation but they can also be due to errors in the back translation. Furthermore, the general conception of back translation assessment (BTA), which is based on a close relationship between the original and the back translation version, is inappropriate if cultural adaptation – i.e. intentional modifications to items going beyond translation – is the only means to produce a valid measure for a new language and culture. In addition, if back translation is a part of the assessment structure this may foster a too close or literal forward translation, which reduces the instrument quality in many cases. Equally important, back translators may make sense of a poorly worded translation and thus render the whole exercise of detecting target text problems futile. Also, forward and back translators may share the same set of translation rules for a given language pair (word A always becoming word B always becoming word A) but this may not be suitable in the given measurement context. Furthermore, the actual translation may retain many of the grammatical features of the source questionnaire. It may thus be comparatively easy to back translate. A back translation may then closely resemble the original version and this would lead to the potentially false conclusion that the actual translation is appropriate. However, the actual translation could lack naturalness, be difficult to comprehend or may simply be downright wrong. Also spelling errors, punctuation mistakes or missed diacritical marks, such as accents, will not be caught in a back translation. Those issues can only be covered by additional commenting or by an additional copy-editing step of the actual translation itself. Furthermore, back translations increase the costs of the whole process of transferring an instrument from one language and culture into another language and culture. In addition, if several forward and back translations are produced and subjected to a comparison process, this process becomes ‘unmanageable due to the sheer volume of text generated’ (Swaine-Verdier et al., 2004, p. S27). Eventually, any discrepancies found between original and back translation automatically mean that native speakers of the target language will have to be consulted. Back translation therefore is not a way to substitute for translators working on the actual translation itself.

3. Research gap

While major survey programs and centers, such as the European Social Survey or the U.S. Census Bureau, have replaced back translation altogether by a multi-step team-based forward translation approach, others have come to argue for the need of empirical evidence before either maintaining or abolishing back translation as a good practice (McKenna & Doward, 2005; pp. 89, 90). Thus, notably in health research, experiments are increasingly been conducted whereby questionnaires produced according to different translation and assessment methods are assessed both in terms of their psychometric properties and of respondent preferences. Even though it needs to be acknowledged that the methodological set-ups of these studies differ in more than just the inclusion of back translation or not, first results suggest that at least in terms of preference of the target group the version assessed by back translation falls behind other methods (e.g. Hagell, Hedin, Meads, Nyberg, & McKenna, 2010). This article adds to the small body of empirical research on back translation by providing answers to two different kinds of questions: First, do different evaluators of original and back translated versions come to similar conclusions on the quality of the actual translation? Second, do evaluations of the back translation and evaluations of the actual translation produce different results?

4. Data and methods

The author of this article was granted access to the translation documentation of the European Quality of Life Survey (EQLS) 2012 by the European Foundation for the Improvement of Living and Working Conditions (Eurofound), a Dublin-based European Union Agency.[1] The translation process is described in detail in the translation report of the study (Eurofound, n.d.). The process includes for each EQLS language version parallel translation by two translators, reconciliation by a third person, back translation, comparison of back translation to the original version, native speaker assessment of the post-back translation version, piloting, and finalization. The different steps were thoroughly recorded in an Excel file. This article focuses on the translation of the EQLS into German for Austria and Germany due to German being the author's mother tongue. For parts of the analyses, the French translation for France and Belgium are also considered. Since in EQLS 2012 only modified or new items were newly translated and assessed using the back translation method, only these items were included in this analysis.[2]

5. Results: comparing the back translation to the original questionnaire

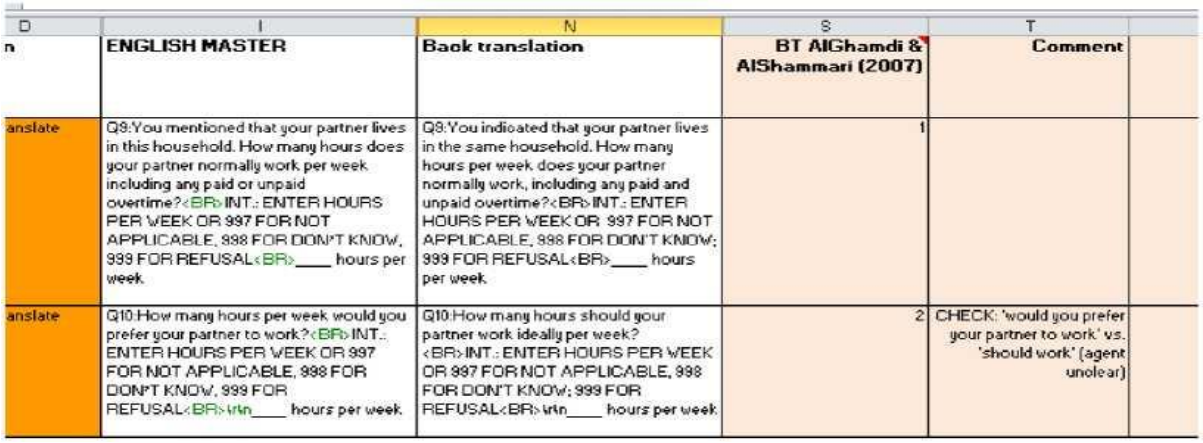
The first analysis step consisted of looking into the reliability of BTA by analyzing what different evaluators produce when confronted with the same back translation and original. The database consists of the following elements: First, the feedback by the *official* back translation evaluators was subjected to the analysis; these were the evaluators commissioned by Eurofound (#1 in Table 1). Second, the author of this article, a researcher focusing on questionnaire translation, compared the back translation to the original for both German and French language versions (#2 in Table 1). Third, four survey researchers at GESIS (#3 in Table 1) conducted each yet another comparison of back translation and original questionnaire. These survey researchers included three cognitive interviewing researchers and one survey researcher focusing on cross-national migration research. Each of these four persons received a briefing and written instructions on how to go about the comparison and each one made a back translation comparison for one language version.[3] Thus, in the end, there were three independent back translation evaluations available for each language version: German (Austria), German (Germany), French (Belgium), and French (France). In comparing the results of three independently conducted back translation comparisons for each language version, the consistency of a back translation comparison was assessed.

The comparison was done in Excel (see Figure 1): a given back translated segment, e.g. a question, an interviewer instruction, or an answer category, was compared to its counterpart in the original

Table 1. Issues flagged through back translation assessment, across languages and across evaluators.

	German (AT)			German (DE)			French (BE)			French (FR)		
Evaluators	#1	#2	#3	#1	#2	#3	#1	#2	#3	#1	#2	#3
Codes												
1	156	134	138	149	124	120	168	150	143	170	136	102
2	–	16	10	–	13	22	–	9	10	–	22	50
3	3	4	5	5	9	4	6	14	15	5	16	14
4	–	5	6	–	8	8	–	1	6	–	1	9
Sum of codes 2 and 3	3	20	15	5	22	26	6	23	25	5	38	64
Total sum of codes	159	159	159	154	154	154	174	174	174	175	175	175

Notes: All comments in comparison #1, which was not based on a coding scheme, were assigned to code 3 for ease of comparison across evaluators. Code 1: satisfactory agreement between back translation and original; Code 2 = almost satisfactory agreement but one or two words uncertain; Code 3 = doubtful translation; Code 4 = see comment above in excel file.



D	I	N	S	T	
n	ENGLISH MASTER	Back translation	BT AlGhamdi & AlShammari (2007)	Comment	
anslate	Q9:You mentioned that your partner lives in this household. How many hours does your partner normally work per week including any paid or unpaid overtime? INT.: ENTER HOURS PER WEEK OR 997 FOR NOT APPLICABLE, 998 FOR DON'T KNOW, 999 FOR REFUSAL ____ hours per week	Q9:You indicated that your partner lives in the same household. How many hours per week does your partner normally work, including any paid and unpaid overtime? INT.: ENTER HOURS PER WEEK OR 997 FOR NOT APPLICABLE, 998 FOR DON'T KNOW; 999 FOR REFUSAL ____ hours per week	1		
anslate	Q10:How many hours per week would you prefer your partner to work? INT.: ENTER HOURS PER WEEK OR 997 FOR NOT APPLICABLE, 998 FOR DONPT KNOW, 999 FOR REFUSAL \r\n____ hours per week	Q10:How many hours should your partner work ideally per week? INT.: ENTER HOURS PER WEEK OR 997 FOR NOT APPLICABLE, 998 FOR DON'T KNOW; 999 FOR REFUSAL \r\n____ hours per week	2	CHECK: 'would you prefer your partner to work' vs. 'should work' (agent unclear)	

Figure 1. Back translation comparison.

questionnaire, that is, to the 'English master.' The *official* evaluators for each language version used either an 'OK' in case of no problem identified, or a comment referring to a problem identified. The evaluations by the author of this article and the GESIS survey researchers were based on a coding scheme by AlGhamdi and AlShammari (2007), which included the codes 1 = satisfactory agreement between back translation and original, 2 = almost satisfactory agreement but one or two words uncertain, and 3 = doubtful translation. A fourth code, that is, 4 = see comment above, was added for the purpose of this study to cater for issues that came up repeatedly.[4] In case of codes 2, 3, and 4, a brief comment was required to explain the issue(s).

Table 1 shows to what extent the different evaluators flagged discrepancies. All problems identified by the *official* back translation evaluators (#1), who did not use a fixed coding scheme, were assigned to code 3 (doubtful translation) for ease of comparison in this study. In addition, since the other evaluators (#2 and #3) noted that it was not always easy to clearly distinguish between codes 2 (almost satisfactory agreement but one or two words uncertain) and 3 (doubtful translation), it seemed wise to look at 2 and 3 together when addressing problematic issues. Remarkable is the consistently higher number of potentially problematic cases flagged by evaluators #2 and #3 compared to evaluators #1. This may be due to the different backgrounds of the evaluators and also to different approaches to the comparison task. In addition, the common coding schemes used by evaluators #2 and #3 may have

had an influence in this regard, especially the 'caution code' 2. All in all, these first results show the subjectivity of the back translation comparison task.

When looking in more detail at the issues flagged, we learn that there is not necessarily convergence between the evaluators. Issues assigned to code 2 or 3 by one evaluator did not always come up as code 2 or 3 by the other evaluators. Some examples in this regard shall now be presented:

5.1. Example for code 2/3 shared by all evaluators

All three back translation evaluators noted that 'in the last 12 months' was included in a given source item but not in its back translation. The time frame was indeed missing in the actual translation.

5.2. Example for code 2/3 shared by two evaluators

'I am a citizen of other EU member state' became 'I'm a citizen of an EU member state' in the back translation. Two evaluators pointed to the lack of 'other' in the back translation, which was indeed also missing in the actual translation.

5.3. Example for problem code 2/3 by only one evaluator

Only one evaluator stumbled over the back translation 'I can accrue hours, in order to file flexidays' because of the unusual verb 'accrue.' The source version used the verb 'accumulate.' The actual translation behind 'accrue' was fine ('ansammeln') and did not come across as odd to native speakers of the respective language.

In a nutshell, sometimes all three back translation evaluators agreed on a problem, sometimes only two, and sometimes none. For this reason, it seemed advisable to look at the agreement rate between back translation evaluators. The first column in Table 2 for each language version pertains to the agreement for evaluators #2 and #3 (those that used the same coding scheme and had at least similar numbers of issues identified). The second column shows the agreement rate when all three evaluators were considered. 'Agreement' means that the evaluators in the respective comparison groups were consistent in not signaling any problem (code 1) or that they *all* signaled that there potentially was a problem with the translation segment (codes 2, 3, or 4). 'Disagreement' means that the outcome of the evaluation differed. For instance, one evaluator suggested that all was fine (code 1) while another flagged a discrepancy that he or she found worth following up on (codes 2, 3, or 4). While the evaluators #2 and #3 differed in 10–27%[5] of cases, the disagreement rate approximately doubled when a third back translation comparison, the official comparison #1, was considered in addition. This increase is particularly due to the official back translation evaluator hardly pointing out any issues and the translation evaluators employed for this study pointing out both uncertainties (code 2) and clearly doubtful translations (code 3). However, the disagreement rate is likely to be higher. Evaluators may have had different reasons for assigning a code, such as 2, to a given translation segment. The issues that have resulted in a code were not considered in the agreement/disagreement rate.

Table 2. agreement rates between back translation evaluators.

Comparison evaluators	German (AT)		German (DE)		French (BE)		French (FR)	
	#2, #3	All three	#2, #3	All three	#2, #3	All three	#2, #3	All three
Agreement	143 (90%)	129 (81%)	136 (88%)	113 (73%)	154 (89%)	139 (80%)	127 (73%)	97 (55%)
Disagree-ment	16 (10%)	30 (19%)	18 (12%)	41 (27%)	20 (11%)	35 (20%)	48 (27%)	78 (45%)
Total	159	159	154	154	174	174	175	175

Note: The total in each column is the number of back translation segments that were compared to the original version.

6. Results: comparing the outcome from the back translation comparison to assessments of the actual translation

While the above analysis 'merely' shows that different people react differently when they compare the back translation version to the original, the following analysis looks at the quality of the assessment method. The second analysis step thus consists of comparing the problems identified by back translation comparison to problems identified by native speaker assessment of the actual translation. The native speaker assessment came from two sources. First, Eurofound members of staff assessed the actual translation after the completion of the back translation process (the step is called 'validation' in their translation report, Eurofound, n.y.). Since problems identified by the *official* back translation had in most cases led to the overwriting of the erroneous translation, the additional validation step could 'only' help to spot mistakes or problems beyond what the *official* back translation had found. Second, the author of this paper, a German native speaker, conducted an additional native speaker assessment of the two German-language versions. The results of all three back translation comparisons *taken together* (as presented above) and the results of both actual translation assessments *taken together* were compared in the context of this study, using the categories as shown in Table 3. Please note that this comparison only concerns the German-language versions from Austria and Germany, not the French version included in the previous stage of analysis.

6.1. Category 1

In more than half of the cases for both German-language versions (AT and DE), back translation comparison and actual translation assessment converged, meaning that they both signaled that the actual translation seemed to be fine.

6.2. Category 2

In about 10% of cases, both types of assessment, i.e. back translation comparison and actual translation assessment, identified a problematic issue. It needs to be noted, though, that only the additional BTAs by the author of this article and GESIS colleagues contributed to category 2. The issues falling under category 2 had gone unnoticed in the *official* back translation comparison. An example of category 2 is the noun 'organizations' as in 'Please look carefully at the list of organisations and tell us, how often did you do unpaid voluntary work through the following organisations in the last 12 months?' 'Organizations' was translated as 'soziale Einrichtungen' in the actual translation, which then became 'social establishments' in the back translation. The German term did not fit to the corresponding list of organizations, since 'political parties, trade unions,' among the organizations referred to, are not typically subsumed under 'soziale Einrichtungen.' Another example for category 2 is 'campaigning,' as

Table 3. Categories for comparing results from back translation and actual translation assessment for German language versions (AT/DE).

Category	Category description	German (AT)	German (DE)
1	Both forms of assessments, i.e. back translation assessment and assessment of the actual translation, signaled that all was fine	102 (56%)	97 (53%)
2	Both forms of assessment correctly identified a problem	14 (8%)	17 (9%)
3	Back translation assessment correctly identified a problem that was not uncovered by the actual translation assessment	1 (2%)	4 (2%)
4	Back translation identified a discrepancy - the issue was not pointed out by any of the actual translation assessors and was usually a false alarm	14 (8%)	23 (13%)
5	The actual translation assessment correctly identified a problem, while the back translation assessment did not point it out	40 (22%)	31 (17%)
6	Repetition of above coding due to reoccurring issue	11 (6%)	10 (5%)
Total		182	182

Note: The total refers to the number of codes assigned.

occurring in the answer category ‘Social movements (e.g. environmental, human rights), charities (e.g. fundraising, campaigning).’ ‘Campaigning’ was translated as ‘Wahlkampf,’ and returned as ‘election campaigning’ in the back translation. This was a clear-cut error identified both by *additional* back translation evaluators and native speakers of the actual translation.

6.3. Category 3

In rare cases, the back translation identified issues that the actual translation assessment failed to see, such as missing brackets around an answer category. Category 3 essentially meant that the BTAs conducted by the author of this article and a third GESIS researcher spotted a problem that was not spotted by any of the actual translation assessors.

6.4. Category 4

Category 4 can be seen as false positives: More or less 10% were discrepancies that back translation evaluators flagged wrongly. An example is the response category ‘not applicable,’ which became ‘nicht zutreffend’ in the actual translation, which then became ‘not relevant’ in the back translation. The discrepancy between the English terms ‘not applicable’ vs. ‘not relevant’ was noted by back translation evaluators; the German translation was deemed fine, though, and received no comment from native speakers. In another example of category 4, ‘participate in social activities of a *club*, society or an association not related to your work’ (original) was contrasted with ‘participation in the social activities of an *association*, a collective or a federation (without any connection to the professional activity)’ (back translation). The discrepancy between ‘club’ and ‘association’ was questioned by one back translation evaluator, but it was also argued by another back translation evaluator that ‘club,’ ‘association,’ etc. can be translated in numerous ways and that the right translation in context can best be assessed in the original language version. The German translation of ‘club,’ which was ‘Verein,’ was considered to be correct.

A careful note seems apt. Category 4 was assigned when the back translation comparison spotted a discrepancy that was *not* spotted by the existing actual translation assessments and that also did *not* cause a problem upon as second assessment by the author of this article. While most of the category 4 cases were clear false positives, a few could be debatable and could lead to different judgments, depending on the inclusion of further persons and supplemental information on the measurement goal. One such example is the following: Two back translation evaluators flagged the discrepancy between ‘care services’ (original) and ‘care facilities’ (back translation) and ‘long-term care services’ (original) and ‘long-term care facilities’ (back translation), respectively. It was not fully clear whether the relevant questions should tap all kinds of services, including the more mobile ones that come into people’s homes, or whether they should tap institution-based services only. In the German language at least, depending on which translation you choose (‘Einrichtungen’ – facility) or (‘Dienste’ – ‘services’), you can make one reading more salient to respondents than the other. Due to the lack of knowledge of the measurement goal, the discrepancy that had been noted was assigned to category 4 rather than category 3.

6.5. Category 5

Category 5 issues can be regarded as false negatives in relation to the back translation: Around 20% of issues were discovered by assessment of the actual translation alone. As Tables 4 and 5 further below show, among those were many issues regarding meaning. One reason for back translation failing to uncover these issues is that back translation can in fact conceal problems. For instance, ‘I have felt downhearted and depressed’ became ‘… habe ich mich niedergeschlagen und depressive gefühlt,’ which then became ‘I’ve felt low and depressed.’ Only through assessment of the actual translation did it become visible that the German term for ‘depressed’ (‘depressiv’) was not suitable due to it carrying

Table 4. Break down of issues noted for German (Austria), according to categories.

	German (Austria)				
	Category 2	Category 3	Category 4	Category 5	Total
Meaning	5	0	8	9	22
Missing information	2	0	1	1	4
Added information	0	0	0	0	0
Consistency	0	0	0	4	4
Text/questionnaire conventions	3	1	2	2	8
Register/wording	2	0	3	7	12
Grammar/syntax	0	0	0	8	8
Spelling	0	0	0	7	7
Layout (incl. brackets)	2	0	0	2	4
Total	14	1	14	40	

Notes: Category 2: Back translation assessment (BTA) and actual translation assessment correctly identified a problem; Category 3: BTA alone correctly identified a problem; Category 4: BTA identified a discrepancy, which was not noted during actual translation assessment and usually a false alarm; Category 5: actual transaltion assessment correctly identified a problem, while BTA did not point out anything.

Table 5. Break down of issues noted for German (Germany), according to categories.

	German (Germany)				
	Category 2	Category 3	Category 4	Category 5	Total
Meaning	10	2	15	8	35
Missing information	0	0	0	0	0
Added information	1	0	1	0	2
Consistency	0	0	1	3	4
Text/questionnaire conventions	4	0	5	0	9
Register/wording	1	0	1	8	10
Grammar/syntax	0	0	0	7	7
Spelling	0	0	0	4	4
Layout (incl. brackets)	1	2	0	1	4
Total	17	4	23	31	

Notes: Category 2: Back translation assessment (BTA) and actual translation assessment correctly identified a problem; Category 3: BTA alone correctly identified a problem; Category 4: BTA identified a discrepancy, which was not noted during actual translation assessment and usually a false alarm; Category 5: actual translation assessment correctly identified a problem, while BTA did not point out anything.

the clinical meaning of depression. Interestingly, the Austrian-German back translator rendered the German word ‘depressiv’ as ‘depressive,’ which is why it was caught by at least two back translation evaluators. Another illustration of concealment is the term ‘care services,’ as used in ‘Could you please tell me for each of the following care services if you or someone close to you have used it or would have liked to use it in the last 12 months?’ The two ‘care services’ subsequently referred to in the questionnaire were ‘child care services’ and ‘long-term care services.’ ‘Care services’ was translated into German as ‘Pflegedienste,’ which was returned as ‘care services.’ Thus, for back translation evaluators at least, all seemed fine. However, while in English one general term for both services is appropriate, in German you will have to be more careful. Depending on how you translate the ‘care’ bit, you may focus on caring for the ill and the elderly or you may focus more in general on attending to someone, regardless of ill health. The German term ‘Pflegedienst’ for ‘care services’ did not fit the questionnaire context since it is only used in the context of the ill and/or the elderly and is thus not fitting to general child care services.

Tables 4 and 5 categorize the issues, or errors, noted according to the following dimensions: meaning, missing information, added information, consistency, text/questionnaire conventions, register/wording (improved style, flow of item), grammar/syntax, spelling, and layout. Both tables show that back translation certainly can pick up larger meaning issues, but back translation is equally successful in suggesting meaning problems without those being true (category 4) or in concealing meaning

problems that indeed exist (category 5). Not surprisingly, actual translation assessment picks up register/wording, grammar, and spelling issues, while these issues remain unnoticed in BTA (category 5).

7. Discussion

7.1. Back translation comparison is open to interpretation

Several aspects influencing the process of back translation comparison have emerged. Most importantly, comparing a back translation to an original version is open to interpretation. As the study has shown, there is quite some variation as to whether discrepancies between the original and the back translation draw attention. Several factors seem to play a role in the decision-making process when comparing the back translation to the original version – the factors are similar to factors influencing good questionnaire translation in general.

First, the background and skillset of back translation evaluators influence the outcome. For instance, the cognitive interviewing researchers in this study, all cognizant of questionnaire design principles, made various comments in view of the technical aspects of a questionnaire, such as non-response categories or scale labels potentially not being translated correctly, or the seeming lack of balancing in the question stem. The migration researcher, on the other hand, flagged a discrepancy regarding 'people from other countries coming here to live and work' (original version) vs. 'foreigners who come to live and work here' (back translation). The researcher noted: 'In many languages "foreigner" has a negative connotation and is therefore avoided (as in the original) […].' These examples serve to illustrate that it is relevant knowledge of questionnaire design, of measurement properties, and of the subject matter that help to spot potential problems – whether these problems are indeed errors in the actual translation or merely produced by an inaccurate or otherwise dissimilar back translation is another issue.

Second, briefing on what to look out for in BTA is likely to affect the outcome. Briefing can point out to issues that deserve particular attention – and potentially to issues that can be neglected when doing the comparison task. Some of the deviations between the official back translation evaluators and the back translation evaluators for this study is likely to be attributable to the instructions (or lack thereof) given to the back translation evaluators.

Third, the language competencies of the back translation evaluator seem to have an effect. The better the back translation evaluator knows the target language, the more this knowledge can be brought to bear on the back translation comparison. For instance, if one is aware that the target language is somewhat more complex, one may be more accepting complexity in the back translation and not report any discrepancy.[6]

Finally, general knowledge of what translation is all about and what types of changes are typically involved when doing a translation seems to influence the process. For instance, the more one expects identity of syntactical or grammatical features in a translation, the more one will possibly flag issues. Or vice versa, the more one is aware of syntactical, grammatical or other differences between languages, the more one will possibly ignore these discrepancies in the comparison task.

These conclusions suggest that a BTA, especially if implemented by unsuitable personnel, can lull researchers into a false sense of security, with a serious impact on comparability. However, even if the right personnel are entrusted with the task, BTA fails to meet the requirements of a comprehensive assessment tool, as summarized below.

7.2. Back translation is not a substitute for thorough assessment of the actual translation

The comparison between BTA and assessment of the actual translation produces the following key results: Back translation can successfully identify errors (codes 2 and 3); however, most of these issues were identified by actual translation assessment as well. In addition, the high number of issues *only* spotted as part of the actual translation assessment (such as an additional 'validation' step) makes it

unmistakably clear that back translation should not be regarded as a substitute for a thorough review of the actual translation. This may not be a surprising result and, in fact, repeats Brislin's conclusions from 1970, namely that back translation should not be regarded as the ultimate test of a questionnaire translation. However, in many studies back translation is still regarded as a (sole) guarantee for translation quality. This study shows the opposite and points to the need to thoroughly review and verify the actual translation rather than its back translation and to employ people experienced in questionnaire translation and suitable for the task in other respects, to perform this assessment.

Having said this, one still needs to go back to the translation steps prior to back translation and validation and see what could be improved here. After all, the errors were not introduced through back translation but they were produced and neglected at the earlier translation steps. The EQLS by Eurofound is no exception in this regard, quite a number of major international studies struggle with quality issues at the very first process steps so that more guidance and better monitoring and/or additional (external) validation steps are increasingly seen as necessary. The EQLS translation process included parallel translation and reconciliation. Reconciliation according to good practice (Harkness, 2003) is a team-based activity where the translators as well as survey and subject matter experts come together to produce a reconciled version. The translation approach chosen for the EQLS included parallel translation, but reconciliation was done by a single person. Thus, there was no discussion, no sharing of expertise, etc. Here there is certainly room for improvement. In fact, more recent Eurofound surveys have adopted a more team-based approach in this respect (Curtarelli & van Houten, 2013).

7.3. Limitations

This study has limitations in that the author of this article compared back translation versions to the original and also actual translations to the original. These different types of assessments may have affected each other. Care was taken, however, to have a time lag between BTA and assessment of the actual translations.[7] In addition, the inclusion of several people in the back translation evaluation steps should have increased the validity of the assessment.

8. Conclusions and outlook

In this study, no assessment was made of the extent to which the errors identified were likely to seriously affect the quality of measurement; however, the fact that errors can be overlooked by back translation, or spurious issues thrown up, should be a cause for concern. This also applies to the fact that different back translation evaluators bring up different issues and lack consistency in their results. This study has provided empirical evidence that back translation is not a guarantee that the actual translation is equivalent or linguistically appropriate. The research community needs to rethink their approach towards translation: A methods description along the lines of 'We translated and back translated the questionnaire to check for equivalence,' which is all too common, should not be regarded as sufficient evidence of a flawless and equivalent translation. Efforts should be directed towards ensuring quality *in* the translation itself – by committee or team approaches; by the involvement of suitable translation, content, and survey experts; and by thorough documentation of the translation process, including problems and intentional deviations from a source questionnaire. These steps would only be one side of the coin, though. Equivalent translations in a cross-national study require as a first step a source questionnaire that is suitable and possibly annotated for translation and cross-cultural transfer – which calls for cross-cultural cooperation during the development of the questionnaire (Smith, 2003). Furthermore, translated versions of questionnaires need to be as thoroughly tested as any other questionnaire in order to assure comprehensibilty and, in the case of cross-cultural studies, equivalence (Braun & Johnson, 2010; Harkness, Pennell, & Schoua-Glusberg, 2004). Thus, the actual production of translated text is not sufficient in order to ensure equivalence, but it is a necessary precondition for all other steps that are to follow. Many good translation processes and methods are

available nowadays, and yet the research community still has a lot to learn about the actual impact of differences in a translation.

Notes

1. http://www.eurofound.europa.eu/about/index.htm, http://www.eurofound.europa.eu/surveys/european-quality-of-life-surveys.
2. The modified questions pertaining to current and past occupation were omitted from this analysis since mere jobs lists assessed by means of a back translation turned out be a very difficult task. One back translation evaluator also noted: 'Personally, I think that the benefit of back translations in the case of occupational descriptions is doubtful. In this case it is normally not a mere translation of the given job description that is needed but the correct term used for an equivalent position in another economical (and linguistic) system. Therefore, it is likely that the back translation of such a "translation" (rather adaption) would more often than not result in the use of a more or less different term than in the original translation.'
3. A key instruction was: 'If there is anything that strikes you (meaning, style, design principles, consistency, etc.), this should be recorded by using the coding scheme and adding a comment on the issue (if possible in English). In "real life" these issues would go back to the translators/a project manager and they would be asked to reconsider the original translation. So, anything where you think one should go back to the original translation should be flagged.'
4. If several different issues were noted for a given translation segment, the following applied for reasons of simplicity: codes 2 and 4 became code 2; 3 and 4 became code 3; 2 and 2 became code 2; 3 and 3 became code 3. This ensured that problematic translation segments were flagged. The details behind a code were looked at in the subsequent analysis step, as presented in a later section of this article.
5. 27% for French (FR) is due to one evaluator using code 2 to a great extent. This evaluator compared the two English versions on a broader basis, e.g. also in the sense of 'English master is easier.'
6. In practice, back translation is typically done because someone interested in the actual translation does not speak the language of the actual translation. Back translation is then the means to grant access to the actual translation.
7. An additional check revealed that there was hardly any overlap in codes between back translation comparison and actual translation assessment.

Acknowledgements

The author would like to thank Eurofound for their openness towards research, their transparency, and their readiness to provide her with the entire EQLS translation documentation. Furthermore, the author would like to thank the four GESIS researchers who each made one back translation comparison to the original version.

Disclosure statement

No potential conflict of interest was reported by the author.

Notes on contributor

Dorothée Behr is a senior researcher at the department of Survey Design and Methodology at GESIS - Leibniz Institute for the Social Sciences in Germany. Her research interests include questionnaire translation, comparability in cross-national survey research, and web probing.

References

Acquadro, C., Conway, K., Hareendran, A., & Aaronson, N. (2008). Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value in Health, 11*, 509–521.

AlGhamdi, K. M., & AlShammari, S. A. (2007). Arabic version of Skindex-16: Translation and cultural adaptation, with assessment of reliability and validity. *International Journal of Dermatology, 46*, 247–252.

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

Braun, M., & Johnson, T. P. (2010). An illustrative review of techniques for detecting inequivalences. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph Mohler, B.-E. Pennel, & T. W. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 375–393). Hoboken, NJ: Wiley-Blackwell.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185–216.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Beverly Hills, CA: Sage.

Curtarelli, M., & van Houten, G. (2013). *Questionnaire translation in the 3rd European Company Survey. Conditions conducive for the effective implementation of the TRAPD approach.* Paper presented at the ESRA conference, Ljubljana, Slovenia.

Eurofound. (n.d.). *3rd European Quality of Life Survey translation report*. Retrieved April 22, 2014, from http://www.eurofound.europa.eu/surveys/eqls/2011/questionnaire.htm

Hagell, P., Hedin, P. J., Meads, D. M., Nyberg, L., & McKenna, S. P. (2010). Effects of method of translation of patient-reported health outcome questionnaires: A randomized study of the translation of the Rheumatoid Arthritis Quality of Life (RAQoL) instrument for Sweden. *Value in Health, 13*, 424–430.

Harkness, J. (2003). Questionnaire translation. In J. Harkness, F. J. R. van de Vijver, & P Ph Mohler (Eds.), *Cross-cultural survey methods* (pp. 35–56). Hoboken, NJ: Wiley.

Harkness, J., Pennell, B.-E., & Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In S. Presser, J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, & E. Singer (Eds.), *Methods for testing and evaluating survey questionnaires* (pp. 453–473). Hoboken, NJ: Wiley.

Leplège, A., & Verdier, A. (1995). The adaptation of health status measures: Methodological aspects of the translation procedure. In S. A. Shumaker & R. A. Berzon (Eds.), *The international assessment of health-related quality of life: Theory, translation, measurement and analysis* (pp. 93–101). Oxford: Rapid Communications.

Maneesriwongul, W., & Dixon, J. K. (2004). Instrument translation process: A methods review. *Journal of Advanced Nursing, 48*, 175–186.

McKenna, S. P., & Doward, L. C. (2005). The translation and cultural adaptation of patient-reported outcome measures. *Value in Health, 8*, 89–91.

Smith, T. W. (2003). Developing comparable questions in cross-national surveys. In J. Harkness, F. J. R. van de Vijver, & P. Ph. Mohler (Eds.), Cross-Cultural Survey Methods, (pp. 69–91). Hoboken, NJ: Wiley.

Swaine-Verdier, A., Doward, L. C., Hagell, P., Thorsen, H., & McKenna, S. P. (2004). Adapting quality of life instruments. *Value in Health, 7*(Suppl I), S27–S30.

Wild, D., Grove, A., Martin, M., Eremenco, S., McElroy, S., Verjee-Lorenz, A., et al. (2005). Principles of good practice for the translation and cultural adaptation process for patient-reported outcomes (PRO) measures: Report of the ISPOR task force for translation and cultural adaptation. *Value in Health, 8*, 94–104.