RECSM

Research and Expertise Centre for Survey Methodology

A procedure to prevent differences in translated survey items using SQP

Diana Zavala-Rojas

RECSM Working Paper Number 38

June 2014

http://www.upf.edu/survey/_pdf/RECSM_wp038.pdf

Diana Zavala-Rojas[1]

Abstract

Survey translation has developed best practice procedures to translate survey instruments aiming that the same stimuli and measurement properties should be provided. However, it is very difficult to check in a systematic way these requirements. Current procedures in translation assessment do not link the quality of the translation with a formal test of measurement equivalence. In addition, monitoring the formal structure of translated questionnaires in cross-sectional surveys is challenging. This paper presents a procedure to prevent differences in the form of translated survey instruments using Survey Quality Prediction program (SQP). SQP asks users to code a large set of properties of a survey item. Deviations in translations are detected by comparing the codes of a source questionnaire and targeted languages. The paper summarizes the findings of this procedure implemented in a set of items in the Round 5 of the European Social Survey (ESS).

The problem of equivalence in survey translation

In the last couple of decades, comparative survey research has become more appealing in the social sciences. Survey methodology has made a distinction between *comparing national surveys* and implementing *comparative surveys from design* (Harkness et al. 2010a). The difference is that a *comparison of national surveys* often involves comparing surveys designed for a specific country, whereas *comparative surveys from design* are surveys thought to have the same procedures and characteristics taking into account the different contexts in all the settings where they are administered with the idea to match the findings in each population of study. In this type of survey research, it is assumed that by trying to keep survey features the same to the maximum extent, the data would remain comparable.

For instance, the objective of the Organisation for Economic Co-operation and Development (OECD) Programme for International Student Assessment (PISA) is to survey how well 15-year old students are prepared regardless the curricula taught in different schools across participating countries (PISA 2009). The tests are designed in such a way that they aim to reflect the differences in the analytical tools of the students and not the cultural context in which schooling education is embedded.

[1] Please send correspondence to: diana.zavala@upf.edu
Research and Expertise Centre for Survey Methodology, Universitat Pompeu Fabra (RECSM-UPF), Barcelona, Spain

In this context, this paper focuses in one aspect of comparative survey research: design of measurement instruments across populations in different languages. A prerequisite for *comparative survey research* is that the measurement instruments administered across populations are in fact *comparable*.

According to Scheuch (1968), this apparently obvious requirement should not be understood in terms of "whether [questions] are identical or equivalent in the commonsense meaning, but whether they are functionally equivalent for the purposes of analysis" (Scheuch 1968: 113). For Mohler and Johnson (2010) Scheuch definition implies that "functionally equivalent indicators are revealed in analysis, they cannot be judged on the basis of face value similarity. (…) they should behave in a similar manner in statistical analysis" (2010: 23). This implies that the responses obtained from questions should represent, across groups, the same concepts they intend to measure (Scheuch 1993, Mohler & Johnson 2010).

For a survey questionnaire, equivalence has two conditions: 1) respondents should understand survey questions in the same way across languages, i.e. they should understand the same concepts of interest asked via questions and 2) they should express themselves in the same way, i.e. a same opinion should correspond to the same observed answer across cultural/linguistic groups (Saris 1988, Saris & Gallhofer 2007a, 2014).

Survey translation has developed best practice procedures to translate a questionnaire to get *functional equivalent* survey instruments in multilingual contexts. Procedures bring together the state of the art in translation studies and the particular needs of survey research. In translation studies, the concept of functional equivalence has been already discussed for a long time. It requires that the message embedded in a text is received by the receptor in the same way as it would be received in the source language (Nida 1964).

Translation guidelines suggest that a good translation aiming functional equivalence would avoid changing deliberately other semantic components than those necessary because of language differences (Harkness 2003, Harkness et al. 2003, Harkness et al. 2010b). This means that a translation should keep the concepts of interest the same across languages, preserve the item characteristics and maintain the intended psychometric properties. But guidelines do not suggest how to formally test that a resulting translation is equivalent. In other words, guidelines and procedures in

translation assessment do not offer methods to test that a same question is measuring a same concept in all contexts where it is being asked.

With some exceptions in the context of translation of psychological instruments and educational testing, there is little research on how to assess statistically cross-cultural instruments before they are administered to respondents (see for reference Brislin 1970, 1976, Hui & Trandis 1985, John et al. 1984, Benet-Martinez & John 1998, Ramirez-Esparza et al. 2006, Dean et al. 2007, Willis & Lessler 1999, PISA 2009).

In practice, it is very difficult to check empirically if requirements set by translation guidelines –to maintain the intended psychometric properties and to keep concepts the same— are achieved because one cannot understand all languages. As Tom Smith (2004: 446) points out "perhaps no aspect of cross-national survey research has been less subjected to systematic, empirical investigation than translation."

Empirical methods are mostly used for detecting flaws once data is already collected. Procedures to check the equivalence of measurement instruments across countries are improving and becoming more sophisticated (Horn & McArdle 1992, Meredith 1993, Steenkampt & Baumgarter 1998, Vandenberg & Lance 2000, Byrne & Van de Vijver 2010, Saris & Gallhofer 2007a, 2014). Their application is increasing rapidly in social sciences (Braun & Johnson 2010, Davidov et al. 2011, Jowell et al. 2007).

This shortage of methods to empirically test questionnaires motivated the research question in this article: *How to detect deviations, in terms of functional equivalence, of a survey instrument in different languages before it is administered to respondents?*

This paper aims to provide an answer to this question, and to do so, it first presents a framework of functional equivalence in cross-cultural research. Secondly, it reviews the literature in survey translation and translation quality assessment to conclude that current procedures do not link translation evaluation to any framework of equivalence. Thirdly, it presents an approach to check for differences in translated survey instruments by comparing item characteristics in a systematic way and it provides the arguments of why this procedure is directly linked to cross-cultural equivalence.

The article is organised as follows, section 1 after this introduction defines measurement equivalence -or measurement invariance- across populations. Section 2 summarises current procedures in the field of survey translation and translation

quality assessment to achieve equivalence and argues that they do not help to formally test equivalence because either they rely on subjective judgements and/or they do not have a direct link with invariance testing.

Section 3 defines the *formal characteristics* of a survey item (domain, concept, wording, response scale, polarity, labelling, symmetry, balance of the request, introduction, instructions, linguistic complexity, etcetera). It is argued that if these characteristics vary across two language versions, it is likely that measurement invariance will not be achieved. The section describes a coding scheme in the Survey Quality Prediction program to collect information in a systematic way about a comprehensive number of characteristics of a survey item.

Section 4 proposes a procedure to compare the formal characteristics of a source and a translated questionnaire as a means to detect deviations in translation before the instrument is administered to respondents.

Section 5 presents the findings of a first implementation of this procedure to compare a group of items of the fifth round of the European Social Survey (ESS) in more than 21 languages. Section 6 concludes and points out some recommendations for future developments on cross-cultural questionnaire design and survey translation.

1. Definition of measurement equivalence

Comparative survey research is more complex than national survey research due to larger conceptual, technical and practical considerations (Armer 1973, Lynn et al. 2006, Jowell 2007). Although each national survey should be implemented on high methodological rigour, efforts should also be focused on preserving measurement features constant across countries.

There is consensus among survey methodologists that *measurement invariance* – or measurement equivalence- is a prerequisite to derive substantive conclusions from data collected in diverse populations. It should not be assumed but tested that survey instruments measure the same constructs in exactly the same way across groups (Jöreskog 1971, Vandenberg & Lance 2000, Harkness et al. 2010, Saris & Gallhofer 2007a, 2014, Davidov et al. 2011, Presser et al. 2004, Jowell et al 2007).

One way to discuss the issue of measurement equivalence is by representing the response function for *n* groups of interest (Equation 1). In this model the observed response, y, collected by asking a survey item in each group, i, is assumed to be a linear function of the opinion (unobserved), x_i, and an error component, ξ_i.

$$y_i = \tau_i + \lambda_i x_i + \xi_i \text{ , } i = 1 \dots n \text{ (Equation 1)}$$

In Figure 1a, it is indicated that the response functions are different for three different linguistic groups A, B and C which have the same opinion on a certain topic but they differ in the intercept, τ_i, and/or in the slope, λ_i, and therefore, they differ in their response, y.

Group A expressed its opinion in a more extreme way, while C did the opposite, and Group B is somewhere in between. The data cannot be interpreted in the same way across the groups because each of them gave different answers for the same opinion. In contrast, Figure 1b shows how the response function looks when it is functionally equivalent for three groups. In this case, the relationship between observed responses and opinions is the same for the 3 groups. The groups are invariant or their answers equivalent because loadings and intercepts are the same across groups.

Figure 1a

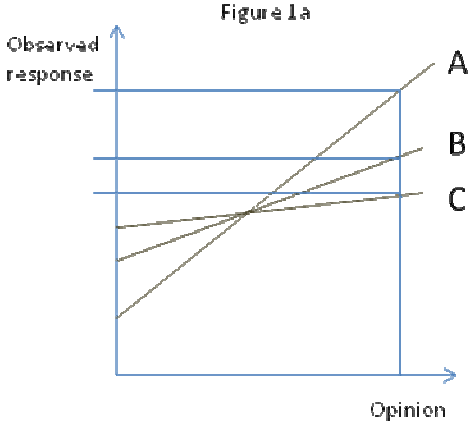
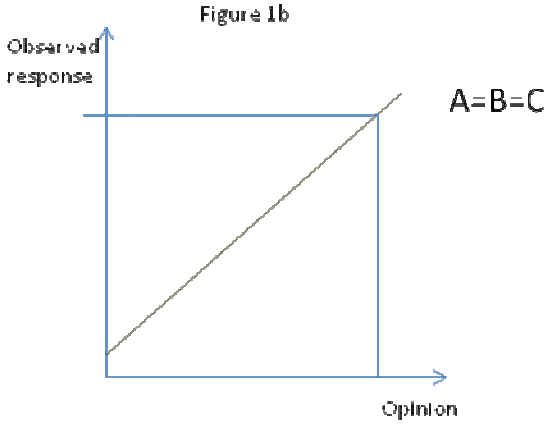


Figure 1b



A test for *measurement invariance* has three steps. As each step is more restrictive, they are called weak, strong and strict equivalence tests (Meredith 1993). In the first, weak invariance, a model called *configural invariance* is tested for all groups (Horn & McArdle 1992). The idea is to check if the configuration of the factorial structure is the same across the groups of interest. In the second step, *metric invariance* –or strong invariance- is tested by restricting the configural model to one where the factor loadings (λ) are invariant across the groups. When the test holds, comparisons of relationships across groups can be done (Horn & McArdle 1992).

The last step, strict invariance, implies that the intercepts (τ) are the same in addition to invariance of the factor loadings. This test is also typically referred as a test of *scalar invariance*. If this model is not rejected, comparisons of means can also be done across groups (Meredith 1993, Steenkamp & Baumgartner 1998, Vandenberg & Lance 2000).

When scalar invariance is not achieved, the responses are said to be affected by *item bias* and/or *method bias* (Van de Vijver & Tanzer 2004). The sources of item bias are many, Van de Vijver and Tanzer suggested that the most frequent causes of item bias are "item translation, ambiguities in the original item, low familiarity/appropriateness of the item content in certain cultures, or influence of cultural specifics such as nuisance factors or connotations associated with the item wording" (2004: 127). Method bias occurs when the observed answers are affected by a factor that is independent of the construct of interest and related to the characteristics of the measurement instrument, e.g. the response scale, layout of batteries, et cetera (Van Herk et al. 2004, Saris & Gallhofer 2007b, Krosnick and Fabrigar, *forthcoming*).

Equivalence in cross-cultural survey research is only confirmed by formally testing invariance. Several studies have identified translation deviations as a source of non-equivalence in assessments of survey data (Hambleton et al. 2005, Harkness et al. 2010a, Mallinckrodt & Wang 2004, Oberski et al. 2007, Saris & Gallhofer 2007a, Van de Vijver & Leung 1997, Villar 2009, Zavala Rojas & Dorer 2013). Unfortunately non-invariance was detected once data was collected and survey organisations had already spent a lot of resources in data collection. A procedure to foresee translation problems that could affect equivalence would have been extremely useful in order to prevent them.

The next section reviews current practices in survey translation and translation quality assessment. It is argued that most of these methods do not help in preventing non-equivalence because they do not have a direct link with measurement invariance and because most of the assessment requires the judgement of evaluators. Judgements may be subjective or in some other cases evaluators may focus on just a set of elements to judge an item (Saris 2012). The final decision about the appropriateness of a translation relies on one person or a team of experts, but not on model-based (or empirical) evidence.

2. Measurement equivalence and translation

2.1 Cross-cultural survey translation

The most widely used approach in questionnaire design for multiple cultures is frequently referred as the "Ask-the-same-question" model (Harkness 2003). In this model, questionnaires developed in a specific population are adopted and exported to other settings via questionnaire translation. A measurement instrument is designed in one language and it is called *source* (source language, source questionnaire, source item, source instrument, et cetera). This instrument is translated into other languages also called *target* languages.

The objective of survey translation is not to achieve *literal*, word-by-word translations but a *functional equivalent* formulation. Even if the process is under strict guidance, each translation has specific cultural elements, grammatical characteristics and a subjective inherent component. This is unavoidable and includes not only the target versions but the specific context of the source language in which the instrument is designed.

Scientific debates relevant in the context of translation studies and linguistics have arrived with delay to survey research. Only recently, manuals on cross-cultural surveys methodology have fully incorporated translation procedures (Presser et al. 2004, Harkness 2010a, Survey Research Center 2010). There, it has been suggested that translation is a "methodological landmark" (Mohler & Johnson 2010: 21). An increasing number of research projects look for benchmarks to prevent effects and bias from translation –such as the European Social Survey (ESS), the Survey of Health, Ageing and Retirement in Europe (SHARE) and the PISA- providing comprehensive guidelines on how to do it (Harkness, 2003, 2007, Harkness et al. 2004, 2010a, 2010b).

Although, very few analyse systematically which translation elements or language properties matter in survey questionnaires (Alwin 2007, Schuman & Presser 1981, Hambleton et al. 2005, Oberski et al. 2007, Saris & Gallhofer 2007a, Villar 2009, Harkness et al 1997, Harkness et al 2004, Zavala Rojas & Dorer 2013). Best practice procedures recommend full documentation of the translation process (Harkness 2010, Survey Research Center 2010, Harkness et al. 2004) although there is little insight on the best way to gather and analyze such information (Mohler & Uher 2003, Mohler et al. 2008).

Best practices in survey translation recommend that translation should be integrated as part of questionnaire design rather than implemented after it when the questionnaire in the source language is finished (Harkness et al. 2010, Erkut et al. 1999). This can be done by making available more than one source instrument. For instance, the Eurobarometer survey and PISA study design their source questionnaires in both English and French and they are taken jointly to produce target versions in other languages (European Commission 2013, OECD/PISA 2006, 2009).

Another advisable practice is to define the unit of translation (the survey item), its goal (match intended meaning and intended measurement properties) and its audience (respondents) and solve problems from this perspective rather than centring the discussion at the level of words (Harkness et al. 2010).

More recent research suggests that translations should be done by a team -or committee- approach in a multistep process where different members provide expertise to arrive at a final translation (Schoua-Glusberg 1992, Harkness 2003, 2008). Harkness (2003) suggests a procedure called TRAPD (translation, review, adjudication, pre-testing and documentation) as the most complete method. The ESS uses a typical

composition of the translation team in the TRAPD: it includes two translators, one reviewer and one adjudicator. It should combine people with survey knowledge and linguistic expertise. The two translators make parallel translations from the source version to the country's language. The reviewer assesses the translation and the adjudicator is responsible for the decisions on the different translation options. The whole process is documented and the translated questionnaire is pre-tested.

It is claimed that the in TRAPD procedure quality monitoring is part of the process as changes are approved by a team and documented at each step (Harkness et al. 2003). However, how adjustments are decided remains subjective. Willis et al. (2010) showed in an evaluation of the TRAPD procedure in five projects that the success of it pretty much depended on the team members' familiarity (translators, reviewers, adjudicators, cognitive interviewers) with the purpose of the translation. A second problem in the TRAPD is that the documentation step produces a large amount of information that lacks systematic analysis. For instance, in the ESS the amount of information available varies from country to country because documentation is conducted by the national teams at the country level (Harkness 2007, Dorer 2012). Several scholars have pointed out that the amount of information available in large scale cross-cultural projects is burdensome for the average user (Mohler & Uhler 2003, Mohler et al. 2008).

A complementary procedure to the TRAPD is to conduct *advance translation,* such as it has been recently incorporated in the ESS (Harkness 1998, Dorer 2010). In this approach, a survey questionnaire is translated using the TRAPD approach during the questionnaire design stage into two languages to foresee potential difficulties in the formulations. If advance translation shows challenges in the formulation of the source text, it could be modified to convey that. One limitation of this approach for multilingual studies with more languages than the ones used in advanced translation is that if a problem is detected in one language and this leads to changes in the source questionnaire, another problem could appear in another language that did not participate in advance translation. Questionnaire designers would remain unaware of this second problem.

2.2 Translation assessment

This section introduces current procedures for translation assessment: back translation, translation verification and pretesting. Ideally, methods to assess a translation in survey research should take care to assess whether the target text kept

semantic content and the psychometric properties determined by the item characteristics the same across languages (Harkness 2003, Harkness et al. 2003, Harkness et al. 2010).

A very common possibility to evaluate translated measurement instruments is by *back translation* (Brislin 1970, 1976). In this procedure the target questionnaire is translated back into the source language. Differences between the two texts are rendered as potential translation problems. This approach is necessary in order to make it possible that a translation is understood in the same way by different members involved in the survey design process. However, as an assessment method it is not exempt from limitations. The main criticism to this approach is that the target text is not evaluated but a version of it in the source language. Other criticisms are that translators may use words that make a translation closer to the source but incoherent in the target language, because their own performance is evaluated taking back translation as a standard rather than as a tool. Deviations may not relate to the translations but to unmatched linguistic structures in both languages (de la Puente 2002, Schoua-Glusberg 1997, Harkness 2003, Schoua-Glusberg & Villar 2013).

A recently borrowed method from linguistics and translation studies to survey translation is outsourcing of semantic verification of target instruments. This procedure has been called *translation verification* or *semantic quality control*. It is used in projects such as the PISA, the Trends in International Mathematics and Science Study (TIMMS) and the ESS (cf. OECD/PISA 2006, 2009, IEA/TIMMS 2007, ESS 2013 respectively). An external provider verifies a questionnaire or a selection of items in all participating languages based on categories for potential interventions -among them: additional information, missing information, grammar/syntax, consistency —to recommend changes in a translation when they are considered necessary. Verifiers give suggestions for improving countries' translations and the overall comparability of data; they also check compliance with annotations provided in the source questionnaire to produce more precise translations.

The Translation Expert Task Group of the ESS has reviewed this practice in the translation process of Round 6 questionnaire (Dorer 2013). They found that there are differences between the verifiers' scope across languages. Some of them were more inclined to stylistic interventions while others were more inclined to verify content related to the concepts of the measurement instruments. They suggested improving the categories of intervention and urge verifiers to use them homogeneously. National

coordinators[2] in the ESS Round 5 have also suggested that the usefulness of the interventions was related to the verifiers' knowledge of each country's context. For example, Russian is the language shared by most countries in the ESS. Speaking populations use different forms and words to assign meaning in Russia, Israel, Estonia and Lithuania. Verifiers need to be familiar with several but at the same time proper usages and forms of the language in each country.

A procedure addressed to directly test equivalence of survey instruments is *pretesting*. For instance, the PISA has conducted a pilot study with an average of over 200 student responses for each item in most participating countries in each round. The data was used to eliminate from the main study such items that showed non-equivalence across countries using common differential item functioning (DIF) and item response theory (IRT) techniques (OECD/PISA 2009). This is a costly and time-consuming procedure but it provides the equivalence tests which are necessary in comparative research before the fieldwork of the whole project.

However, pretesting in many survey projects has mostly meant to administer the questionnaire to a small group of respondents before starting fieldwork. This is an important step; it is very useful to detect flaws in routing, layout, comprehension, length, et cetera. A questionnaire must always be read out and answered completely before starting the fieldwork. However as a translation assessment method, it does not allow to draw statistical conclusions about the measurement instrument.

Saris (2012) reviews procedures to evaluate the design of survey questions depending on two criteria: first, if they are based on personal judgements or if they are model based. Secondly, depending if collection of a large amount of data is needed; if only collection of few data or if data collection is not necessary. Among these procedures, a very common pretesting method for multilingual instruments is *cognitive interviewing,* because it claims to test equivalence with a small amount of data. In its typical design, probing questions are used in a face-to-face interview to get information about item comprehension and response formulation, by making respondents think aloud while they answer or/and by making them tell how they arrived to their answer (Harkness 2010, Willis 2004, 2005, Pan et al 2005, Beatty 2004, ESS 2013, US bureau of the Census 1998, Fitzgerald et al. 2011).

[2] National coordinators oversee the whole implementation of the European Social Survey in a specific country.

Pan et al (2005) showed that respondents participating in cognitive interviewing in four different languages had in each group specific patterns of linguistic behaviour and communicative style. This meant that for the same probing questions, respondents differed in the way they answered the cognitive interview and not in the way they understood the survey item that were evaluated. Other criticisms are about the large effects of interviewers (Beatty 2004, Goerman & Caspar 2010), the thresholds for problem acceptance and the reliability of respondents in problem detection (Conrad & Blair 2004).

Psychological research has used for a long time bilingual individuals to test the equivalence of items in two languages (John et. al. 1984). Split ballot experiments designed for bilinguals were a very common procedure. Each random group received the questionnaire in a different language. The reliability of the instrument was assessed considering the differences between the two groups (Kroll & de Groot 2005, Mallinckrodt & Wang 2004, Egisdottir et al. 2007).

Benet-Martinez and John (1998) and John et. al. (1984) used multi-trait multi-method (MTMM) experiments to assess cross-language validity in personality measures because, as long as the repetitions of the same traits in different languages were answered by bilinguals, it was possible to estimate the effects of language differentiating it from other sources of measurement error. But research has showed evidence that bilinguals do not use languages in the same way as monolinguals do (Bond & Yang 1982, Yang & Bond 1980, Blais & Gidengil 1993, Ellis et al. 1989, Harzing 2006, Gibbons et al 1999).

Experiments on bilingual individuals have revealed that answers to instruments measuring personality traits change depending on the cultural frame activated in each language (Xiaohua Chen & Bond 2010, Hong et al. 2000). Responses to attitudes and personality traits have varied depending on how integrated or conflicted are the different cultural schemas in bilinguals (Benet-Martinez &Haritatos 2005). Thus, bilinguals seemed to follow different response patterns in each language depending on how integrated were both cultures in their own identities. Therefore, switching responses because of language is not expected for all kinds of bilinguals (Ramirez-Esparza et al. 2006). These results have decreased the validity of experiments using bilingual individuals as a means to test equivalence of translated instruments.

Dean et al. (2007) have suggested the Question Appraisal System (QAS) as a coding tool for pre-testing cross-cultural instruments. The QAS is defined as a

"taxonomy" of the cognitive demands of a question. It is a coding system based on four cognitive processes for response formation: comprehension, memory retrieval, judgement, and response selection (Tourangeau 1984). Results of the appraisal are used to revise question wording, questionnaire format and question ordering (Lessler & Forsyth 1996). Although the system is useful to detect the complexity of a survey item, this depends on the coders' ability to provide impartial judgements. The assessment includes many subjective categories such as if an item is difficult to read, if there are complicated instructions, or if a respondent is unlikely to know an answer. If coders are used to technical language or are highly educated they could dismiss the complexity of a survey question.

To sum up, the first part of this section elaborated on the problem of equivalence in cross-cultural survey research and how to formally test for measurement invariance. Unfortunately, testing procedures are only available once data has been collected. The second part revised current and best practice procedures in the field of survey translation and translation assessment. These methods have shortfalls when assessing the quality of a translation. Most of them relied on subjective judgements to evaluate a translation and they do not have a direct link with a test of measurement invariance.

Harkness and Schoua-Glusberg (1998) pointed out that assessment in survey translation is challenging because methods do not specify the criteria of assessment i.e. what is assessed and how. Saris (2012) reviews methods to evaluate survey questions and concludes that "all procedures based on personal judgments provide information about the validity, social desirability, and knowledge of the respondents about the issue of the question and much less about the effects of the form of the questions (Saris 2012: 548)."

In other words, procedures lack of systematic assessment and judges look at different elements that matter for comparability, concentrating on content but paying less attention to effects of question wording on equivalence. Pilot studies as a pretesting strategy have a direct link to measurement invariance, but this procedure is not affordable for most surveys because it requires collecting a large amount of new data.

3. Formal characteristics of a survey item

Thanks to many years of research, we know to a large extent, which item characteristics are likely to affect a measurement instrument. They are also known as formal or measurement properties of a survey item. Their effects on measurement have been studied largely in the tradition of questionnaire design. Starting in 1951, Payne's book on the art of the formulation of survey questions already considered the consequences of different question formats and answer scales (Payne 1951).

This tradition evolved and included experimental research to show how responses change between different formulations of a same concept (Schuman & Presser 1981, Bradburn & Sudman 1979), research has also shown an account of the cognitive processes behind a survey response (Tourengeau et al 2000, Sudman et al. 1996, Schwartz & Sudman 1987) and how different properties, for instance qualifiers in answer scales, affect this cognitive process (Krosnick & Fabrigar 1997, Saris 1988). Research has shown that item characteristics –such as layout, question form, response scale, labelling of response options, don't know option, length of the interview, among many others- may increase or decrease item bias and method effects (Költringer 1995, Krosnick & Fabrigar 1997, Saris & Gallhofer 2007a, Saris & Gallhofer 2007b Alwin 2007, Tourangeau et al. 2000).

A related line of research, measurement quality, made it possible to estimate to what extent observed answers change when specific characteristics in a survey item also change and how serious this is in terms of measurement error (Andrews 1984, Költringer 1995, Saris and Andrews 1991, Scherpenzeel 1995, Scherpenzeel and Saris 1997, Alwin 2007).

When survey questions are designed, researchers take decisions of which item features are to be chosen. Saris and Gallhofer "made an inventory" (2007a: 29) of those decisions (over 60). They developed a coding scheme for this inventory to collect comprehensive information about the characteristics of a survey item and use them as predictors for measurement quality –interpreted as the variance of the observed variable explained by the variable of interest (Saris & Gallhofer 2004, 2007). This paper proposes to apply this coding scheme to translation evaluation. If the characteristics of source and target survey items are coded and compared using this scheme, differences in the codes mean that features that determine invariance are different across language versions. This procedure gives a simple way to assess language versions before data collection.

This coding scheme is incorporated in Survey Quality Prediction (SQP) (Saris 2003, Oberski et al. 2005, Oberski et al. 2011, Saris & Gallhofer 2013). SQP is a survey software which takes the characteristics of an item as predictors of measurement quality. The next section summarizes the current inventory of item features in SQP and gives a brief introduction to the program.

3.1 Survey characteristics in SQP

In their inventory, Saris and Gallhofer (2004, 2007a, 2014) have included a comprehensive list of features that scholars in survey methodology have identified as the characteristics that affect a survey item. It is not the objective of this paper to go further on how these characteristics affect survey responses. Specialised literature in this regard is available (cf. Saris & Gallhofer 2007a, Bradburn & Sudman 1979, Krosnick & Fabrigar 1998, Alwin 2007, Tourangeau 2000, Schuman & Presser 1981, Couper 2008, Dillman 2011) and the codebook available at www.sqp.upf.edu provides an in-depth definition of each of survey property included in the program.

Saris and Gallhofer (2007a, 2007b) first divided the classification of item characteristics in two groups: 1) features that are inherent to the topic of interest and cannot be changed by the designer of the questionnaire and 2) the characteristics that are the product of decisions taken by the researcher when is formulated.

Across those two groups, the list of survey characteristics in the coding scheme of SQP can be summarized into eleven subgroups shown in Table 1 below: 1) the characteristics of the trait; 2) the characteristics associated to the trait 3) the characteristics on the formulation of the request for an answer, 4) the characteristics of the response scale; 5) the presence of instructions to respondents or interviewers; 6) the presence of additional information or definitions; 7) the characteristics of the introduction; 8) the linguistic complexity of the request for an answer, the response scale and the introduction; 9) the method of data collection; 10) the language and, 11) the characteristics of the showcards or visual aid. In this way, the classification of survey characteristics in SQP is a comprehensive list of features that matter in order to produce a survey item.

Table 1. Summary of characteristics inventoried by SQP

	Group	Specific characteristic	Level of decision
Group 1	On the trait	Domain	Features that are inherent to the topic of interest and cannot be changed during questionnaire design
		Concept	
	Associated to the trait	social desirability	
Group 2		centrality of the topic	
		time specification	
Group 3	Formulation of the request for an answer	trait requested indirectly, direct or no request and presence of stimulus ((battery)	Features that are decisions taken during questionnaire design
		WH word and what type of WH word	
		Type of the request (interrogative, Imperative question-instruction, declarative or none (batteries).	
		Gradation	
		Balance of request or not	
		Encouragement to answer	
		Emphasis on subjective opinion	
		Information about the opinion of other people	
		Absolute or a comparative judgment	
Group 4	Characteristics of the response scale	Categories; yes/no answer scale; frequencies; magnitude estimation; line production and, more steps procedures.	
	If the selection is "categories":	number of categories	
	Characteristics of labels:	full or partial labels	
		labels in long or short text	
		Order of labels	
		Correspondence between labels and numbers	
		theoretical range of scales (bipolar or unipolar)	
		Range of scales used	
		Fixed reference points	
		Don't know option	
Group 5	Instructions	Respondent instructions	
		Interviewer instructions	
Group 6	Additional information about the topic	Additional definitions, information or motivation	
Group 7	Introduction	Introduction and if request is in the introduction	
Group 8	Linguistic complexity	Number of sentences	
		Number of subordinated clauses	
		Number of words	
		Number of nouns	
		Number of abstract nouns	
		Number of syllables	
Group 9	Method of data collection		
Group 10	Language of the survey		
Group 11	Showcards or visual aid	Categories in horizontal or vertical layout	
		Text is clearly connected to categories or if there is overlap	
		Numbers or letters shown before answer categories	
		Numbers in boxes	
		Start of the response sentence shown on the showcard	
		Question on the showcard	
		Picture provided.	

The characteristics of the trait are four. The 'domain' is determined by the topic of the research. Saris and Gallhofer (2007a, 2014) compiled an extensive list of

domains in survey research. The 'concept' is the abstract subject that the request is intended to get information about. The choice of domain and concept determines 'other associated characteristics' that are coded in SQP such as the presence of 'social desirability', the 'centrality' of the topic in the mind of the respondents, and the 'time specification' of the survey items.

The second group of codes specifies formal characteristics of the request for an answer. The coder gives information on the 'basic choice': if it is a direct or an indirect request or if there is no request (in a battery). It is also coded if there is a 'WH word' and its 'type', if it measures quantity, extremity, intensity, place, time, etcetera. The request for an answer is classified as 'interrogative question', 'imperative question or instruction', 'declarative statement' or 'none of three' (subsequent items of batteries).

Other properties in this group are if 'gradation' is used, if the request is balanced', if there is an 'encouragement to answer', if there is 'emphasis on subjective opinion', if the request contains 'information about the opinion of other people', if it demands an 'absolute' or a 'comparative judgment' and, if it uses 'stimulus or statement' (batteries).

The third group of codes in SQP are the measurement properties of the response scale. The program asks to code which is the 'basic form of the response scale', options are 'categorical' when the number of categories is between 3 and 12; 'yes/no answer scale or dichotomous choice'; 'frequencies', where amounts such as percentages, time, probabilities are requested; 'magnitude estimation', when size of numbers means opinion; 'line drawing' and, 'more steps procedures'.

In each case, the program asks for other specific characteristics. For instance, 'categories' options are 'number of categories', 'use of full' or 'partial labels', 'long' or 'short texted labels', 'order' and 'correspondence between labels and numbers'. The coder also reports if the scale has a 'bipolar' or 'unipolar theoretical range' and 'range used' in the questionnaire. There are codes for the 'number of fixed reference points' and if there is an explicit, implicit or there is no 'don't know option'.

Codes in group 5 ask about the presence of 'instructions for interviewers or/and respondents'. Coders report in Group 'additional information' about the topic or the scale, such as, 'extra motivations, information or definitions'. The seventh group of characteristics asks about the 'presence of an introduction' and if the question is repeated on it. Group 8 asks on the linguistic complexity of the item using as indicators

the ‘number of sentences’, of ‘subordinated clauses’, of ‘words’, ‘nouns’, ‘abstract nouns’ and ‘syllables’ in the request for an answer, the answer scale and in the introduction (if present). SQP also asks about the ‘method of data collection’ (group 9) and the language of the survey (group 10).

Finally, SQP asks information about the showcards or visual aid (if used). If the layout is ‘horizontal’ or ‘vertical’, if the ‘text is clearly connected to categories’ or if there is ‘overlap’. If ‘numbers’ or ‘letters are shown before answer categories’ or if ‘numbers are in boxes’, if the start of the ‘response sentence’ is shown on the showcard, if the ‘question is shown the showcard’ and if a ‘picture’ is provided.

Participating languages in a cross-cultural survey may be very different, being their structures closer or different from the source language. Empirical research across countries (after data collection) has identified that when item characteristics vary across source and translated items measurement invariance does not hold (Saris & Gallhofer 2007a, 2014; Saris et al. 2011, Oberski et al. 2008, Billiet 2006, Zavala-Rojas & Dorer). Therefore, the proposal in this paper is that item characteristics in different language versions should be compared systematically before data collection to prevent non-equivalence.

It is proposed that this comparison should be done using the coding scheme of SQP program because the codes are independent of the languages. As one cannot be familiar with all languages participating in a cross-cultural project, this coding scheme allows that trained individuals in survey research and proficient in the respective languages provide information about item characteristics. Once the characteristics are coded, these are the only elements that need to be compared to detect deviations across language versions.

As an illustration, consider this item taken from the ESS Round 5 source questionnaire in English:

If a violent crime were to occur near to where you live and the police were called, <u>how slowly or quickly</u> do you think they would arrive at the scene? Choose your answer from this card, where 0 is extremely slowly and 10 is extremely quickly.

Extremely slowly										Extremely quickly	(Don’t know)
00	01	02	03	04	05	06	07	08	09	10	88
				(Violent crimes never occur near to where I live)							55

When coded into the coding scheme of SQP, the 'domain' of this request is about 'local institutions' and the 'concept' specifies it is a 'judgement'. Other item characteristics that can be coded in SQP are that it is a 'direct request' in an 'interrogative' format using a 'WH word' and with a 'balanced' concept because it shows the two poles 'slowly/quickly.' About the response scale, it can be said that it is 'categorical' the 'number of categories' is 11, it is 'partially labelled'; labels are 'short texts' and it has 'three fixed reference points' because the qualifier 'extremely' denotes for an absolute ending point in the scale and there is a 'neutral' category (5).

When looking at the codes across the same item in different languages, it is obvious that characteristics such as 'Domain' and 'concept' should be kept the same. If they are different the questions are referring to different topics, but in the translation process there are other characteristics reflected in SQP codes that translators may vary. This variation will affect the equivalence with the source and other target versions.

It can also be said that this request has an 'instruction for the respondent': '*Choose your answer from this card…'*, a 'definition for the scale': '*…where 0 is extremely slowly and 10 is extremely quickly'* and a 'don't know option' which is not explicitly showed but only registered. The list of characteristics (approximately 60) allows having a very detailed map of the formulation of the item regardless the language. Therefore one can use these codes to detect differences in the formulation of a question in different languages.

For instance, this request in the source language has some formal properties that cannot be used in Lithuanian language at the same time: 1) use of a question word *'how'*; 2) gradation and, 3) balance in the request where both poles of the scale are present, '*slowly or quickly'*. The expression *'kaip lėtai ar greitai (*how slowly or quickly) introducing these properties would be completely inappropriate in Lithuanian language. Therefore, in order to keep a request balanced the translation team opted for omitting the WH word. They could include the WH word if gradation were the most important characteristic and rephrase the question. Then, they must omit either *"quickly" or "slowly"* in order to make this question sound fluent resulting in an unbalanced request. Unfortunately both are not possible in Lithuanian.

This example identifies important challenges for questionnaire translation as balance of requests; gradation and, WH words are item characteristics that have an impact in measurement instruments (Schuman & Presser 1981, Saris & Gallhofer

2007a, Alwin 2007, Andrews 1984). How can translators decide on which item characteristics should be the same across languages? How can they make a systematic assessment of all trade-offs that could appear in many languages? An answer to the first question is possible if one knows the potential effect of many features in the comparability of the requests. For answering the second question one needs a tool that makes it possible to compare the characteristics of all languages in a systematic way and detect possible deviations.

A second helpful illustration of how the coding scheme of SQP would help to detect deviations across languages, is the information that the 'linguistic characteristics' provide for comparing translated items. It is true that the number of words, syllables and subordinated clauses vary depending on the structure of each language. But outliers can be detected using very simple thresholds, for instance, one can check the number of languages which items are above (or below) one and two standard deviations in the number of words, nouns, syllables; or simply those which exceeds the number of sentences. Without knowing the meaning, this indicates an additional complexity (or simplification) of the items that could easily be confirmed in terms of content with the translation teams.

4. A procedure for comparing item characteristics across languages

It has been said on previous sections that survey research in questionnaire design has studied how different features of question wording and scaling affect responses. When a questionnaire is translated, the translation team faces different options of wording to remain equivalent with the source text. Currently, there is very little research on objective criteria to decide among different translation options (cf. Behr 2012). Translation assessment has remained a very subjective exercise. This paper suggests that the criteria to decide among translation options should be to preserve the item characteristics constant in both source and target versions (as long as the structure of the target language allows it). Those item characteristics have been defined by the tradition of questionnaire design in survey research and they are summarized in the coding scheme of SQP program.

4.1 Use of SQP to compare item characteristics in survey translation

For survey questions in different languages, one can check if their characteristics are the same when the questions are coded into a same coding scheme and the codes are compared. This makes it possible to compare the characteristics

independently of the languages. This paper explains a five-step procedure to compare different language versions of a survey item using SQP program. After describing it, the paper shows the main findings of its implementation in a sample of items from the ESS Round 5 questionnaire.

1) Introducing questions in SQP

Each question in the source and target languages should be introduced into the program SQP. This can be done by any user at no cost after signing up and logging in the program at sqp.upf.edu webpage. When coding, the program displays a help option on each screen indicated by a yellow box, which defines each item characteristic asked and gives examples (a complete codebook is also available in a PDF version).

2) Coding the source questionnaire

The information regarding the item characteristics of the source questionnaire must be accurate because target versions will be compared against it. It should be coded independently by two individuals with deep knowledge about questionnaire design; differences should be reconciled in collaboration with a third individual which plays the role of a reviewer.

3) Coding a target questionnaire

The translated questionnaire should be coded by a proficient speaker of the target language, preferably an individual involved in the translation process.

4) Comparison of measurement properties

The codes of the characteristics of source items should be compared with those in the target language. Any difference should be clarified with coders, first, to rule out coding errors in the target questionnaire. True differences in the codes should be reported to the translation team.

5) Interpretation of deviations and actions taken in the target text

The translation team should clarify any difference in the codes in terms of the definition of the features. In other words, it should justify the reasons behind a deviation in the item characteristics.

Depending on the type of difference, they may fall into one of three categories as shown in Table 2. Each category results in a suggested action for the translated text.

Table 2. Categories for differences in the SQP codes for two languages	
Type of deviations found (source vs. translation)	Action taken
A) A difference that cannot be warranted, for instance a different number of response categories, leaving out a "don't know" option or/and an instruction for the respondent.	The translation should be amended
B) A difference that may or may not be warranted e.g. use of complete sentences in the scales instead of short texts. In some languages it is necessary, in some others this may be a fact of stylistic choice	Amendments in the translation are recommended to keep the principle of functional equivalence in translation if the language structure allows keeping the item characteristic the same.
C) A difference in the linguistic characteristics that may be warranted e.g. different number of words, syllables. Also, a difference in the codes of linguistic characteristics that may not be warranted e.g. different number of sentences, nouns, extreme deviations in the number of words.	Amendments in the translation are recommended to keep the principle of functional equivalence in translation if the language structure allows it. If the differences are unavoidable due to linguistic characteristics, no change is recommended.

4.2 The questions evaluated in the ESS R5

In Round 5 (R5) of the ESS, 27 items of the main and supplementary questionnaires in 24 languages were coded in SQP. Although in some countries the questionnaire was translated into more than one language (for instance in Switzerland it was translated into French, German and Italian), in most of them only one language participated in this procedure. Participating languages were Bulgarian, Catalan, Croatian, Czech, Danish, Dutch in the Netherlands and in Belgium, Estonian, Finnish, French in Switzerland and France, German, Greek, Hebrew, Hungarian, Lithuanian, Polish, Portuguese, Russian, Slovak, Slovenian, Spanish, Swedish, and Ukrainian. Questionnaires in countries that share a language such as French in Switzerland and France are taken separately because in the ESS countries sharing the language do not implemented the same target versions, they are allowed to translate their own instruments to fit the target texts best into the country's context.

The items selected in the main questionnaire were part of the rotating module: “Trust in Criminal Justice: A Comparative European Analysis”. The items from the supplementary questionnaire were repetitions of the items in this module designed as experiments. Annex 1 shows the exact formulation of the items in the main and supplementary questionnaires as designed in the English Source version.

A member of the Core Scientific Team (CST) of the ESS introduced the items to be coded into SQP in the source language (English) and in the target languages. The source questionnaire was coded independently by two individuals with experience in survey research and differences were reconciled by a third reviewer (a survey methodologist).

National Coordinators (NCs) in participating countries were asked to provide codes on the formal and measurement characteristics of translated versions. The national coordinator as the person overseeing the survey in the specific country is the ultimate responsible for the quality of the translations.

Differences in the codes were first checked by the translation team for possible mistakes in the coding. All remaining differences represent deviations in item characteristics between the English source questionnaire and other language versions.

All deviations were reported to NCs in each country and they were asked about the reasons for the differences in the translation e.g. if it was a decision taken due to the characteristics of the language, if it was a cultural problem, if it was a mistake in the translation process, etcetera. To minimize deviations, recommendations were provided when changes to the translation were not fundamental to the structure of the language.

5. Findings

Differences between the codes of the characteristics of the source version and the translated versions falling in categories A (a difference that cannot be warranted) and B (a difference that may or may not be warranted) were found in 21 out of 24 languages. This means that in 87.5% of the questionnaires some item characteristics between the source and the target versions were different. This proportion does not include category C, the number of countries were the linguistic characteristics such as the number of nouns, syllables or sentences are different.

For the first category of deviations (A) suggestions to amend the questionnaire were made and the translations were changed, preventing unintended differences.

For the second category (B), when translation teams justified the reasons behind a deviation but they were not warranted, amendments in the translation were recommended to keep the principle of functional equivalence. However, most differences in this group remained unsolved because translation teams had strong arguments to keep them.

Table 3 shows all participating countries summarising if all differences in the codes were corrected; if corrections were only for some characteristics or if they were not implemented. It is shown that at the end of the process, the number of languages in which the item characteristics were the same as in the source questionnaire increased from 3 to 11 out of 24 participating languages. In 9 cases some deviations were corrected whereas others were kept and, in 4 cases they were not corrected at all.

Table 3. Summary information of participating countries

Country	Language	Deviations found	Deviations corrected
Belgium	Dutch	YES	YES
Bulgaria	Bulgarian	NO	---
Croatia	Croatian	NO	---
Czech Rep.	Czech	YES	NO
Denmark	Danish	YES	YES
Estonia	Estonian	YES	Partially
Finland	Finnish	YES	Partially
France	French	YES	NO
Germany	German	YES	YES
Greece	Greek	YES	YES
Hungary	Hungarian	YES	NO
Israel	Hebrew	YES	Partially
Lithuania	Lithuanian	YES	Partially
Netherlands	Dutch	NO	---
Poland	Polish	YES	Partially
Portugal	Portuguese	YES	YES
Russia	Russian	YES	YES
Slovakia	Slovak	YES	YES
Slovenia	Slovenian	YES	Partially
Spain	Spanish	YES	Partially
Spain	Catalan	YES	Partially
Sweden	Swedish	YES	NO
Switzerland	French	YES	Partially
Ukraine	Ukrainian	YES	YES

5.1 Category A: Differences that cannot be warranted

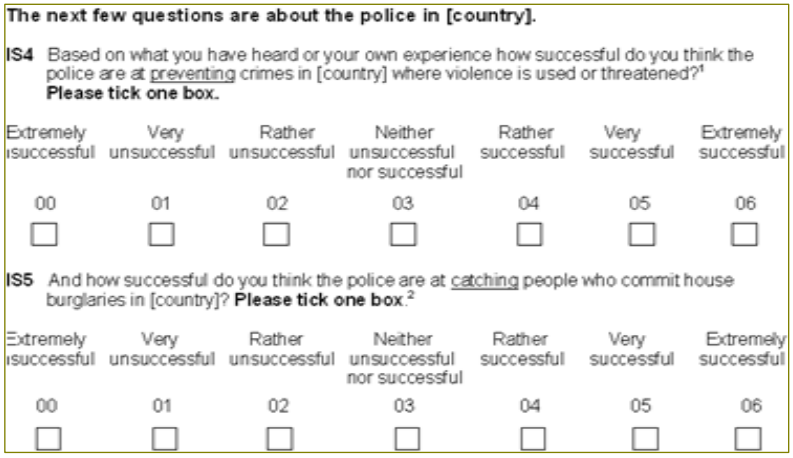
Most common problems in translations that were prevented were incorrect layout in self-administered questionnaires, inconsistent translations in formulations that were used in several items, increased complexity and, missing parts of the items. All the examples below were back translated in English to make them understandable for this paper.

Layout of direct questions in self-administered questionnaires

Krosnick (1990), Sanchez (1992), Saris and Gallhofer (2007a) among others have found a negative effect of batteries in the quality of responses. The effect is larger in self-administered modes of data collection and it cannot be assumed that respondents are answering the battery in the same way as they do when the questions are separated. Saris et. al. (1984), Neijens (1987) among others have found that complex batteries affect consistency of answers.

Through comparing the characteristics of the items across language versions, it was prevented that separate questions were formulated as batteries. The complexity would be different between the first and the subsequent items for some countries. Figure 1 presents how the items looked in the source questionnaire and Figure 1.1 how the translated version would look back-translated in English.

Figure 1. Layout of coded Items in source questionnaire



The next few questions are about the police in [country].

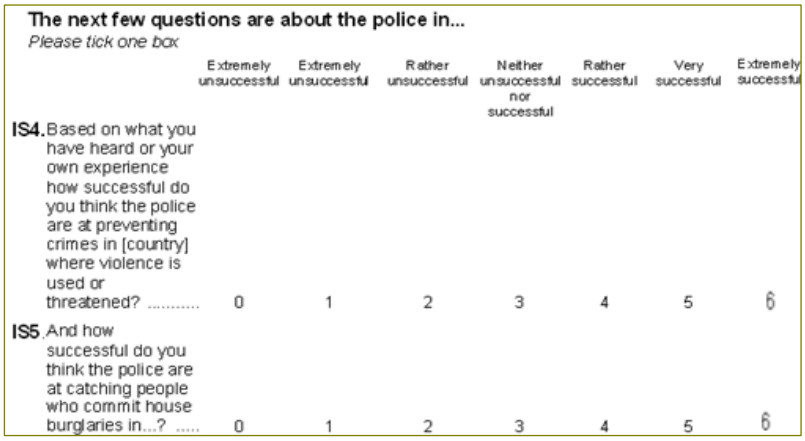
IS4 Based on what you have heard or your own experience how successful do you think the police are at preventing crimes in [country] where violence is used or threatened?[1] **Please tick one box.**

Extremely successful	Very unsuccessful	Rather unsuccessful	Neither unsuccessful nor successful	Rather successful	Very successful	Extremely successful
00	01	02	03	04	05	06
☐	☐	☐	☐	☐	☐	☐

IS5 And how successful do you think the police are at catching people who commit house burglaries in [country]? **Please tick one box.**[2]

Extremely successful	Very unsuccessful	Rather unsuccessful	Neither unsuccessful nor successful	Rather successful	Very successful	Extremely successful
00	01	02	03	04	05	06
☐	☐	☐	☐	☐	☐	☐

Figure 1.1 Layout of coded Items in the translated version



The next few questions are about the police in...

Please tick one box

	Extremely unsuccessful	Extremely unsuccessful	Rather unsuccessful	Neither unsuccessful nor successful	Rather successful	Very successful	Extremely successful
IS4. Based on what you have heard or your own experience how successful do you think the police are at preventing crimes in [country] where violence is used or threatened?	0	1	2	3	4	5	6
IS5. And how successful do you think the police are at catching people who commit house burglaries in...?	0	1	2	3	4	5	6

Differences in showcards layout were also prevented, for instance, the visual presentation of the scale in a vertical or horizontal format or the absence of numbers in front of categories as it was designed in the source questionnaire.

Missing characteristics of items

It was prevented that definitions of the scale and introductions formulated in the source questionnaire were missing in the translated versions. For example in

'Based on what you have heard or your own experience how unsuccessful or successful do you think the police are at preventing crimes in [country] where violence is used or threatened?'

If the item has no definition of the scale, then the respondent does not know what to answer. Therefore an instruction is necessary in this item: *'Choose your answer from this card, where 0 means very unsuccessful and 4 means very successful.'* This deviation was detected in one target version and a correction was introduced in the questionnaire

5.2 Category B: differences that may or may not be warranted

In the process of comparing the characteristics of the English source questionnaire and the translated versions some deviations could not be prevented, most of them related to the formulation of scales. They are challenging because they may impact comparability but at the same time, countries had strong arguments to keep them.

Balance/unbalanced items

It has been discussed for a long time in the literature that unbalanced requests can mislead answers (See Alwin 2007 and Saris & Gallhofer 2007a for a review). As this paper showed in a previous section, in Lithuanian language it was not possible to balance a question and combine some other characteristics at the same time. It is not appropriate to use a question word *'how'* to give gradation and balance a request showing the two poles of the scale, '*slowly or quickly'*. The expression *'kaip lėtai ar greitai (*how slowly or quickly) introducing these properties is not appropriate in the use of the language. In order to keep a request balanced the translation team opted for omitting the WH word, omitting gradation as well. They could include the WH and omit either *"quickly" or "slowly"* resulting in an unbalanced request.

Translation of bipolar/ unipolar concepts

In the item *'And how successful do you think the police are at…?'* the labels for the ending points of the scale *'extremely unsuccessful/successful'* were translated as a *'extremely inefficient/efficient', 'extremely ineffective/effective'* and '*extremely bad/well'* in Spanish, Catalan, French and Finnish respectively. Translation teams interpreted the adjective *successful* as bipolar which opposite pole was unsuccessful.

They argued that there was not an equivalent formulation to successful and unsuccessful in a bipolar range and they adapted the adjective. In English, this adjective is understood as unipolar and translations could have been formulated ranging from not successful/ extremely successful. But translation teams in these languages argued that their translation represented the same meaning as successful in English.

Labels for categories

Another deviation that was not corrected was the use of long texts for labels instead of short texts in French in Switzerland and France and in Polish. Using as an example the English source survey item

'If a violent crime were to occur near to where you live and the police were called, how slowly or quickly do you think they would arrive at the scene? Choose your answer from this card, where 0 is extremely slowly and 10 is extremely quickly'.

Extremely slowly										*Extremely quickly*	*(Don't know)*
00	*01*	*02*	*03*	*04*	*05*	*06*	*07*	*08*	*09*	*10*	*88*
				(Violent crimes never occur near to where I live)							*55*

Translations for the scale labels were formulated as *'It will arrive extremely late to the place"* and *"It will arrive extremely quickly to the place.'* The main argument was that *'Extremely slowly/quickly'* were difficult labels for less educated people. Respondents may think in driving fast or slow. To solve this issue, the translation included a long sentence to clarify the scale. Pretesting in Poland (N=50) indicated that the understanding of this scale was difficult for people with low levels of education; however if this was the case in Poland, it is highly likely that the same explanation would be needed for other countries as well.

Fixed reference points

Another frequent deviation was the translation of qualifiers defined as *fixed reference points*. A fix reference point is an anchor; there is no doubt about its position on the subjective scale in the mind of the respondent (Saris & Gallhofer 2007a). In contrast the position of an unfixed reference point, a vague qualifier which can be interpreted differently among respondents.

Fixed reference points of the scale such as *'extremely slowly'*, *'neither slowly nor quickly*' or *'extremely quickly'* were used in the source questionnaire as anchors, whereas some languages used the equivalent form in English of *'very slowly'* and *'very quickly'*. This second form is not a fixed reference point, because respondents can have a different idea of what *'very'* means depending in their own subjective reference. The same deviation was found for end-points of the scale such as *'extremely successful'* and *'extremely unsuccessful'*.

This deviation was seen in five languages: Slovak, Russian, Hebrew, Polish and Hungarian. In the later three, NCs argued that there were no equivalent adverbs to '*extremely'*, thus they reformulated the labels of the end-points as *'very*'. A second argument given by the countries was that the formulation was possible but it was difficult to be understood by less educated people because it could be understood as *'extremist'.* As in the case of long sentences for labels, if some formulations needed additional explanations; this could be likely true for more languages.

A last example of this frequent deviation was found in four languages: Czech, Slovakia, Russia and Poland translated the scale *'not at all likely'* as *'very unlikely'* and the scale *'not at all often'* into *'very rare'*. The reason was that the expression *not at all* is idiomatic in English. Therefore, it is difficult to decide how it can be represented in target languages. A solution was to use the equivalent to 'never' instead of trying to reproduce not at all.

5.3 Category C: Differences in the linguistic characteristics

Inconsistent translation in repeated characteristics

Repeated expressions should be translated in the same way; there is no need to develop a new translation for instructions or concepts that are used several times in a questionnaire. This can be especially problematic in the design of experiments, were varying elements in the formulation of items disturbs an experimental design. A systematic check of repeated wordings is difficult without a program. SQP asks the number of words, nouns and abstract nouns in items. It is easy to detect a deviation if these numbers are different in expressions that are repeated several times in a questionnaire. In this way, instructions or concepts had a coherent translation in other parts of the questionnaire. A variation of this kind of problem was prevented in four languages.

An example is the instruction for the respondent *'Use the same card'* which was translated inconsistently in one language into an equivalent of *'Please, use this card to answer.'* In another language, the translation of *'violent crime'* was suggested as *'aggression'* in a first occasion and as *'a crime or an offense'* in a repetition of the same concept. These deviations were detected because the number of words was different for a same sentence.

Increased complexity of the items

SQP coding made it possible to check for deviations that would vary the complexity of the items when extra explanations were included making the item more complex. For example, in the request:

'Based on what you have heard or your own experience, how often would you say the police generally treat people in [country] with respect? Choose your answer from this card, where 0 means almost never and 10 means almost always.'

An additional question at the end of the item was phrased as *'Would you say that this is the case…?'* If this second question would remain the respondent were asked twice the same question. This was an additional unnecessary repetition.

In summary, there were three main arguments to keep a deviation in a target version. The first common argument was related to the characteristics of the language: an equivalent formulation was impossible, in this case the difference was considered as warranted. The second argument was directed to the cultural context: some formulations had a negative connotation in the country or they were difficult to be understood by less educated people. A third argument was that the difference was justified because it has improved fluency in the request. These deviations are considered as unwarranted because in principle the structure of the language would allow a closer translation. However, the arguments given by the national teams and their implications for functional equivalence remain under debate and should be considered more in detail in future research.

6. Conclusions

Testing equivalence in survey translation

This paper has focused on a current problem in comparative survey research. Survey translation has developed best practice procedures to translate functionally equivalent survey instruments. But in practice, it is very difficult to check empirically requirements set by translation guidelines because one cannot understand all languages. Empirical methods are mostly used for detecting flaws once data is collected. There is little research on how to assess empirically cross-cultural instruments before they are administered to respondents.

Best practice procedures to translate and assess the quality of a survey item do not have a direct link to testing equivalence (or invariance). They rely on judgements that may be partial, only focused in some characteristics or cognitive processes, or subjective, because they depend on the evaluators' knowledge of the context in which the survey is embedded or even stylish preferences about the language. The procedures that are thought to test equivalence, such as the pilot study in the PISA, are not affordable for most survey projects.

This paper suggested a procedure to detect deviations relevant for comparability of different language versions of a survey instrument before it is administered to respondents. It requires comparing the item characteristics of source and target survey items in a same coding scheme. This coding scheme is developed in the form of a semiautomatic program called Survey Quality Prediction (SQP). Once survey items are coded into SQP, their characteristics can be compared regardless the item languages in a systematic way.

This procedure was applied on a set of items in the ESS Round 5. Findings were classified into three categories: In the first were differences in item characteristics that were not warranted and could be prevented. This led to changes in the translation of

some items in some languages. The result was that the items' form after the correction was closer to the one in the source questionnaire.

In the second category, there were two possibilities either the deviations were warranted or they were not, this depended on the arguments given by the national translation teams. Amendments in the translation were recommended to keep the principle of functional equivalence in translation if the language structure allowed keeping the item characteristic the same.

There were some differences in this category that in principle could be avoided, but they remained unsolved because national teams involved in the translation processes had strong arguments to keep them, stating for instance that they were more helpful to less educated people and that fluency was improved in the interview. There is not an answer to this issue. It opens a line of debate for further research on comparative questionnaire design.

A third category of deviations in the item characteristics are unavoidable: differences in the linguistic characteristics such as different number of words or syllables, this depends on the language structure. However, they can be indicators of the complexity of an item, the coding process makes it possible to detect in a systematic way if there are differences in the number of sentences or if there are extreme deviations in the number of words.

Flaws in the source questionnaire

The coding process showed some of the implications that flaws in the source text have in the translation. There were problems in the formulation of the source questionnaire that affected several linguistic or cultural groups. Problems related to the concepts *'successful'* and *'unsuccessful'* are clearly of this type. They were expressions easily understood in English language but very difficult to translate into other languages. According to the documentation on the questionnaire design of the ESS, this item measured *trust in police effectiveness focused on achievements or*

outputs (ESS 2011. A more general term *effective* or *efficient* would be more appropriate for the source questionnaire.

A second problem in the design of the source questionnaire was the ending point of the scale '*not at all often.'* It is an expression used to indicate non occurrence, the adverb *not at all* is used repeatedly in the questionnaire as an ending label in combination with other adverbs or adjectives, however the combination with *'often'* makes it very specific for English and even there seldom used. A solution could be the adverb *'never'* to keep the characteristic of a fix reference point and also a label for a zero probability of occurrence. Through consultation with different national teams, this could be a general solution for Slavic languages in the future.

Other common problems are cultural in nature where items do not function properly in specific contexts such as misunderstanding of the qualifier 'extremely' as 'extremist' in some countries. A solution is to improved guidelines to indicate what item characteristic is expressed by the adverb extremely (a fixed reference point). This would allow national teams to search and look for other possibilities such as fully, completely, etcetera.

Finally, the coding process brings advantages for an effective communication between the survey designers and the translators. The item characteristics set the framework of what is expected from a translator regarding a survey text. If the framework of the elements that need to be fixed across languages is clear, it would be easier for the translators to take decisions on item formulation.

References

Armer, M. (1973). "Methodological problems and possibilities in comparative research." *Comparative social research: Methodological problems and strategies* pp: 49-79.

Behr, D. (2012), Translationswissenschaft und international vergleichende Umfrageforschung: Qualitätssicherung bei Fragebogenübersetzungen als Gegenstand einer Prozessanalyse, GESIS. [Translation: Research and cross-national survey research: Quality assurance in questionnaire translation from the perspective of translation process research.]

Benet-Martínez, V., & John, O. P. (1998). "Los Cinco Grades" across cultures and ethnic groups: Multitrait–multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology*, 75, 729–750.

Braun, Michael; Johnson, Timothy P. (2010) "An illustrative review of techniques for detecting inequivalences." In: Harkness, Janet A.; Braun, Michael; Edwards, Brad; Johnson, Timothy P.; Lyberg, Lars; Mohler, Peter Ph.; Pennel, Beth-Ellen; Smith, Tom W. (Hrsg.) *Survey methods in multinational, multiregional, and multicultural contexts*. NJ: John Wiley and Sons.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of cross-cultural psychology*, *1*(3), 185-216.

Brislin, R. W. (1976). Comparative research methodology: Cross-cultural studies. *International journal of psychology*, *11*(3), 215-229.

Byrne, B. M., & van De Vijver, F. J. (2010). Testing for measurement and structural equivalence in large scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107-132.

Davidov, E., J. Billiet and P. Schmidt (2011) (Eds.) "Methods and applications in cross-cultural analysis". NY: Routledge.

Dean, E., Caspar, R., McAvenchey, G., Reed, L., and Quiroz, R. (2005), Developing a lowcost technique for parallel cross-cultural instrument development: The Question Appraisal System (QAS-04), in J. H. P. Hoffmeyer-Zlotnik and J. A. Harkness (Eds.), *Methodological Aspects in Cross-National Research,* pp. 31-46, Mannheim, Germany: ZUMA Nachrichten Spezial, Volume 11.

Dorer, B. (2012) *Round 6 Translation Guidelines*. Mannheim: European Social Survey, GESIS

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

European Social Survey (2010). ESS Round 5 Translation Guidelines. Mannheim, European Social Survey GESIS.

European Social Survey (2011a) *Round 5 Module on Trust in the Police & Courts– Final Question Design Template*. London: Centre for Comparative Social Surveys: City University London.

European Social Survey (2011b) *Round 6 Specification for Participating Countries*. London: Centre for Comparative Social Surveys, City University London.

John, O. P., Goldberg, L. R., & Angleitner, A. (1984). Better than the alphabet: Taxonomies of personality-descriptive terms in English, Dutch, and German. In H. Bonarius, G. van Heck, & N. Smid (Eds.), *Personality psychology in Europe: Theoretical and empirical developments.* Lisse, The Netherlands: Swets & Zeitlinger.

Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (Eds.). (2007). "Measuring attitudes cross-nationally: Lessons from the European Social Survey." London: Sage Publications.

Hambleton, R. K., Merenda, P. F. & Spielberger, C. D. (Eds.). (2005). Adapting educational and psychological tests for cross-cultural assessment. Mahwah, NJ.: Lawrence Erlbaum Associates.

Harkness, J. A. (2010). "Translation" In Survey Research Center (ed). *Guidelines for Best Practice in Cross-Cultural Surveys.* Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved December 13, 2012, from http://www.ccsg.isr.umich.edu/

Harkness, J. A. (2007). Improving the comparability of translation. In R. Jowell, C. Roberts, R. Fitzgerald & G. Eva (Eds.), *Measuring attitudes cross-nationally: Lessons from the European Social Survey* (pp. 79-94). London: Sage Publications.

Harkness, J. A. (2003). Questionnaire Translation. In: Harkness, J. A., van de Vijver, F. and Mohler, P. Ph. (eds.). *Cross-cultural Survey Methods*. Hoboken, NJ: John Wiley & Sons.

Harkness, J. A. (Ed.) (1998). "Cross-cultural survey equivalence" Mannheim, Germany: ZUMA.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L. E., Mohler, P. P., Pennell B.E., & Smith, T. W. (Eds.). (2010a). Survey Methods in Multicultural, Multinational, and Multiregional Contexts (Vol. 552). John Wiley & Sons.

Harkness, J. A., Villar, A., & Edwards, B. (2010b). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. E. Lyberg, P. Ph. Mohler, B-E. Pennell & T. W. Smith (Eds.), *Survey methods in multinational, multicultural and, multiregional contexts*. Hoboken, NJ: John Wiley & Sons.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. and Villar, A. (2010c). "Designing Questionnaires for Multipopulation Research." *In Survey Methods in Multinational, Multicultural and Multiregional Contexts,* edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, & Tom W. Smith. Hoboken, NJ: John Wiley and Sons, 31-57.

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

Harkness, J.A. and Schoua-Glusberg, A. (1998). "Questionnaires in Translation" in: Harkness, J.A. (Ed.). *Cross-Cultural Survey Equivalence*. ZUMA-Nachrichten Spezial No. 3.. Mannheim: Zentrum für Umfragen, Methoden und Analysen.

Harkness, J. A., van de Vijver, F. J. R., & Johnson T. P. (2003). Questionnaire design in comparative research. In J. A. Harkness, F. J. R. van de Vijver & P. M. Mohler (Eds.), *Cross-cultural survey methods*. Hoboken, NJ: John Wiley & Sons.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research*, *18*(3), 117-144.

Hui, C. H., & Triandis, H. C. (1985). Measurement in Cross-Cultural Psychology A Review and Comparison of Strategies. *Journal of cross-cultural psychology*, *16*(2), 131-152.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika*, *36*(4), 409-426.

Krosnick, J. A. (1990). The impact of satisficing on survey data quality. In Proceedings of the Bureau of the Census 1990 Annual Research Conference (pp. 835-845). Washington, D.C.: U.S. Government Printing Office.

Krosnick, J. A., & Fabrigar, L. R. (forthcoming). *The handbook of questionnaire design*. New York: Oxford University Press.

Likert, R. (1932). "A Technique for the Measurement of Attitudes". *Archives of Psychology* 140: 1–55.

Lynn, P., Japec, L., and Lyberg, L. (2006) What's so special about cross national surveys? (1998) in: Harkness, J.A. (Ed.). *Conducting Cross-national and Cross-cultural Surveys*. ZUMA-Nachrichten Spezial No. 12. Mannheim: ZUMA.

Mallinckrodt, B., & Wang, C. C. (2004). Quantitative Methods for Verifying Semantic Equivalence of Translated Research Instruments: A Chinese Version of the Experiences in Close Relationships Scale. *Journal of Counseling Psychology*, *51*(3), 368.

Meredith W. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543.

Michalos, A. C. (Ed.) Encyclopedia of Quality of Life Research. Springer. Forthcoming 2013.

Mohler, Ph. P., and Johnson, T. (2010). "Equivalence, comparability and, methodological process" In *Survey Methods in Multinational, Multicultural and Multiregional Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy Johnson, Lars Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, & Tom W. Smith. Hoboken, NJ: John Wiley and Sons, 31-57.

Neijens P. 1987. *The Choice Questionnaire. Design and Evaluation of an Instrument for Collecting Informed Opinions of a Population*. *PhD thesis*, Amsterdam: Free University.

Nida, E. A. (1964). *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Oberski, D. Saris, W.E. and Hagenaars, J. 2007, "Why are there differences in measurement quality across countries?", In Measuring Meaningful Data in Social Research (Geert Loosveldt, Swyngedouw, eds.), Acco, Leuven.

PISA (Programme for International Student Assessment) (2012), *Translation and Adaptation Guidelines for Pisa*, Paris: OECD.

PISA (Programme for International Student Assessment) (2009), *PISA 2009 Technical Report*, Paris: OECD.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. Ò., Martin, E. A., Martin, J., et al. (Eds.) (2004), *Methods for Testing and Evaluating Survey Questionnaires,* Hoboken, NJ: John Wiley & Sons.

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J., & Pennebaker, J. W. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of Research in Personality*, 40, 99–120

Sanchez, Maria Elena (1992) "Effects of questionnaire design on the quality of survey data." *Public Opinion Quarterly* 56.2: 206-217.

Saris, W. (2012) "Evaluation procedures for survey questions". RECSM Working Paper 26.

Saris, W.E. (2013) "The prediction of question quality: The SQP 2.0 software", in B. Kleiner, I. Renschler, B. Wernli, P. Farago and D. Joye (eds.) *Understanding Research Infrastructures in the Social Sciences*. Zurich: Seismo Press; pp. 135-144.

Saris W. E. (ed.) (1988) *Variations in Response Functions: a Source of Measurement Error in Attitude Research*. Amsterdam: Sociometric Research Foundation.

Saris W. E., P. Neijens, and J. A. de Ridder 1984. Resultaten van de keuze enquete in het kader van de B.M.D. In Stuurgroep Maatschappelijke Diskussie Engergiebeleid: *Het Eindrapport, Appendix*. Leiden: Stenfert Kroese.

Saris, W.E., and Gallhofer, I.N. (2014) *Design, Evaluation and Analysis of Questionnaires for Survey Research*. New York: Wiley.

Saris, W.E., and Gallhofer, I.N. (2007a) *Design, Evaluation and Analysis of Questionnaires for Survey Research*. New York: Wiley.

Saris, W.E., and Gallhofer, I.N. (2007b) "Estimation of the effect of measurement characteristics on the quality of survey questions" *Survey Research Methods*, 1(1); pp.29-43.

Saris, W.E., and Gallhofer, I.N. (2004) "Operationalization of Social Science Concepts by Intuition", *Quality and Quantity*, 35; pp.235-258.

Saris, W.E., Oberski, D., Revilla, M., Zavala, D., Lilleoja, L., Gallhofer, I.N., Grüner, T. (forthcoming 2011) "Final report about the project JRA3 as part of ESS Infrastructure" European Social Survey.

Saris W. E., W. van der Veld, and I. N. Gallhofer 2004a. Development and improvement of questionnaires using predictions of reliability and validity. In S. Presser , J. M. Rothgeb, M. P. Couper, J. T. Lessler, E. Martin, J. Martin, and E. Singer (eds.), *Methods for Testing and Evaluating Survey Questionnaires*. Hoboken: Wiley, 275–299.

Scherpenzeel A. C., and W. E. Saris 1997. The validity and reliability of survey questions: A meta-analysis of MTMM studies. *Sociological Methods and Research*. 25, 341–383.

Scherpenzeel A. C. 1995. A Question of Quality. Evaluating Survey Questions by Multitrait-Multimethod Studies. KPN Research: Leidschendam.

Scheuch, E. k. (1968), The cross-cultural use of sample surveys: Problems of comparability, in S. Rokkan (Ed.), Comparative Research Across Cultures and Nations, pp. 176-209, Paris: Mouton.

Scheuch, E.K. (1993) "The Cross-Cultural Use of Sample Surveys: Problems of Comparability", *Historical Social Research*, Vol. 18 No. 2, pp: 104-138.

Smith, T. W. (2004), Developing and evaluating cross-national survey instruments, in S. Presser et al. (Eds.), *Methods for Testing And Evaluating Survey Questionnaires,* pp. 431^152, Hoboken, NJ: John Wiley & Sons.

Steenkamp, J. B. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of consumer research*, *25*(1), 78-107.

Survey Research Center. (2010). Guidelines for Best Practice in Cross-Cultural Surveys. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved December 13, 2012, from http://www.ccsg.isr.umich.edu/

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational research methods*, *3*(1), 4-70.

Van der Veld W., W. E. Saris, and I.N. Gallhofer 2000. *Survey Quality Prediction: SQP.* Paper presented at the ISA Methodology Conference in Cologne, Germany.

Van de Vijver, F. J., & Leung, K. (1997). *Methods and data analysis for cross-cultural research* (Vol. 1). Sage.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology*, *54*(2), 119-135.

Van Herk, H., Poortinga, Y. H., & Verhallen, T. M. (2004). Response Styles in Rating Scales Evidence of Method Bias in Data From Six EU Countries. *Journal of Cross-Cultural Psychology*, *35*(3), 346-360.

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

Willis, G. B., & Lessler, J. T. (1999). *Question appraisal system QAS-99*. Rockville, MD: Research Triangle Institute.

Zavala Rojas, D. (2012) "Evaluation of the concepts 'Trust in institutions' and 'Trust in authorities' (European Social Survey Deliverable 12.4: Evaluation of questions and concepts - report 2 (Political Trust))". RECSM Working Paper 29.

Zavala Rojas, D. & Dorer, B. (2013) "A Mixed Method Approach to Link Translation Deviations and Measurement Quality" Mimeo.

Software:

Oberski D., L. Kuipers, and W.E. Saris 2007. *SQP 1.0 Survey Quality Predictor*. www.sqp.nl