A Versatile tool? Applying the Cross-national Error Source Typology (CNEST) to triangulated pre-test data

Rory Fitzgerald[1], Lizzy Winstone[2] and Yvette Prestage[3]

Centre for Comparative Social Surveys, City University London[4]

Acknowledgements

The authors would like to thank Sally Widdop for her helpful comments on earlier drafts of this paper. In addition we would like to thank the Russian ESS National Coordinator, Anna Andreenkova and her colleagues and the team at TNS BMRB UK for their pilot fieldwork and reports. All errors of interpretation of their work and omissions are entirely the responsibility of the authors.

[1] e-mail: r.fitzgerald@city.ac.uk
[2] e-mail: lizzy.winstone.1@city.ac.uk
[3] e-mail: yvette.prestage.1@city.ac.uk
[4] Corresponding author: Rory Fitzgerald, Centre for Comparative Social Surveys, City University London, UK

Abstract

There are certain error sources that are unique to cross-national questionnaires, or occur less frequently in single nation studies. Tools that help to identify these errors and separate them from measurement errors that only occur in single nation studies assist the cross-national survey researcher in producing a higher quality source questionnaire. In turn, this supports translators in producing functionally equivalent translations that work well in the target languages and cultures. The Cross-national Error Source Typology (CNEST) was developed as a tool for improving the effectiveness of cross-national questionnaire design and has already proved useful when applied to cognitive interview data. This paper assesses the consistency and versatility of the tool by applying it to triangulated cross-national pre-test data collected in Russia and the UK as part of the development of questions for the European Social Survey (ESS). The benefits and challenges of triangulating pre-test data in a cross-national setting are also highlighted and discussed.

1. Introduction

Compared to single nation, single language studies, cross-national and cross-cultural survey methodology with regards questionnaire design and pre-testing is still relatively underdeveloped (Harkness, 2010). There is a debate as to whether cross-national researchers should employ the best practice developed for single national studies, employ and adapt it, or even develop a completely new methodology (Harkness, 2008; Jowell, Kaase, Fitzgerald and Eva, 2007).

In recent years there have been important improvements in many areas of cross-national survey methodology (Mohler and Johnson, 2010). Yet one of the most central tasks, questionnaire design, has received less attention. Questionnaire design is at the very heart of the social survey, its measurement aims embodying the rationale for the scientific enquiry. Without an effective questionnaire that can generate reliable and valid estimates to meet the measurement aims, it is impossible to draw firm conclusions from the data gathered.

Questionnaire design is a complex task and poses a series of challenges. For the researcher designing a single country, single language study, there is a wealth of guidance on how to create and pre-test a questionnaire (Converse and Presser 1986; Fowler 1995; Presser, Rothgeb, Couper, Lessler, Martin, Martin and Singer, 2004). However, even though such resources are available, there is a clear view that each new questionnaire poses bespoke challenges, and thorough pre-testing is therefore always recommended (Sudman and Bradburn, 1982).

There is an acknowledgement too that cross-national questionnaire design is significantly more challenging than for a single nation situation (Smith, 2004). At the same time, the

literature is far less developed. The lack of documentation about the questionnaire design process in many large scale cross-national surveys hampers the development of best practice and prevents the building of knowledge about designing questionnaires on specific topics (Mohler, Pennell and Hubbard 2008). While many large scale cross-national surveys provide some information about the pre-testing they undertake, this tends to be a descriptive summary of the method rather than full information about the findings themselves (Harkness, 2008). Recent efforts at increasing transparency regarding the questionnaire design for each round of the European Social Survey (ESS), using a questionnaire design template, are going some way to address this (http://www.europeansocialsurvey.org/methodology/questionnaire/rotating_questionnaire.html), as are efforts to develop a cognitive interviewing databank in the United States (Miller, 2006). The ESS is a biennial cross-national social survey, established in 2001, which employs a high quality comparative methodology (Jowell, et al., 2007; Fitzgerald and Jowell, 2010). The provision of the detailed design documentation on the ESS has facilitated this paper since it is possible to reconstruct the development of different concepts and individual question items.

The ESS adopts a parallel 'ask the same questions (ASQ)' approach, allowing only necessary or essential adaptation where it is required in order to realise a meaningful question in the target language (Harkness, 2008). The source questionnaire is produced in a single language and finalized before other translated versions are produced. Significant cross-cultural input is incorporated during the design process, including at the conceptual, drafting and testing phases (ibid). The source questionnaire has a dual function. It must first work as a field ready questionnaire in the source language and culture. At the same time, it must serve as the template for supporting directly comparable measurement in other languages and cultures through effective translation. This dual role places unique demands on the cross-national

questionnaire designer. Of course, even questionnaire designers working on a single nation, single language study must facilitate comparative measurement between different groups in the target population. It is necessary, for example, to ensure that respondents with different educational levels understand the questions in the same way (Jowell et al., 2007). However, the cross-national researcher must also try and ensure shared understanding across multiple languages and across varying national contexts (Smith, 2004). Without a source questionnaire that is effective in both its roles, even the best translation procedures will be unlikely to facilitate comparable, high quality target language versions.

A critical stage in the development of any source questionnaire is the pre-testing employed to assess its effectiveness (Presser et al., 2004). Pre-testing of questionnaires designed for cross-national implementation needs to explore the effectiveness of the instrument in the source language as well as in other target languages and cultures. Inevitably, on surveys fielded across large numbers of countries, such testing is often restricted to a sample of countries due to cost and logistical limitations (Smith, 2004). This pre-testing allows the reliability of the source questionnaire in the language and culture in which it was designed to be tested. It also enables the researcher to assess the effectiveness of the source question in generating equivalent measurement tools in other languages and cultures. As such, pre-tests are likely to raise further issues to those that emerge during a single nation, single language pre-test. Cross-national researchers need additional skills and tools to deal with these extra factors. This paper examines one such tool, the Cross-national Error Source Typology (CNEST).

2. Cross-national pre-test finding typologies

In recent years there has been parallel but independent work to develop typologies to assist researchers in interpreting the findings from cross-national cognitive interviewing (Levin,

Willis, Forsyth, Norberg, Kudela, Stark and Thompson, 2009; Forsyth, Kudela, Levin, Lawrence and Willis, 2007; Kudela, Levin, Tseng, Hum, Lee, Wong, McNutt and Lawrence, 2004; Willis, Lawrence, Kudela and Levin, 2005a; Willis, Lawrence, Thompson, Kudela, Levin and Miller, 2005b; Carrasco, 2003; Schoua-Glusberg, 2006; Goerman and Caspar, 2007). Such typologies help to disentangle distinct problems found with cross-national questionnaires from those that would only apply to single nation, single language instruments. By identifying these different error sources, the typologies facilitate development of an appropriate solution.

Typologies developed by Willis (Willis and Zahnd, 2007; Willis, Lawrence, Hartman, Kudela, Levin and Forsyth, 2008) and Fitzgerald, Widdop, Gray and Collins (2009; 2011) are very similar, with differences limited largely to terminology and some further detail in respect of translation in the typology from Fitzgerald et al. (Willis, 2009). These two typologies have independent development histories, with the Willis typology arising during the analysis of cognitive and behaviour coding projects (Willis et al., 2005a; Willis et al., 2005b), whilst the Fitzgerald et al. (2011) typology was based on experience of designing and analysing the ESS, drawing on evidence of poor quality questions, the results of the translation process and findings from Multi-Trait Multi-Method (MTMM) experiments.

The Cross-national Error Source Typology (CNEST) (Fitzgerald et al., 2011) has already proved useful and comprehensive when applied to cognitive interviewing findings. Most similar typologies have also been tested on cognitive interviewing data (Carrasco, 2003; Schoua-Glusberg, 2006, Goerman and Caspar, 2007). This paper goes beyond qualitative interviewing data to evaluate efforts to apply the CNEST to triangulated qualitative and quantitative pre-test findings from the development of questionnaire modules for Round 6 of

the ESS. In addition to testing the typology, the advantages and challenges of applying triangulated pre-testing data cross-nationally are highlighted and discussed.

3. The Cross-national Error Source Typology (CNEST)

The background to and development of the CNEST is described in detail in a paper by Fitzgerald et al. (2011). In essence, the authors of the CNEST aimed to map all the sources of error that can occur when fielding a survey question cross-nationally in a summative typology. Drawing on evidence from early rounds of developing the ESS the CNEST includes three major sources of error (see Table 1). The first category is '**Poor source question design**', where the question in the source language and culture has problems that are likely to lead to poor measurement quality. A question using an agree / disagree scale, for example, could be described as having this error source because the scale is known to be problematic regardless of where it is used (Fowler, 1995; Saris and Gallhofer, 2007). The second category refers to '**Translation problems**'. Here, an important distinction is made between cases where translators have made a simple error (e.g. translating 'healthy' instead of 'wealthy') or where a sub optimal choice is made (e.g. using a very formal and therefore unfamiliar term, when a less formal term is available), and those where the design of the source questionnaire makes a functionally equivalent translation difficult (e.g. using a word that has no clear equivalent and therefore requires a lengthy or complicated explanation in the translated question). The final category is '**cultural portability**'. In some cases, a concept exists in certain countries but not others; making a question about this futile cross-nationally (e.g. asking about 'General Practitioners', doctors who practice general medicine in the UK, when there is no clear functional equivalent in certain European countries). Alternatively, a concept might exist but in such a different form as to make the use of the same source question impractical (e.g. highest level of educational attainment), requiring instead

consideration of an 'ask different questions' (ADQ) approach to measurement (Harkness, 2008).

Table 1 The Cross National Error Source Typology (CNEST)[5]

		Error found in:	
Error classification	**Description**	**Source language testing**	**Non source language testing**
1) Poor source question design	All or part of the source question has been poorly designed, resulting in measurement error	Always	1 or more countries
2) Translation problems…	Errors occur in translation, resulting in a loss of functional equivalence		
(a) resulting from translator error	Errors stem from the translation process (i.e. a translator making a mistake or selecting an inappropriate word or phrase) rather than from features of the source question that make translation difficult	Never	1 or more countries
(b) resulting from source question design	Features of the source question, such as use of vague quantifiers to describe answer scale points, are difficult / impossible to translate in a way that preserves functional equivalence	Occasionally	1 or more countries
3) Cultural portability	The concept being measured does not exist in all countries. Or the concept exists but in a form that prevents the proposed measurement approach from being used (i.e. you can't simply write a better question or improve the translation). For example, to measure religiosity a different question might be needed in a Christian country compared to a Muslim one.	Less likely*	1 or more countries

[5] Table Reproduced from Fitzgerald et al. (2011).

Note: *Cultural portability problems should be less likely in the source country (language). This is because the question designers should have a greater familiarity with this culture. However, this is not always the case and is complicated further by within-country diversity in cultural practices.

4. Background to the application of the CNEST to triangulated pre-testing data from the ESS

This paper discusses the application of the CNEST to pre-test data obtained during the development of two modules included in Round 6 of the ESS, namely 'Measuring personal and social well-being' (Huppert, Marks, Siegrist, Vazquez and Vittersø 2010) and 'Europeans' understanding and evaluations of democracy' (Kriesi, Morlino, Magalhães, Alonso and Ferrin, 2010).

In Round 6, efforts were made to ensure multi-national input at key stages, as a broad mix of countries can increase the diversity of cross-cultural insights (see Table 2). This input included design by a multi-national questionnaire design team, multi-national expert review groups, cognitive interviewing in five countries, initial quantitative piloting of specific items in three countries, advance translation in three countries and a large scale two nation quantitative pilot.

Recent efforts to improve and evaluate questionnaire pre-testing have noted the benefits of triangulating different methods to complement the insights gained from each (see for example, Padilla, Benítez and Castillo, 2013; Reeve, Willis, Shariff-Marco, Breen, Williams, Gee, Alegría. Takeuchi, Stapleton and Levin, 2011; Thrasher, Quah, Dominick, Borland, Driezen, Awang, Omar, Hosking, Sirirassamee and Boado, 2011; Yan, Kreuter and Tourangeau, 2012). Denzin (1970; p.297) describes triangulation as "the combination of

methodologies in the study of the same phenomena" and also notes its use as "strategy for resolving the inherent biases in one measurement technique" (Denzin, 1970; p.x). This is the case in the examples explored in this paper where the results from various pre-testing sources were considered simultaneously. This paper describes and evaluates the application of CNEST to those questions included in a two nation pilot conducted in Russia and the UK. Quantitative pilot findings were triangulated (i.e. considered in combination) with feedback from the pilot survey agencies (including feedback from the Russian translation process) and respondent debriefs (Andreenkova, 2011; Sullivan, Hamlyn and Hanson, 2011). Feedback from advance translation and National Coordinators was also considered.

The pilots were conducted using demographically controlled quota sampling to be largely reflective of the demographic profile of each country and the achieved samples in both countries were around 400 cases. Although Russia used Paper and Pencil Interviewing (PAPI) and the UK Computer Assisted Personal Interviewing (CAPI), few mode differences were anticipated. In both countries all questions were administered by an interviewer and respondents had the same stimulus - the full question read out verbatim with a showcard to read if applicable at that question.

The questionnaire for the Russian pilot was translated using the ESS committee approach (Harkness, 2007). Interviewers in the UK and Russia were debriefed by survey agency researchers and a small number of respondents in both countries were 'debriefed' using a semi-structured questionnaire to follow-up on pre-specified questions thought to be potentially problematic by the research team.

The advance translation exercise (in which translators are asked to comment on the translation process) in Turkey and the Czech Republic was based on the ESS committee

approach, with a less thorough ‘light’ translation implemented in Germany (Dorer 2012; see also Dorer, 2011 for a background to advance translation).

Finally, comments from ESS National Coordinators (NCs) based in each of the countries where the questionnaire was to ultimately be fielded were sought at a plenary meeting and subsequently in writing. NCs commented on questions that had been amended after the pilot phase. In addition to providing general expert review, they were also asked to identify any translation or cultural barriers specific to their country.

Information from all of these sources was then considered simultaneously, depending on which tests were available for each question, with it presented ‘side by side’ in the questionnaire design template. Essentially a ‘committee approach’ was then taken to deciding on how to interpret any reinforcing or conflicting evidence.

Table 2 ESS Round 6 Questionnaire Development and Pre-testing Schedule

Stage	Process		Description
1	Proposals for new question modules, identifying key concepts, definitions and measurement aims	Iterative	Question design template
2	Proposals reviewed by multi-disciplinary specialist panel		Expert review
3	Survey Quality Predictor Program (SQP) – usually used once		Program used to predict reliability and validity of new items
4	Cognitive Interviewing		In Austria, Bulgaria, Israel, Portugal and the UK
5	Quantitative pre-testing using commercial omnibus surveys		In the UK, Portugal and Hungary
6	Revised proposals submitted in light of stages 2 to 5		Revised question design template submitted
7	ESS National Coordinators consulted on substantive and translation issues		Comments fed into process via email and face-to-face meeting
8	Split ballot MTMM experiments developed		Tests of alternative question wording
9	Advance Translation		In Czech Republic, Turkey and Germany
10	Large-scale, two-nation quantitative pilot run containing MTMM experiments		In the UK and Russia
11	Analysis of pilot data - including examination of item non-response, scalability, factor structure, correlations, analysis of the MTMM experiments and assessment of translation		Conducted by question designers and members of the ESS Core Scientific Team (CST)
12	Further specialist review of the proposed questions in light of stage 11		Expert review
13	Further consultation with the National Coordinators		Comments fed into process via email and face-to-face meeting
14	Final source questionnaire is produced and translated according to a committee approach following the ESS committee translation procedures		Source questionnaire finalised in British English then translated into target languages

5. Examples of the application of the CNEST to triangulated ESS pre-test data

The previous application of the CNEST to cognitive interviewing data has shown that the typology can comprehensively categorise all problems identified, including those instances where multiple problem types have occurred at the same question (Fitzgerald et al., 2011).

Section 5 gives examples of where triangulating and pretesting evidence leads to a clear categorisation of error and can point to a possible solution. More complex examples of classification, including multiple error types and less straightforward classifications, are given in Section 6.

Table 3 Examples of CNEST application to quantitative and qualitative pre-test findings

Error type	Question wording	Quantitative pre-test findings	Qualitative pre-test findings
5.1 Poor source question design	How difficult or easy do you think it is for immigrants* to get the right to vote in national elections in [country]? Use this card where 0 is far too difficult and 10 is far too easy. (0 = far too difficult; 10 = far too easy) **Translation Annotation: People who come to live in [country] from another country*	Mean scores: Russia 2.98; UK 6.75. Very high item non-response: Russia 33%; UK 23%. High use of mid-point: Russia: 9%; UK 19% - suggesting some UK respondents chose the mid-point rather than saying 'don't know' (see Kalton, Roberts and Holt, 1980). Findings suggest that many respondents in both countries had difficulty answering the question.	Respondent debrief: there were respondents in both countries who interpreted this as a 'knowledge' question, referring to their own lack of knowledge about the relevant legal situation for immigrants. Others mentioned a lack of relevant experience (i.e. they were not immigrants, therefore could not judge whether getting the 'right to vote' was easy). There were also respondents from both countries who felt the question referred to the ease of obtaining citizenship' rather than the right to vote.

Error type	Question wording	Quantitative pre-test findings	Qualitative pre-test findings
5.2 Translation problems resulting from translation error	There are differing opinions on whether or not everyone should be free to express their political views openly in a democracy, even if they are extreme. Which **one** of the statements on this card describes what you think should happen in a democracy? In a democracy: Everyone should be free to express their political views openly, even if they are extreme / Those who hold extreme political views should not be free to express them openly / (Neither of these / it depends)	Much higher item non-response in Russia than UK: Russia 11% neither of these/it depends + 10% don't know; UK 6% neither of these/it depends + 2% don't know.	Feedback from Russian fieldwork report: a direct translation of 'be free to express' ('иметь свободу публично выражать') was used, which the report said in hindsight made the question 'sound unusual' in Russian. The Russian fieldwork report suggested a better translation of 'be free to express' would have been (roughly back translated) 'have the right to express' ('иметь право выражать'). Had this been used, the question would, the report suggested, have been much clearer for respondents and still equivalent to the source question. Whilst there was also a source question problem related to the term 'extreme views' at this item, which probably contributed to the item non-response, it was thought that the sub-optimal translation had significantly increased levels of 'don't know' responses in Russia compared to the UK.

Error type	Question wording	Quantitative pre-test findings	Qualitative pre-test findings
5.3 Translation problems resulting from source questionnaire design	And on how many of these days were you physically active for 20 minutes or longer in a way that made you breathe somewhat harder than normal? (WRITE IN)[6]	Much higher item non-response in Russia (14%) than UK (1%).	Feedback from Russian fieldwork report: the structure of the source question was difficult to reproduce in the Russian language, leading to a question that was confusing for respondents. For example interviewers reported having to repeat the question several times. In the UK there were no reported difficulties with this question, suggesting that the high item non-response in Russia was related to translation problems which had resulted in a long and complex question that was trying to replicate the source stimuli. Advance translation and comments from several NCs also led to queries about the meaning of other terms and phrases in the question (such as the scope of 'physical activity'). This suggested that longer, more complicated questions may be required in other target languages, perhaps leading to uneven question quality cross-nationally.

[6] Preceding question: "Using this card, please tell me on how many of the last 7 days you were physically active for 20 minutes or longer?"

Error type	Question wording	Quantitative pre-test findings	Qualitative pre-test findings
5.4 Cultural portability	When governments and public opinion in [country] disagree on what is best for the country, do you think governments change their policies or plans too rarely or too often? (0 = far too rarely; 10 = far too often)	Pilot data show relatively low item non-response in the UK (5%) but much higher levels in Russia (14%). The data are well distributed in both countries, although notably use of the mid-point was high in both the UK (20%) and Russia (19%).	While the UK fieldwork report did not highlight any issues with this question, feedback from the Russian fieldwork report suggests that respondents found the question particularly difficult to answer. This was due to a perceived lack of transparency in Russian government decision making, which made it difficult for many respondents to evaluate the decision making process and decide how often Russian governments change their plans. This problem is intrinsically linked to respondents perceptions of the transparency of their country's political system, suggesting difficulties in transporting this item into contexts where there is little, if any, openness regarding political decision making.

5.1 Poor source question design

"Source question problems may emerge if all or part of the source question has been poorly designed, resulting in measurement error" (Fitzgerald et al., 2011; p574).

It appears that the first question (Table 3, error type 5.1) was asking some respondents about an issue they simply do not know about. By triangulating this quantitative and qualitative pre-test data there was sufficient evidence to suggest that there was a 'source question problem'. Best practice in questionnaire design suggests that respondents should be asked about topics where they can reasonably be expected to provide an answer (Fowler, 1995). For a substantial number of respondents who gave a 'don't know' or mid-point response, this topic may simply have been too opaque.

The respondent debrief helped to explain the high item non-response and use of the mid-point found in the pilot data, which indicated that these two issues were likely to be a serious problem in both countries. Having only the respondent debrief information would not have allowed the researchers to be sure of the likely extent of the item non-response that would be found; having only the pilot data would have meant the researchers would have been unclear as to the reasons why so many respondents did not answer. Having both sources therefore provided greater certainty as to the problem.

It was decided to drop this item from the module entirely, as no solution to the lack of knowledge on this topic could be found.

5.2 Translation problems resulting from translation error

Translation errors occur when questions in the target language are not functionally equivalent to the source questions they are expected to mirror. This sometimes results from human error or from the fact that the translated phrase or word used is sub-optimal. In both cases, this can result in a lack of functional equivalence. Whilst there were no examples of simple translator human error in the ESS Round 6 testing, the adoption of a sub-optimal translation was discovered. More than one error source was identified at the second example question given in Table 3, but the focus here is on translation (error type 5.2). Combining information from the Russian fieldwork report with the quantitative data provided a compelling case for concluding there had been a translation error.

The ESS uses annotations to provide additional information for translators in cases where part of the question needs clarification including where part of the question wording has connotations in the source language that might be less obvious to a non-native speaker. For example, 'household' might be annotated as 'all those who share a living space' so that it is clear that 'the family' is not intended (Harkness, 2007). In the case of this item, 'free to' was annotated, as 'are allowed to' in the final questionnaire, to try and avoid ambiguity and assist translators in finding an optimal translation.

5.3 Translation problems resulting from source questionnaire design

"When this type of problem is discovered it suggests that although the question could function reasonably well in the source (and possibly some target) language(s), there is something about it that makes translation particularly difficult" (Fitzgerald et al., 2011: 574).

The feedback from the Russian pilot agency, the advance translators and NCs clearly suggests there may be difficulties with the third example in Table 3 (error type 5.3) that stem from trying to reproduce the source question stimulus in target languages. Triangulating this with evidence of non-response from the Russian pilot helps to quantify the possible impact this may have on the data.

It was decided that the two questions measuring physical activity should be combined in a simplified, single item. By adding the term 'continuously' to the question, it was felt that this single item would be sufficient to capture 'moderate physical activity'[7].

5.4 Cultural Portability

"This CNEST error source applies when there are cultural barriers to equivalent measurement. Maybe the concept being measured does not exist at all in some of the countries where testing is taking place. Or the concept does exist but in a form that prevents the proposed measurement approach from being used (i.e. you can't simply write a better question or improve the translation)" (Fitzgerald et al., 2011: 570).

There is a debate in the cross-national questionnaire design field about the extent to which all questions should be equally applicable or relevant in a cross-national study. Smith (2004) discusses this issue and highlights an example in an International Social Survey Programme (ISSP) module. In this study, 'car use' questions were included even though car ownership was very low in some participating countries. A similar challenge is posed for the fourth example in Table 3 (error type 5.4) by the potential irrelevance of evaluating government

[7] Final question wording: Using this card, please tell me on how many of the last 7 days you were physically active continuously for 20 minutes or longer? **INTERVIEWER NOTE:** include household tasks such as housework or gardening if mentioned, as long as performed for 20 minutes or longer.

decision-making in the Russian context, when the process is arguably less visible than in other countries.

The high item non-response figures from the pilot in Russia suggest a problem with the item in Russia. Triangulating this with the feedback from the survey agency highlighted that this arose due to Russian respondents perceptions of their country's political system – highlighting the cultural portability issues with certain items from the democracy module which were conceived more with the European democratic model in mind. The concept measured by this question ('responsiveness to citizens') was felt to be a core aspect of democracy from a theoretical perspective. Despite the difficulties, it was therefore agreed this item could not be dropped from the module entirely. Two separate 'evaluation' questions were included instead. Respondents were asked tailored questions[8] depending on their preference for government responsiveness to citizens in an ideal democracy. It was hoped this would make the questioning a little less 'alien' for respondents in countries where government transparency was limited. However, it was acknowledged that the question would be likely to continue to cause some difficulty in Russia and possibly also in some other countries.

6. Complex examples of the application of the CNEST to triangulated pre-test data

The examples shown above have been presented to demonstrate examples of a particular error, even if more than one error was found. In addition, classification was relatively straightforward for these examples. There are often multiple error sources and classification can sometimes prove to be difficult (Fitzgerald et al., 2011). In this section, more complex

[8] Final question wording: "Using this card, please tell me how often you think the government in [country where interview taking place] today [changes/sticks to] its planned policies [in response to/regardless of] what most people think?"

cases are outlined, including cases where triangulation of different sources provided conflicting perspectives.

6.1 Two Errors: Poor source question design and Translation problem resulting from source questionnaire design

Question wording: To what extent do you think governments in [country] take into account the demands of minority groups as well as following the demands of the majority?
(0 = Not at all – 10 = Completely)

Pilot data showed higher item non-response in Russia (12%) than in the UK (3%). Use of the mid-point was higher in the UK (17%) compared to Russia (10%). Feedback from the Russian fieldwork agency identified significant problems with the meaning of this question, and in particular the ability to translate ‘demands’ in a functionally equivalent way. Many alternatives were suggested by the Russian translators before an appropriate term was agreed. This ambiguity in the source questionnaire design made an effective, clear, equivalent translation difficult. Whilst the concept of ‘group demands’ is fairly commonly used in the British English context, this is perhaps a culturally specific term that might not be immediately apparent to translators. It was therefore agreed this was probably a translation problem resulting from the design of the source question.

Respondent debriefs in both the UK and Russia suggested there were also problems with the terms 'minority' and 'majority'. Their meanings were not immediately clear to respondents, and there was great variety in the interpretation of these terms. The French NC also commented on the possible confusion with a ‘parliamentary majority’. Such specific interpretations are likely to be problematic when the question refers to all minorities (not only some) and ‘the majority’ is not intended to refer to ‘the government’. Despite the lack of

obvious quantitative impact, this was clearly a source question problem affecting respondents in the UK and Russia and would probably impact respondents in other countries too.
The item was dropped from the module due to the lack of clarity in the source question. The terms 'minority' and 'majority' were considered too broad in this context, and the cognitive task too challenging.

6.2 A 'Non' Error

This next example demonstrates how triangulation can sometimes lead to contradictory results.

Question wording: Using the same card please tell me how important you think it is for a democracy that governments are voted out of office when they do a bad job?
(0 = Not at all important – 10 = Extremely important)

The Russian pilot report suggested that this item (measuring 'retrospective accountability') posed a cultural portability problem as it asked about an alien democratic mechanism. . The report noted that as Russian governments are appointed and dismissed by the President, the concept of 'voting governments out of office' is unfamiliar (Andreenkova, 2011). The Prime Minister for example is appointed by the President and then approved by the state parliament (White, 2011). However, there was low item non-response in Russia (less than 4%) as well as in the UK (less than 2%), suggesting that there were no obvious comprehension issues in Russia. In addition, at a related item most Russian respondents answered (appropriately for their context) that governments in Russia are unlikely to be 'voted out of office if they do a bad job'. The mean score in Russia at this related item was 3.11, compared to 6.11 in the UK, with higher scores reflecting higher perceived likeliness of a government being voted out of

office when they do a bad job. Despite the suggestion in the Russian fieldwork report that a cultural portability error was present, the quantitative data (from the two related items) suggest that the concept was reasonably well understood in Russia.

Triangulation here posed a challenge, since the quantitative data was in some senses at odds with the expert opinion of the field agency report. However, combining data from a related question with this one provided reasonable evidence to assert that the questions could be fielded without serious comprehension issues resulting from cultural portability. Unfortunately, no respondent debrief information was available to help shed further light on this matter. In the end, a judgement had to be made, which is so often the case with questionnaire design when dealing with cognitive understanding.

Although the pilot data showed no particular problems, additional comments received from National Coordinators in Belgium and Germany later in the design process led to a decision to change this item. The NCs identified translation difficulties with 'voted out' (translation error) and that the question suggested that governments are voted out immediately or automatically when they do a bad job, rather than at elections (source questionnaire error).

To address these issues, and reflect more closely the concept description of governments being punished or rewarded in elections, the item was changed to: And still thinking generally rather than about [country], how important do you think it is for democracy in general that governing parties are punished in elections* when they have done a bad job? (0 = Not at all important for democracy in general – 10 = Extremely important for democracy in general)

* Translation annotation: Punished in elections' in the sense of 'getting fewer votes than in the previous election'

6.3 An Unusual 'Cultural Portability' Problem

Question wording: Countries differ in whether their governments are generally formed by a single party or by two or more parties. Do you think governments in [country] are formed by two or more parties too rarely or too often?
(0 = Far too rarely – 10 = Far too often)

Pilot data show moderate item non-response in the UK (9%) with high levels in Russia (16%). Use of the mid-point was also high in both countries, with a quarter of respondents in the UK and a fifth in Russia choosing the mid-point from the scale.

In Russia, information from the respondent debrief suggested that the Russian political system made this question particularly difficult to understand. In Russia, individual members of the government are appointed by the President on the advice of the Prime Minister rather than being formed by parties (White, 2011), which makes a question about 'parties forming governments' somewhat odd. This perhaps contributed to the high item non-response in Russia.

There is further evidence of a potential lack of cultural portability from the comments provided by the Swiss National Coordinator. They suggested that there would be comprehension difficulties at this question because in Switzerland governments are only ever formed by multiple parties (i.e. 'several' rather than 'two or more'). So although the question might make sense in an abstract way, it is so far from the reality that in Switzerland that it would probably create problems.

Could there be a cultural applicability problem that accounted for the quantitative findings in the UK? Unfortunately, nothing from the UK fieldwork report suggests why this question posed a problem for respondents. However, there was clearly a problem with the question in both pilot countries. The researchers concluded that the question may work in a country that alternates, fairly frequently, between single- and multi-party government. In the UK, however, this is extremely rare. Although there was a coalition government in the UK at the time of the ESS pilot fieldwork, this had only been in place for a year or so, preceded by a period of sustained single party governments in the 1980's, 1990's and 2000's. It is possible that UK respondents were lacking in confidence about answering an item that requires a normative assessment of the situation over an (unspecified) period of time. This, in turn perhaps explains the high use of the mid-point in the UK, although there was no corroborating evidence for this from the respondent debrief.

Although cultural portability problems are not usually found in the source language country, this can happen when multinational design teams develop new questions based on theoretical aims and pan-European measurement concerns. This source question therefore may well work in some contexts but pose challenges in many others, including in the source language itself.

The lack of qualitative input from the UK poses a challenge to interpretation. A respondent debrief was not performed on this question and therefore the high mid-point use was not illuminated. However, the high item non-response, high mid-point use, feedback from the Russian pilot and comments from the Swiss National Coordinator (in combination) reinforce

that this question is likely to be more challenging to respondents residing in countries where there is little, if any, alternation between single- and multi-party government.

In the end, a less normative question format than that used in the pilot was adopted. Respondents were first asked what they thought was better for a democracy (single or multi party government). Depending on their answer, they were then asked tailored follow-up questions[9]. The follow-up question was therefore more likely to be contextually relevant to the respondent's own country and refer to an option which they have already indicated they understand. It was hoped this would help, to some extent, with the challenge of cultural portability.

7. Reflections on Triangulation

As noted at the start of this paper, the cross-national researcher has additional issues to address compared to their colleagues working in a single nation context. This is because the source questionnaire has a dual role: to function in the source language and culture, and to support translation into multiple cultures and languages. Harkness (2008) has outlined the benefits of ensuring cross-cultural input into the source questionnaire design process. The examples given in this paper reinforce the benefits of this approach, but particularly highlight the strength of including triangulation as part of this process at the pre-testing stage. Combining simple, quantitative analysis from a relatively large two nation pilot (n = 400 in each country) with feedback from interviewers and respondent debrief feedback from semi-structured interviews conducted immediately after the main pilot survey, assisted the research team in drawing their conclusions, enabling more fully informed decisions to be made during

[9] Final question wording: **Now for the last question on this topic.** Using this card, please tell me how often you think the government in [country] is formed by [a single party/ two or more parties in coalition]? (0 = Never – 10 = Always)

the questionnaire design process. Having only the pilot data and limited interviewer feedback (the staple of the 'traditional' pilot) would have been more restrictive. Conversely, having only feedback from respondents (for example only cognitive interviewing) would have left the researchers unsure of the likely impact of particular problems on the data in the main stage survey.

Combining the pilot information with advance translation and feedback from NCs helped to contextualise the findings beyond the two main test countries. These two methods are significantly cheaper compared to pre-testing methods that require direct contact with respondent, therefore for relatively little additional cost they helped to reassure the researchers about whether the findings were country specific or of more general relevance. Triangulating these methods assisted not only in applying the CNEST, but also in assessing the likely quantitative impact of particular problems. This is important in guiding the researcher in developing an appropriate solution, as well as helping to decide whether taking such action is really necessary. This is helpful since there is always a risk that making changes can introduce new errors that will not be discovered without further testing. It was notable that when less evidence was available (e.g. no respondent debrief) conclusions were sometimes more difficult to draw.

More complex quantitative analysis, such as factor analysis, could be included in future triangulations to see how this impacts on the application of the CNEST. More detailed studies assessing triangulation and the application of the CNEST using a range of advanced statistical techniques would also be welcome.

8. Reflections on applying the CNEST

Willis has argued that the conceptual structure of the CNEST maps the range of issues that can potentially damage a question designed for cross-national administration (Willis, 2009) and an earlier application of the CNEST showed that the typology comprehensively categorised all of the emerging problems in a round of cross-national cognitive interviewing (Fitzgerald, et al., 2011).

However, cognitive interviewing is only a single, limited pre-testing method (Miller, 2005) and if typologies like the CNEST are to have more general relevance in the field, they need to be versatile enough to apply to various combinations of pre-testing techniques. The most common pre-testing method undoubtedly remains the 'traditional' pilot, usually a relatively modest quantitative exercise in two or more countries accompanied by feedback from interviewers and researchers in the test countries.

Data outputs alone would never be enough to allow the application of the CNEST, since some qualitative information is required to contextualise any findings and map them to the error sources. However, the application of the CNEST to the ESS triangulated pilot and pre-test information has again shown that the typology is comprehensive in its ability to map emerging problems with the source questions and their translated counterparts. As in the earlier application of the CNEST (Fitzgerald et al., 2011), there were cases where more than one error source applied to a question. In addition, applying the typology was sometimes time consuming, requiring multiple attempts and occasionally further discussions with researchers in the test countries. However, in each scenario, the CNEST was helpful in pointing towards an appropriate solution, even if this was, on occasion, to drop the measure entirely.

References

Andreenkova, A. (2011). *Pilot of questionnaire for European Social Survey Round 6 in Russia*. Available on request from Centre for Comparative Social Surveys. City University London.

Campanelli, P. (2008). Testing survey questions. In Edith.D. de Leeuw, Joop. J. Hox, and Don. A. Dillman (Eds.), *International Handbook of Survey Methodology* (p. 176-200). New York: Taylor and Francis.

Carrasco, L. (2003). *The American Community Survey (ACS) en Español: Using cognitive interviews to test the functional equivalence of questionnaire translations*. Washington DC: Statistical Research Division, U.S. Census Bureau.

Converse, J. and Presser, S. (1986). *Survey questions: Handcrafting the standardized questionnaire*. Newbury park, CA: Sage.

Denzin, N. K. (1970). *The research act: A theoretical introduction to sociological methods*. Transaction publishers.

Dorer, B. (2011). *Advance translation in the 5th round of the European Social Survey (ESS)*. FORS Working Paper Series, paper 2011–4. Lausanne: FORS.

Dorer, B. (2012). *Report on advance translation (R6)* ESS DACE Deliverable 4.5. Available on request from the authors of this paper, Centre for Comparative Social Surveys, London.

Fitzgerald, R., Widdop, S., Gray, M. and Collins, D. (2009). *Testing for equivalence using cross-national cognitive interviewing*. London: Centre for Comparative Social Surveys, City University, CCSS Working Paper 01.

Fitzgerald, R. and Jowell, R. (2010). Measurement Equivalence in Comparative Surveys: The European Social Survey (ESS) – From Design to Implementation and Beyond. In Janet A. Harkness et al. (2010 (Eds.), *Survey methods in Multinational, Multiregional and Multicultural Contexts* (p. 485-497). New Jersey: Wiley.

Fitzgerald, R., Widdop, S., Gray, M., and Collins, D. (2011). Identifying sources of error in cross-national questionnaires: Application of an error source typology to cognitive interview data. *Journal of Official Statistics*, *27*(4), 569-599.

Forsyth, B. H., Kudela, M. S., Levin, K., Lawrence, D., and Willis, G. B. (2007). Methods for translating an English-language survey questionnaire on tobacco use into Mandarin, Cantonese, Korean, and Vietnamese. *Field Methods*, *19*(3), 264-283.

Fowler, F. J. (1995). *Improving Survey Questions. Design and Evaluation*. Thousands Oaks, CA: Sage.

Goerman, P., and Caspar. R. (2007). A New Methodology for the Cognitive Testing of Translated Materials: Testing the Source Version as a Basis for Comparison. Paper presented

at the American Association for Public Opinion Research conference, 17-20 May 2007, Anaheim, California and submitted to *2007 JSM Proceedings, Statistical Computing Section [CD-ROM]*. Alexandria, VA: American Statistical Association, 3949–3956.

Harkness, Janet A., (2007). Improving the quality of translations. In Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey (p. 79-92)*. London: Sage.

Harkness, Janet A. (2008). Comparative survey research: goal and challenges. In Edith D. De Leeuw, Joop J. Hox, Don A. Dillman (Eds.), *International Handbook of Survey Methodology* (p. 299-316). New York: Taylor and Francis.

Huppert, F., Marks, N., Siegrist, J., Vazquez, C. and Vittersø, J. (2010). *Personal and social wellbeing: Proposal submitted to the ESS Scientific Advisory Board.* Available from www.europeansocialsurvey.org.

Jowell, R., Kaase, M., Fitzgerald, R. and Eva, G. (2007). The European Social Survey as a measurement model. In Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey (p. 1-29)*. London: Sage.

Kalton, G., Roberts, J., and Holt, D. (1980). The effects of offering a middle response option with opinion questions. *The Statistician*, 65-78.

Kudela, M. S., Levin, K., Tseng, M., Hum, M., Lee, S., Wong, C., McNutt, S. and Lawrence, D. (2004). *Tobacco Use Cessation Supplement to the Current Population Survey Chinese, Korean, and Vietnamese Translations: Results of Cognitive Testing. Final Report submitted to the National Cancer Institute*. Rockville, MD: National Cancer Institute.

Kriesi, H., Morlino, L., Magalhães, P., Alonso, S. and Ferrin, M. (2010). *Europeans' understandings and evaluations of democracy, Proposal submitted to the ESS Scientific Advisory Board.* Available from www.europeansocialsurvey.org.

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Kudela, M. S., Stark, D., and Thompson, F. E. (2009, March). Using Cognitive Interviews to Evaluate the Spanish-Language Translation of Dietary Questionnaire. *Survey Research Methods*, 3(1), 13-25.

Miller, K.,, Willis, G., Eason, C., Moses, L. and Canfield, B. (2005). Interpreting the Results of Cross-Cultural Cognitive Interviews: A Mixed-Method approach. In Jürgen Hoffmeyer-Zlotnik and Janet A. Harkness (Eds.). *Methodological Aspects in Cross-national Research*. Spezial Band 11. ZUMA; Mannheim.

Miller, K. (2006). Q-BANK: Development of a Tested-question Database. In *Proceedings of the Section on Government Statistics American Statistical Association* (p. 1352-1359). American Statistical Association.

Mohler, P., Pennell, B., and Hubbard, F. (2008). Survey Documentation: Toward Professional Knowledge management in Sample Surveys'. In Edith. D. De Leeuw, Joop J.

Hox, Don A. Dillman (Eds.), *International Handbook of Survey Methodology* (p. 299-316). New York: Taylor and Francis.

Mohler, P. and Johnson, T. (2010). Equivalence, Comparability, and Methodological Progress. In Janet A Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennel Tom W. Smith (Eds.), *Survey Methods on Multinational, Multiregional, and Multicultural Contexts (p17-29).* New Jersey:, Wiley.

Padilla, J. L., Benítez, I., and Castillo, M. (2013). Obtaining validity evidence by cognitive interviewing to interpret psychometric results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *9*(3), 113-122.

Presser, S., Rothgeb, J., Couper, M. P., Lessler, J.T., Martin, E., Martin, J. and Singer, E. (Eds.), (2004). *Methods for testing and evaluating survey questionnaires*. New Jersey: Wiley.

Reeve, B. B., Willis, G., Shariff-Marco, S. N., Breen, N., Williams, D. R., Gee, G. C., Alegría, M., Takeuchi, D.T., Stapleton, M. and Levin, K. Y. (2011). Comparing cognitive interviewing and psychometric methods to evaluate a racial/ethnic discrimination scale. *Field Methods*, *23*(4), 397-419.

Saris, W. and Gallhofer, I. N. (2007). *Design, evaluation and Analysis of questionnaires for survey research*. NewYork: Wiley.

Schoua-Glusberg, A. (2006). *Eliciting education level in Spanish interviews.* Paper presented to the American Association of Public Opinion Research. Montreal Canada: AAPOR.

Sudman, S., and Bradburn, N. (1982). *Asking questions: a practical guide to questionnaire design*. San Fransisco: Jossey-Bass.

Smith, T. (2004). Developing and evaluating Cross-National Survey Instruments. In Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith T. Lessler, Elizabeth Martin, Jean Martin, Eleanor Singer (Eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken, NJ: John Wiley and Sons.

Sullivan, S., Hamlyn, R., and Hanson, T. (2011). European Social Survey Round 6 Pilot – Report (UK). Available on request from Centre for Comparative Social Surveys. City University London.

Thrasher, J. F., Quah, A. C., Dominick, G., Borland, R., Driezen, P., Awang, R., Omar, M., Hosking, W., Sirirassamee, B. and Boado, M. (2011). Using Cognitive Interviewing and Behavioral Coding to Determine Measurement Equivalence across Linguistic and Cultural Groups An Example from the International Tobacco Control Policy Evaluation Project. *Field Methods*, *23*(4), 439-460.

White, S. (2011). *Understanding Russian Politics*. Cambridge: Cambridge University Press.

Willis, G. B., Lawrence, D., Kudela, M. S., and Levin, K. (2005a). *The use of cognitive interviewing to study cultural variation in survey response.* Paper presented to the Questionnaire Evaluation Standards Workshop (QUEST) on Questionnaire Design and Testing. Heerlen, the Netherlands, 19-21 April, 2005. QUEST.

Willis, G. B., Lawrence, D., Thompson, F. Kudela, M. S., Levin, K., and Miller, K. (2005b). *The use of cognitive interviewing to evaluate translated survey questions: Lessons learned.* Proceedings of the Federal Committee on Statistical Methodology Research Conference, Arlington, VA.

Willis, G. B., and Zahnd, E. (2007). Questionnaire design from a cross-cultural perspective: an empirical investigation of Koreans and non-Koreans. *Journal of Health care for the poor and underserved*, 18: 197-217.

Willis, G. B., Lawrence, D., Hartman, A., Kudela, M. S., Levin, K., and Forsyth, B. (2008). Translation of a tobacco survey into Spanish and Asian languages: The Tobacco Use Supplement to the Current Population Survey, *Nicotine and Tobacco Research*, 20(6):1075-1084.

Willis, G. B. (2009). *What Kinds of Problems Does Cross-Cultural Pretesting Reveal*? Paper at 2009 QUEST Meeting. Available from http://www.quest.ssb.no/meetings/QUEST09/Binder1.pdf

Yan, T., Kreuter, F., and Tourangeau, R. (2010). Evaluating survey questions: a comparison of methods. *Journal of Official Statistics, 28*(4), 503-529.