Striving for quality, comparability and transparency in cross-national social survey measurement: illustrations from the European Social Survey (ESS)

Rory Matthew Fitzgerald

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy at City University London, Department of Sociology

November 2015

I grant powers of discretion to the City University London Librarian to allow this thesis to be copied in whole or in part without further reference to me. This permission covers only single copies made for study purposes, subject to normal conditions of acknowledgement.

ABSTRACT

It is well known that even the best conducted surveys often generate significant amounts of error during their design and implementation, best described by the Total Survey Error (TSE) framework. It is also widely accepted that cross-national surveys have the potential to increase that error further. This is because they often comprise of multiple national surveys under some kind of coordinated framework, whilst also having additional sources of error that stem from the cross-cultural nature of that work.

Recently great strides have been made in identifying the sources of error that can impact on social surveys and how these are magnified in a cross-national context. This doctorate presents a body of my own published work that has contributed to the field of cross-national research. It has provided tools and approaches that help in the identification and correction of three overarching aspects of non-sampling survey error: specification error, measurement error and non-response error. In each of these areas my contribution to the field through work on the cross-national European Social Survey (ESS) will be demonstrated, drawing on peer reviewed journal articles, book chapters, a book and published working papers. This academic research has added to knowledge in the field and made a practical contribution by leading to tangible improvements in the methodology of the ESS and other cross-national social surveys.

Contents

List of Tables

Acknowledgements

I would like to thank my supervisors Professor Howard Tumber and Dr Tom Smith for their support and encouragement during the writing of this thesis. Their critical evaluations and suggestions for improvement were extremely valuable.

I would also like to acknowledge how thankful I am to my co-authors from across the world with whom it has been a pleasure to work. I gained much both professionally and personally from many of those collaborations.

I would also like to acknowledge those involved with the European Social Survey whether as scientists, survey practitioners or respondents. Without their professionalism or willingness to participate in the survey this investigation would not possible.

I dedicate this thesis to the late Professor Sir Roger Jowell, my mentor and academic inspiration. His support was instrumental in developing my academic work and ensuring its impact stretched far beyond publications themselves.

Finally sincere thanks to my parents Annie and Mike, my partner John, my friends and of course my assistant Mary. Without their love, support and patience I would never have made it to the end of this journey.

Chapter 1: 'Error sources in survey research and the development of cross-national survey research'

The 20th Century saw the inception and rapid development of survey research and eventually the (almost complete) globalisation of the field (Heath, Fisher, and Smith 2005; Norris 2009; Smith 2010; Smith and Fu 2014). Conducting surveys is, however, a remarkably complex and error-prone task. There is a debate over whether human subjects make social science more challenging than natural sciences where researchers deal with inanimate phenomena (Jowell et al. 2007). More recently great strides have been made in identifying the sources of error that can impact on social surveys (Biemer 2010; Biemer and Lyberg 2003; Groves and Lyberg 2010) and discussing how these are magnified in a cross-national context (Smith 2011).

This doctorate presents a body of my published work that has contributed to the field of cross-national research, by providing tools and approaches that help in the identification and correction of specific elements of survey error. It is well-known that even the best conducted surveys often generate significant amounts of error during their design and implementation (Biemer 2010). It is also widely accepted that cross-national surveys have the potential to increase that error (Smith 2011). This is because they generally comprise multiple national surveys under some kind of coordinated framework (Smith 2011) whilst also having additional sources of error that stem from the cross-cultural nature of that work (Jowell et al. 2007).

This doctorate is related to three overarching aspects of non-sampling survey error: specification error, measurement error and non response error. In each of these areas my contribution to the field through work on the cross-national European Social Survey (ESS) will be demonstrated, drawing on peer reviewed journal articles, book chapters, a book and published working papers. This academic contribution has added to knowledge in the field and made a practical contribution by leading to tangible improvements in the methodology of the ESS and other surveys. An additional component of total survey error – protocol error – will also be introduced and my work to reduce that error source discussed.

This introductory chapter will first outline the key approaches to identifying and mapping error in survey research in order to describe the various 'error sources' that can damage a survey and lead to poor quality estimates. The Total Survey Error (TSE) framework will be discussed to

highlight the various sampling and non-sampling errors that survey researchers should try to overcome or at least be aware of. The way in which this framework relates to cross-national surveys and needs will then be discussed, with particular reference to the specific methodological domains covered in this thesis.

The chapter will move on to describe the historical development of cross-national social survey research, outlining how the field has developed over time and identifying key areas where improvements are still required. Against that background the different approaches to cross-national survey research which have persisted to date will be discussed. The chapter will then position the European Social Survey (ESS) within that framework. The need for rigour, transparency and harmonisation will be emphasised and the contribution of the ESS in these areas highlighted.

European Social Survey[1]

Most of the work forming part of this doctorate by prior publication has been conducted under the framework of the European Social Survey (ESS). The ESS (www.europeansocialsurvey.org) is a cross-national survey of attitudes and behaviour that aims to chart change and stability in the social and moral fabric of Europe, using a rigorous methodology (Jowell et al. 2007). The ESS is led by its Principle Investigator and Director and a Core Scientific Team drawn from 7 institutions in Europe. Collectively they are responsible for the design and coordination of each round of the survey. The ESS headquarters where the Director (who is also the Principle Investigator) of the ESS is situated, is at City University London, UK. The first round of data collection was conducted in 2002 with a biennial frequency since that time. The Director issues a specification for each round of fieldwork and each country then appoints a National Coordinator who is responsible for realising that specification in their country, adapting it where necessary. Most countries also appoint a survey agency to conduct the face-to-face fieldwork, most often a commercial company but sometimes a national statistical institute, non-profit organisation or even self-organised efforts by a university. Data is then deposited to the ESS data archive and made freely available for download. There are over 80,000 registered users of the ESS website and nearly 3000 academic publications to date. Results have also been disseminated directly in policy forums including the European, Italian and Lithuanian Parliaments, at the OECD and within the European

[1] I became the Director of the ESS in January 2012. This followed the death of the ESS founder Director Sir Roger Jowell in December 2011.

Commission. The methodology of the ESS has also influenced the methods used by other cross-national survey programmes such as the European Quality of Life surveys and the European Values Surveys.

In 2013 the European Social Survey became a European Research Infrastructure Consortium (ESS ERIC) in recognition of its contribution to understanding the social condition of Europe and for its scientific excellence. An ERIC is a legal organisation established under European Community law. ERICs are publically owned legal entities responsible for operating research infrastructures, formed and operated by the countries which establish them. The governments of those member countries are then responsible for operating the infrastructure.

In order to understand the focus of my work, its potential impact for reducing TSE and how it has moved the field forward methodologically, it is necessary to understand the background to the ESS and how it fits into the wider landscape of cross-national survey research. This is discussed later in this chapter. First, error in survey research is discussed and the concept of Total Survey Error is introduced.

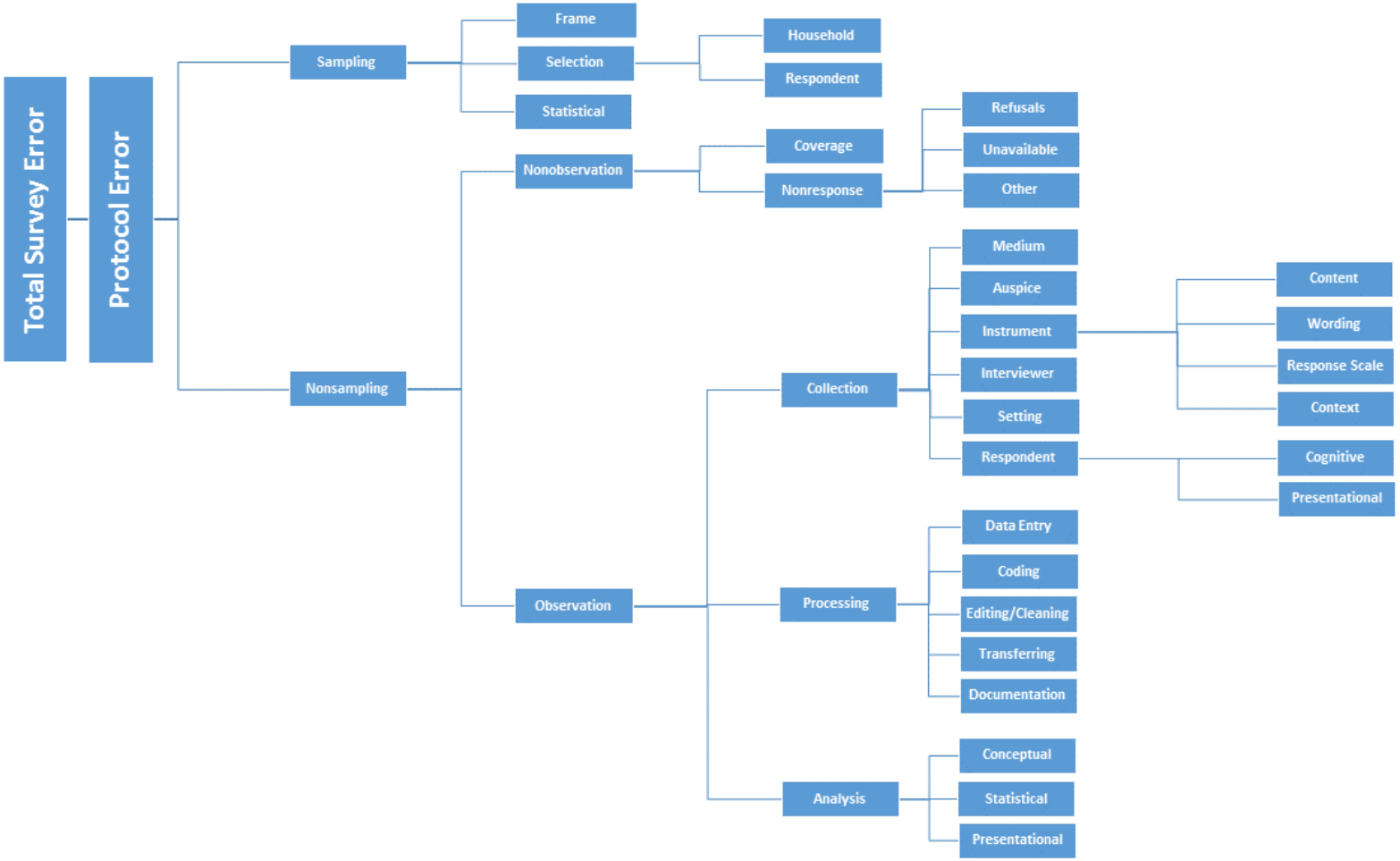
Total Survey Error

In the paper, 'Past, present and future', the authors outline the historical development of the Total Survey Error (TSE) concept. TSE is ''... a conceptual framework describing statistical error properties of sample survey statistics" (Groves and Lyberg 2010, 849). Ideally it would allow a researcher to calculate the statistical accuracy of estimates produced in a probability sample survey, taking account of all possible errors that can prevent that estimate being a perfect reflection of reality. However, despite the general consensus in the field about the overall concept of TSE, it has been acknowledged that there are variations in the precise components that different researchers use to operationalise the concept (Ibid). This results in the existence of a number of different typologies, but of course not all of these can be 'total' in their coverage of error unless the differences are purely descriptive. In fact it is important to remember that the 'total' in the title dictates that TSE must be as comprehensive as possible, which is not the case with all (or perhaps any?) of the existing typologies (Smith 2011, 464) . As we will see below, even the rather comprehensive typology outlined by Smith has omissions.

Despite these differences there is basic agreement that there are two major sources of error that impact on surveys, namely, sampling and non-sampling error (Biemer 2010; Biemer and Lyberg 2003; Groves and Lyberg 2010; Smith 2011). Smith presents a very clear overview of TSE (Figure 1.1) which presents the two types of error as parallel lines flowing from each box or component rather than showing separately the variance (which is random and has no expected impact on estimates) and the bias (which is directional and can distort survey estimates). This method of presentation allows a focus on the error source itself in terms of the dimension of the survey life cycle which is the root cause of the error and prevents the visual portrayal of TSE from becoming too complicated. Figure 1.1 portrays this visually with just one addition by the author of this thesis which will be discussed below (the addition of 'protocol error').

Sampling error, strictly related to probability samples, has three key dimensions: the first is related to the frame itself (eg coverage), the second to the selection process (of the address, household or individual) and the third is related to statistical processes (such as weighting the data for unequal selection probabilities) (Smith 2011). Naturally the exact location of specific dimensions can always be debated. For instance the interviewer's execution of the sampling design – the selection of households and respondents - leaves open the possibility of error stemming from individual interviewers' compliance with the required rules (Stoop et al. 2010). So this element usually sits under the interviewer domain. However it also remains critical to sampling.

Figure 1 TSE (Adapted from (Smith 2011, 476)



It is widely acknowledged that due to their statistical nature, sampling processes are (at least theoretically) fully measurable in terms of error whilst many non-sampling areas can often only be measured with 'significant alteration of the typical survey designs' (Groves and Lyberg 2010, 850) or perhaps cannot even be measured at all with current methods. It is these non-sampling areas which are focused upon in this thesis.

Smith subdivides the non-sampling errors into two types: non-observational and observational. The non-observational errors include coverage and non-response, with the latter subdivided into 'refusals', 'unavailable' and 'other' (Smith 2011). Measuring non response is discussed in this thesis, with the ESS approach to recording each contact attempt by an interviewer and the resulting dataset together representing a major innovation in cross-national research. Two publications (Billiet et al. 2007; Stoop et al. 2010) illustrate my contribution.

The observational domains of TSE include collection, processing and analysis (Smith 2011). Attempting to minimise the error associated with collection through questionnaire development and pre-testing have been central to my academic focus. This work has involved trying to reduce error related to the content, wording, response styles and context through the better specification and design of questionnaires. In addition, the procedures for cross-national pre-testing which I developed with colleagues, and the development of an analytical framework for the interpretation of cross-national pre-testing results which I produced, have both improved the state of the art. Questionnaire development is discussed in a working paper (Fitzgerald 2015a) and pretesting in two journal articles (Miller et al. 2011; Fitzgerald et al. 2011).

The final observational domain of TSE is 'analysis' where errors related to conceptual, presentational and statistical tasks are specified (Smith 2011). Based on two cross-national substantive papers I have co-authored, I touch upon some of the challenges that emerged related to TSE and reflect on possible implications for my methodological work. The first paper looked at acculturation amongst migrants in terms of attitudes towards homosexuality (Fitzgerald, Winstone, and Prestage 2014b) whilst the second looks at cross-national differences in traditional values (Harrison and Fitzgerald 2010).

Some TSE schemes include specification error as one of the components (Biemer and Lyberg 2003). Smith does not use this term, however one key element of what is often labelled

specification error, is located in two of the error boxes in his scheme. First, the inclusion of the correct variables in the survey is covered in "Content". That is, did the survey contain the right variables to fully specify the model? Second, it is also covered under Conceptual. That is, did the analyst include the right variables in the right way in the statistical modelling (Smith 2011). Biemer and Lyberg define the content element as ''…when the concept implied by the survey question and the concept that should be measured in the survey differ" (Biemer and Lyberg 2003, 38). Specification error is "... often caused by poor communication between the researcher (or subject-matter expert) and the questionnaire designer" (Biemer 2010, 822). In many public surveys there is no documented questionnaire specification at all (Fitzgerald 2015a) either because there is no documentation at all or it exists but it is not made available. However, it is particularly important that this process is undertaken and conducted in a cross-national context in order to ensure that concepts are not 'lost in translation' and indeed to be sure that they can in fact be measured cross-nationally. One aim of the questionnaire design template I developed was to try and prevent this error by requiring a detailed specification at the start of the design process, whilst also facilitating clear, well-documented, conceptually organised communication amongst all those involved in the development and pre-testing of the questionnaire.

Total Survey Error in cross-national perspective

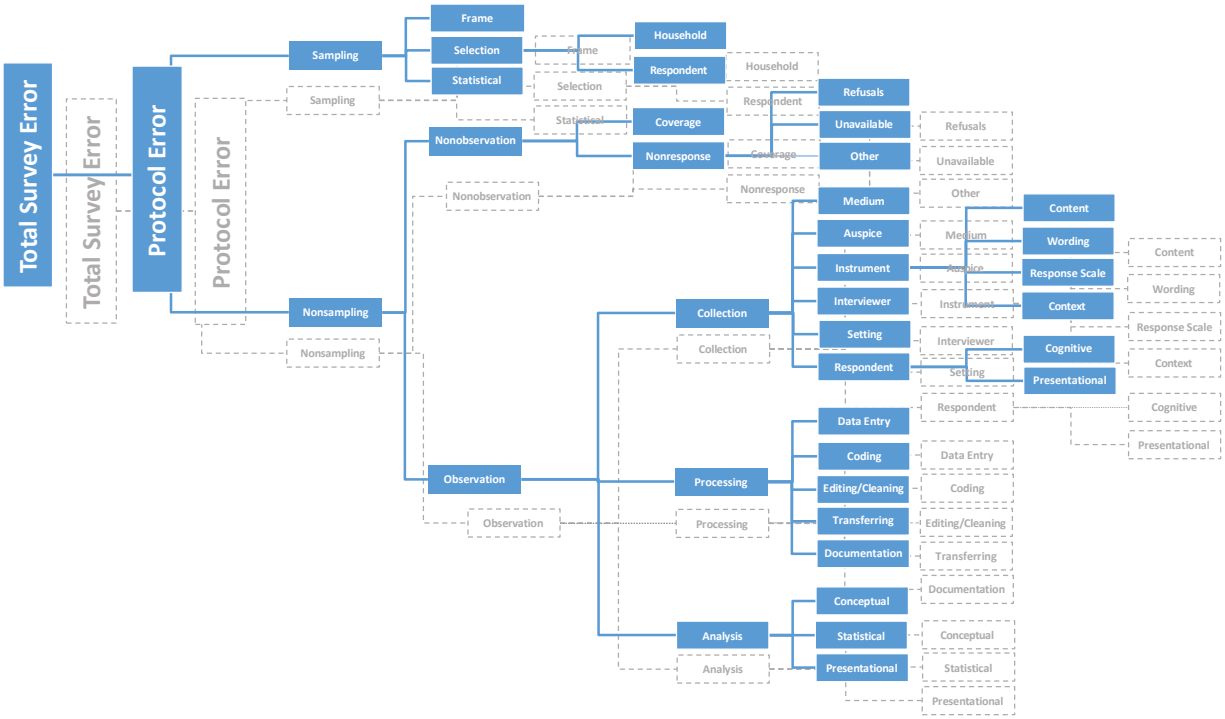
For practical reasons, and to ensure appropriate cross-national input into their design, most cross-national face-to-face interviewer surveys are effectively administered as a series of distinct national surveys organised under varying levels of centralised design, harmonisation and management. As a result of this method the possibility for error is logically multiplied by the number of national surveys. Error sources can interact with one another (Smith 2011) and so in a cross-national survey the potential for error is subsequently multiplied by the total number of possible interactions, even if not all are triggered.

Smith is one of the few scholars to discuss the potential implications for TSE when multiple surveys are being used to compare populations, which is the case with most cross-national surveys. He outlines three additional complications for TSE in this scenario. These include (a) Data collector or which organisations conducted the survey, (b) time or when the data collections occurred, and (c) target population or from whom the data are collected (Smith 2011).

The European Social Survey (ESS) exemplifies each of these complications in a cross-national context. In terms of '(a)' each country chooses its own data collector based on its abilities to conduct high quality face-to-face fieldwork using probability samples (Fitzgerald and Jowell 2010), inevitably leading to different approaches being followed despite attempts at harmonisation. Differences between data collectors, even in the same country, can in turn produce 'house effects' (Smith 2011). The impact of these is often unknown but there can be national differences in survey practices as well as differences between agencies, even in the same country. In the ESS, cross-national differences between countries due to fieldwork differences are apparent, whilst those between agencies in the same country are probably under-explored but exist nonetheless. In terms of '(b)' the ESS has had considerable difficulties arranging for all countries to start their fieldwork at the same time, as national funding and practical issues between countries led to uneven start and finishing times, possibly compromising cross-national comparability. Even in a single country study, different dates of fieldwork can lead to variations in socio-political contexts, which have the potential to impact on how respondents answer specific questions. However this tends to be more pronounced in a cross-national survey context (Stoop 2007). Smith (2011) argues that '(c)' applies most often in cross-national surveys where populations differ cross-culturally, for example in terms of language, culture, context, institutional and structural differences. Additional complications include data protection differences, varying social desirability and differing response styles. The ESS faces all of these difficulties having included over 30 countries and languages (Fitzgerald and Jowell 2010), although compared to studies with a global reach the task is undoubtedly lessened.

Figure 2 demonstrates how Smith (2011) illustrates the additional inputs into TSE from each national survey which form part of a comparative study. Whilst showing these as separate surveys, he points out that not only do the error sources within each survey have the potential to interact amongst themselves, but also that different error sources can interact *between* the various national surveys. To show this visually would be unhelpful and prove difficult to interpret for even the most visual of readers. However the approach nicely demonstrates the additional layers and complexity of the cross-national study in TSE perspective. These additional interactions are considered in my published work, for instance the Cross-National Error Source Typology (CNEST) paper which aims to disentangle sources of error discovered in cross-national pre-testing (Fitzgerald et al. 2011).

Figure 2 – Comparison Error (Adapted from (Smith 2011, 476)



Upon reflection it appears that there is an overarching part of comparison error in TSE that is missing from the various conceptualisations that already exist. In cross-national surveys there are often large, devolved teams involved in the realisation of the survey in their countries. Yet to achieve equivalence of measurement it is surely necessary for the overall study design and each individual component of the survey life cycle to be clearly described in protocols and other documents. The approach taken in the ESS has been to specify the overall design (ESS 2015) and provide detailed protocols for each step in the survey life cycle (Fitzgerald and Jowell 2010). Whilst it is surely undisputed that such an approach should be used to reduce error, the ESS was pioneering in this respect, making clearly specified protocols a key part of its *modus operandi*. My two publications (Jowell et al. 2007; Fitzgerald and Jowell 2010) discuss the importance of this approach. However this additional element of 'protocol error' is introduced here for the first time in terms of its relationship with TSE for comparative surveys. Its position is shown in Figure 2 at the overall survey level, covering sampling and non-sampling error. What is clear is that protocol error can exist both at the survey level (the PI or coordinator failing to detail the requirements adequately) and the national level (the person responsible for national implementation failing to engage with the protocols and / or produce the required national versions and adaptations). Having a protocol in itself is an essential first

step, however the key in a cross-national survey of course is also to ensure that there is high quality adherence to those protocols. For example, having multiple field agencies implement interviewer training may lead to uneven compliance and the delivery will also be impacted by the professionalism and motivation of the various staff involved.

Origins of Comparative Survey Research

Some trace the origins of survey research back to the 19th Century and in particular to Victorian England when surveys of the social conditions were conducted by Charles Booth and Seebohm Rowntree (Norris 2009). Many argue that the establishment of the social survey as we understand it today lies with the political science work of George Gallup in the USA between the two world wars (Heath, Fisher and Smith 2005) and in particular with his early attempts to predict the outcome of elections (Norris 2009)[2]. From the late 1900s and with a somewhat more sociological outlook the USA saw NORC established in Chicago, the Bureau of Applied Social Research at Colombia being set up and the Institute of Social Research at Michigan come into being (Bulmer 1998). The Survey Research Centre at Michigan, within the ISR, broke new ground being able to keep its overheads from government funded research leading to what was the sociologist Rossi reported as envy from other research organisations (Converse 2009). In her seminal book 'Survey Research in the United States: roots and emergence' Converse is clear that it was business and politics that came first in survey research rather than its initial origins being in science or academia. She also argues that the development of survey research was the most important development in social science in the twentieth century (Ibid).

Following the initial polling breakthrough by Gallup and the wider consolidation into academia, a process got underway by which this model for measuring public opinion was transported more widely, first to Europe but later to other parts of the world (Norris 2009). Heath and colleagues charted this process and described it as the 'globalisation of public opinion research', whilst noting the limitations in terms of coverage (Heath, Fisher and Smith 2005), many of which have subsequently been bridged by the establishment of the Gallup World Poll. However, notable exceptions remain, for example in North Korea where polling is forbidden or China where certain content is severely restricted (Smith and Fu 2014).

[2] One of the most inspirational moments of my own career was during a short foreign exchange, whilst I worked at the Gallup Organisation, meeting the sons of George Gallup in Princeton NJ and hearing tales of the early days of polling

Smith (2010) has identified three key phases in the development of comparative survey research. The first *adhoc* phase started with the realisation that comparative data existed and quickly included more concerted efforts to conduct 'one-time' topic-specific studies in two or more countries. Such studies became possible due to the establishment of polling companies outside of the USA, including the UK British Institute for Public Opinion established by Gallup himself (Smith 2010). The first example of a 'dedicated cross-national survey' using a common questionnaire was the 'How Nations See Each Other' study by Buchanan and Cantril in 1948, which was quickly followed by the US*I*A series surveys, many of which were international: the 1956 'International Stratification survey' by the sociologists Ganzeboom and Nieuwbeerta, the 1957 'Pattern of Human Concerns survey' by Cantril, and the 1959 ground breaking 'Civic Culture Study' by Almond and Verba (Norris 2009). The 2nd phase lasted from 1973 to 2002, with this stage representing a shift from *adhoc* cross-national studies run by a small team to more *established sustained programmes*, including researchers from a wider range of countries or formally representing a cross-national association like the EC (Smith 2010). The establishment of the Eurobarometer survey series in 1974 fore-grounded the institutionalisation of many other continuing survey programmes in this period (Heath, Fisher, and Smith 2005). A number of key cross-national social surveys were also established in the period including, but not limited to, The European and World Values Surveys (http://www.europeanvaluesstudy.eu; http://www.worldvaluessurvey.org/wvs.jsp), the International Social Survey Programme (http://www.issp.org/), electoral studies such as the Comparative Study of Electoral Systems (http://www.cses.org/) and the 'loosely related' Globalbarometers (http://www.globalbarometer.net/) . Whilst these continuing programmes marked a departure, *adhoc* studies continued and expanded, with the World Fertility Study covering 61 countries (Smith 2010). The third phase started with the advent of the European Social Survey in 2002, marking the first time there were dedicated central resources for a continuing cross-national social survey, allowing much greater focus on a harmonised methodology and quality control. Other surveys such as the Survey for Health Ageing and Retirement in Europe (SHARE) (http://www.share-project.org/), have built upon that model and introduced even greater centralised management (Boersch-Supan et al. 2010), perhaps marking the start of a fourth phase?

It is this greater centralisation and harmonisation that arrived with the establishment of the ESS, as well as the increased attention to quality in this cross-national social survey that facilitated and stimulated the work presented in this doctorate. This therefore leads to the

question as to why such a shift to centralisation was considered necessary. Why was there a need for a European Social Survey?

The Development of the European Social Survey

The European Social Survey (ESS) was formally established in 2001, following a successful application to the European Commission by Principle Investigator Roger Jowell to cover the central coordination costs, which the EC agreed to provide only on condition that a significant number of individual countries would cover their own national costs. The project was established following concerns from substantive researchers that existing cross-national studies were insufficiently harmonised, thus undermining confidence in the quality of the data. Most notably it was the difficulties faced by Kaase and Newton in the 'Beliefs in government' project which, although able to utilise the Eurobarometers, ISSP, European and World Values surveys as well as national election studies, was thwarted from drawing the required comparative conclusions due to 'discontinuities and internal inconsistencies' (Jowell et al. 2007, 2). Through subsequent discussions within the European Science Foundation a Blueprint was agreed for a European Social Survey. This established the case for the ESS as well as identifying the methodology required to learn from the past and put comparability 'centre stage'. In particular the Blueprint authors stated:

"The social sciences, if they are to make progress, require regular cross-national surveys that are conceptually well anchored, conducted according to rigorous methodological standards and are available at little cost to the entire research and policy community. Such studies must be designed for use by a broad variety of people for a broad variety of purposes. No such database currently exists in Europe, and this is the essential rationale for a regular European Social Survey (ESS)" (European Science Foundation 1999, 5).

The rationale for high quality both in terms of the need to seek equivalence and a call to be 'in defence of rigour' are outlined in 'The European Social Survey as a measurement model' (Fitzgerald and Jowell 2010) and 'Measurement equivalence in Comparative Social Surveys: the European Social Survey (ESS) – from Design to Implementation and Beyond' (Jowell et al. 2007). Together quality and rigour are the cornerstones of the ESS and underline a desire both to tackle sources of survey error and maximise comparability, despite the cross-national challenges. These barriers include one " ...obvious reason …(which is) their expense. But there are other even more compelling reasons, notably that comparative surveys have to deal with

competing cultural norms and national methodological preferences that single nation studies do not begin to face" (Jowell et al. 2007, 1).

The call to be in 'defence of rigour' was most notably manifested in the scientific case for probability sampling in all countries. In addition rigorous questionnaire design and pre-testing, cutting edge translation procedures, single mode data collection, careful fieldwork planning to a single protocol, high response rates targets, and new standards in data delivery, processing and dissemination are at the heart of the ESS project. The need for rigour was underlined by a need to minimise total survey error *per se*, whilst also maximising equivalence.

Sometimes the term equivalence in cross-national surveys is used to refer to the extent to which survey questions are comparable across nations and languages (Mohler and Johnson 2010). However we use the term more broadly to refer to the equivalence of the entire survey process across respondents within and between countries (Jowell et al. 2007). So even in a single nation study the experience of the survey should be the same for all respondents, as should the way their data is processed, in order to ensure that differences discovered reflect substantive differences and not simply methodological artefacts. Therefore the chances of selection at the sampling stage must be known and non zero, the questionnaire must broadly be understood in the same way by differing respondents, the mode of data collection should not impact on how respondents answer the questions, the coding should be rigorous and duplicated to avoid personal bias and the data processing must be transparent and consistent. Of course these issues are made more difficult in a cross-national context due to language, cultural, methodological, financial and political differences. In addition what Roger Jowell used to refer to *informally* as the difficulty of 'herding cats' (Fitzgerald 2015b) brings us full circle to the same issues that Smith identifies as being specific to comparison error.

After three rounds of the ESS we were able to conclude that the survey had made great strides in tackling sources of TSE and that deficiencies in equivalence that had persisted in previous cross-national social surveys had been addressed (Fitzgerald and Jowell 2010). The award of the Descartes Prize in 2005[3] represented external validation of these conclusions (Jowell et al. 2007). Other external commentators are generally agreed that the ESS is a model for improving

[3] The ESS was awarded the Descartes prize for 'Excellence in scientific collaborative research' in recognition of its radical innovations in cross-national surveys. It was the first social science project to be awarded the Prize. See http://ec.europa.eu/research/press/2005/pr0212en.cfm.

quality and improving equivalence, whilst correctly noting that there is a long way to go in terms of reducing and accounting for TSE, even for projects like the ESS (Heath, Fisher and Smith 2005; Norris, 2007; Smith 2009; Smith 2010).

Against this background the mission of the ESS, to 'rectify longstanding deficits in the rigour and equivalence of comparative quantitative research, especially in attitude studies' (Jowell, Kaase, Fitzgerald and Eva, 2007; 9) provided me with fertile ground to work with others both within and outside the immediate ESS to rise to this challenge. In the remaining chapters I will draw on my published work to illuminate my contribution in tackling sources of TSE. Drawing on published work from 2007 onwards I will highlight my particular focus on comparison error. As a data producer this has primarily meant a focus on reducing error *before* it occurs but it has also touched upon methods for correction and reflections on TSE made during analysis.

Specifically the following issues will be addressed:

Chapter 2 will indicate my efforts to measure and understand non-response bias in a comparative perspective. Specifically the non-sampling, non-observation, non-response dimension of TSE will be explored (Stoop et al. 2010; Billiet et al. 2007).

Chapter 3 will show how I contributed to minimising error in questionnaire design. In particular specification error on the ESS was tackled through the design and implementation of a questionnaire design template. A working paper which describes and evaluates the template will be introduced (Fitzgerald 2015a), whilst papers which show examples of the outputs of ESS questionnaire design structured via the template will be discussed to highlight its benefits. One of these papers is formally submitted (Winstone, Widdop, and Fitzgerald 2016) .

Chapter 4 will examine how I reduced instrument and respondent error in a comparative perspective. Specifically it will address how, with colleagues from the USA and across Europe, I developed a methodology for effectively implementing cognitive interviewing cross-nationally. A journal article submitted as part of the thesis will be used to highlight this contribution (Miller et al. 2011) In addition, it will first show how I developed a typology to allow the identification of different error sources identified during cross-national cognitive interviewing and pre-testing more generally. A journal article which introduces and discusses the typology will be introduced (Fitzgerald et al. 2011).

Chapter 5 will introduce the overall procedures for pre-testing on the ESS, which I further developed from more modest beginnings. In addition to discussing this innovation, the benefits of triangulation of pre-testing in a cross-national study will be explored. A working paper will be introduced that highlights the benefits of triangulation in cross-national perspective (Fitzgerald, Winstone, and Prestage 2014a).

Chapter 6 will very briefly introduce 2 substantive analytical pieces that I co-authored, highlighting the influence of methodological limitations and their impact on analysis. A book chapter and journal article will be introduced (Fitzgerald, Winstone, and Prestage 2014a; Harrison and Fitzgerald 2010).

Chapter 7 will make some concluding remarks about my contribution to reducing error in cross-national surveys. I critically evaluate that contribution and identify my future research agenda.

Chapter 2: Survey non-response in comparative perspective

This exegesis now moves on to a discussion of two contributions which looked at efforts to minimise, measure and analyse non-response to cross-national social surveys. The submitted publications 'Survey nonresponse in Europe: Lessons from the European Social Survey' (Stoop et al. 2010) 'Estimation of Response Bias in The European Social Survey: using information from reluctant respondents in Round One' (Billiet et al. 2007) were among some of the first attempts to evaluate cross-national survey response using individual level paradata. Such analysis and reflection became possible due to the specification of common definitions by the ESS and provision of harmonised data across the large number of countries included in the study. This provided a unique and novel opportunity that has furthered knowledge about nonresponse in cross-national surveys. The ESS approach to nonresponse is a clear example of its approach to achieving equivalence of measurement and promoting rigour (Jowell et al. 2007). It also demonstrates the effort put into understanding this potentially very damaging element of TSE.

The approach on the ESS has facilitated an in-depth examination of cross-national survey nonresponse processes and outcomes, as well as allowing methods for subsequent detection and correction of nonreponse bias to be implemented (Stoop et al. 2010). For example, one approach has been to take information from those who finally agreed to cooperate with a survey request, but who had initially been reluctant to do so, and compare them to those who cooperated earlier in order to estimate nonresponse bias. The impact of including such respondents in the survey and potentially using them as the basis for estimating nonresponse bias have then been examined (Billiet et al. 2007).

Nonresponse Theory and TSE

It is generally accepted that non-response to surveys is a major component of Total Survey Error (Groves and Lyberg 2010; Biemer 2010; Smith 2011). In addition it is a component of TSE that is fully measurable and, at least theoretically, fully correctable. It therefore correctly receives a significant focus in the survey literature (ibid), even if this is at the expense of other dimensions of TSE which are less explored or harder to measure. In addition evidence that response rates have been declining, or have been maintained only with significant additional effort, have also led to an increased focus on nonresponse issues (Couper and Leeuw 2003; Stoop et al. 2010).

Target respondents who either never received a request to take part in a survey, were unable to take part in it or were unwilling to do so, *can* differ systematically from those who did take part in the end. Where this has occurred there is a serious possibility that the conclusions drawn from the survey statistics might be biased (Groves and Couper 1998). Groves and Couper developed a conceptual framework which identifies three main types of survey nonresponse: contactability, ability to participate and willingness to participate.

The smallest group of nonrespondents in face-to-face surveys are normally those who are unable to participate. The next smallest group are those where the barrier to participation is contactability, meaning that target respondents never received a request to take part in the survey (or at least one that actually enables them to do so easily). The final group are normally the largest of the nonresponders and include those who decline an opportunity to take part. Groves and Couper (1998) reassure researchers that where such nonresponse occurs at random there is little to fear, as this will not bias estimates, although they note that having fewer respondents overall will naturally reduce precision in the estimates. The more serious point of concern is where those missing are not missing 'completely at random', which reduces the ability of the survey data collected to represent the universe from which the sample was drawn (Stoop et al. 2010). For example if a general population survey measuring attitudes towards immigration has high refusals from younger target respondents, who in turn are more likely to be positive toward immigration, then the overall population estimate *may* be biased.

So how does nonresponse impact on estimates in a cross-national survey? As noted in Chapter 1, most cross-national surveys are essentially a series of separate national surveys, conducted under a harmonising framework. So in terms of 'comparison error' (Smith 2011, 475), it is instantly apparent that there is the clear potential for the type and level of non-response between countries to differ. If such differences were to manifest differently between countries, or over different time points within a country in a repeat survey, or both, then the potential for spurious conclusions to be drawn increases markedly. It is of note that the main way in which the Groves and Couper model is adapted for cross-national application in the 'Improving Survey Response book' (Stoop et al. 2010), is to down play the distinction between factors which are and are not under the control of the researcher. Reflecting the rather federal approach to implementation of cross-national surveys, it is clear that far less is under the control of 'the researcher' in terms

of a single Principle Investigator and that 'house effects' must also be taken account of (Stoop et al. 2010, 17).

Taking the ESS as an example, one can use the three main categories of nonresponse outlined by Groves and Couper (1998) to highlight some of the additional challenges faced in a cross-national context. The category 'unable to participate' is normally the smallest category of nonresponse. It tends to include those who are ill, who cannot participate due to language barriers or who have a disability that prevents participation, for example being unable to hear. Normally this would be a very small proportion of the total issued gross sample and, in the overall scheme of TSE, could arguably be given far lower priority in the overall spectrum of error. However in a cross-national study this could be a source of more notable error.

Those unable to participate due to language barriers: this not only highlights cross-national differences but also shows different sources of TSE colliding with one another and competing for attention. The ESS tries to be inclusive of those speaking minority languages, in order to be as reflective as possible of the sample universe (all adults aged 15+ and resident in the ESS country). However the ESS is also mindful of the need to implement standardised interviewing in order to minimise interviewer effects on the data collection, another potentially very damaging element of TSE. The ESS therefore insists on full translations of the questionnaire into each target language and forbids 'live' translations by bilingual interviewers (which if allowed would essentially lead to a somewhat different translation each time and therefore compromise standardised interviewing principles) (Harkness 2007). However in order to limit costs, translated versions are required only for minority languages spoken by at least 5% of the population as a first language (ESS 2015). This not only reduces translation costs themselves but means that tricky and expensive scheduling of interviewers able to speak minority languages is not required for smaller groups. Table 2.1 shows that in the ESS language barriers are quite a small factor in terms of preventing participation. However it also highlights how this differs cross-nationally. In four countries – Hungary, Poland, Russia and Slovakia - this ground for exclusion does not apply *at all*, whilst in Sweden, Switzerland, Cyprus and Iceland 2% or more of issued cases are excluded on these grounds.

Table 2.1 Language barrier as proportion of total issued sample ESS Round 6

Country	Lang barrier	Gross sample	%	Country	Lang barrier	Gross sample	%
Hungary	0	3194	0.00	kosovo	10	2312	0.43
Poland	0	2706	0.00	Spain	14	2868	0.49
Russia	0	3772	0.00	UK	23	4520	0.51
Slovakia	0	2500	0.00	Israel	17	3230	0.53
Portugal	1	3040	0.03	Finland	41	3296	1.24
Ukraine	2	3692	0.05	Germany	138	8904	1.55
Estonia	4	3707	0.11	Ireland	77	4420	1.74
Denmark	4	3372	0.12	Norway	55	3041	1.81
Slovenia	3	2250	0.13	Belgium	74	3267	2.27
Lithuania	6	4470	0.13	Netherlands	91	3537	2.57
France	8	4200	0.19	Sweden	99	3750	2.64
Bulgaria	7	3200	0.22	Switzerland	83	2907	2.86
Albania	4	1602	0.25	Cyprus	63	1589	3.96
Czech R	10	3010	0.33	Iceland	76	1431	5.31
Italy	11	2778	0.40				

Source: ESS Round 6 Data Documentation Report

This very straightforward example shows us that this source of nonresponse applies differentially across countries, whilst also highlighting how the response angle 'collides' with the measurement domain and the need to ensure standardised interviewing is implemented. All of this has to be considered within the budget available, highlighting the competing demands placed on the cross-national researcher within the TSE framework.

The next main category of nonresponse, identified by Groves and Couper (1998), is 'contactability'. In face-to-face surveys this relates to the target respondent receiving a request to take part in the survey directly from the interviewer. In most studies there is a much smaller proportion of nonresponse from contactabilitythan from 'willingness to participate'. However due to the specific types of target respondents excluded when they are not contacted, there are particular concerns to minimise this source of error where possible (Stoop et al. 2010). The ESS has tried to deal with this problem by specifying that nonconatcts should not exceed 3% of the total issued sample (ESS 2015). In terms of TSE, and in particular comparison error, the same issue with noncontacts arises as with the language barriers discussed above. Levels of noncontacts differ considerably cross-nationally, leading to the strong possibility that bias will differ and the comparative basis of the data will be undermined. In ESS Round 3 for example, noncontact rates ranged from 0.8% to 13.1%, with 10 countries having noncontact rates over the 3% maximum target (Stoop et al. 2010).

The final area of nonresponse identified by Groves and Couper (1998) is 'willingness to participate', with nonresponse identified in face-to-face surveys with a direct refusal to take

part in the survey, usually to the interviewer themselves (or perhaps by contacting the field company directly after receiving an advance letter). This is normally from the target respondent but can be from another household member. As noted earlier this is usually the largest category of nonresponse and there are extensive theories and studies including methods to try and correct for this result (ibid). The ESS sets a minimum target response rate of 70% in order to maximise efforts to secure cooperation in all countries (ESS 2015), despite awareness that this is difficult to achieve in many countries (Stoop et al. 2010). Taking ESS Round 3 as an example, the response rates varied between 46%, in France, and 73% in Slovakia (Stoop et al. 2010). However across later rounds there have been response rates lower and higher than the minimum and maximum in Round 3, including significant variation from particular countries over time. What is instantly clear, however, is that there are great differences *between* countries in the level of cooperation obtained, suggesting a high potential for differences in bias.

From a TSE perspective these findings from the ESS indicate that nonresponse patterns are quite different across parts of Europe. However nonresponse is only a source of error if it leads to bias and only a source of comparison error if there are different levels of bias cross-nationally. So the question that ultimately needs answering is whether a survey has nonresponse bias and whether this differs cross-nationally.

In order to answer this question, it would of course be necessary to have accurate paradata from a range of cross-national surveys and countries. In addition a harmonised approach to the measurement and reporting of response and nonresponse would be needed alongside comparisons with other data, such as benchmarks from official statistics. However, when the ESS was established similar cross-national, cross-sectional social surveys were not generally calculating response rates in a uniform or transparent manner. This made meaningful within and across country comparisons impossible or unreliable (Stoop et al. 2010).

As part of its mission to improve standards of cross-national measurement (Fitzgerald and Jowell 2010), the ESS set out to implement a system of harmonised response rate measurement that could apply across all participating countries. Although the ESS is seen as an input harmonised study (see Chapter 1), this cannot mean identical procedures are used in all countries at every stage in the survey life cycle (Jowell et al. 2007). This is clearly demonstrated when it comes to sampling: here each participating country in the ESS has little choice but to use the best available sampling frame in order to draw their probability sample (Häder and

Lynn 2007). The alternative would be to try and implement the same design in every country, requiring area based sampling with random route address selection. However this would greatly decrease quality and increase the costs. The advice of Kish is salient here:

"Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods: known probabilities of selection for all population elements" (Kish 1994, 173).

However in order to be able to measure nonresponse comparatively across different sampling designs, an equivalent yet not identical method of recording every contact attempt needed to be established. Chapter 3 in '*Survey nonresponse in Europe: Lessons from the European Social Survey'* (Stoop et al. 2010) discusses how the ESS has managed to operationalise a process for doing this, regardless of whether a country is using a sample of named individuals, addresses or households. Specifically, ESS contact forms were developed to enable every contact attempt with every selected sample unit to be recorded by interviewers. Regardless of the sample type that contact data could then be analysed to enable a final disposition code to be assigned and allow direct cross-national comparisons of response rates across Europe. This innovation nicely illustrates the ESS approach to achieving equivalence of measurement cross-nationally, balancing the need for harmonisation with national specifics.

The contact form data collected has enabled conclusions to be drawn about nonresponse in a comparative perspective, with many of these possible for the first time. Although largely limited to the ESS, the findings broke new ground in terms of how we understand nonresponse cross-nationally. Some of the key findings from '*Survey nonresponse in Europe: Lessons from the European Social Survey'* (Stoop et al. 2010) were as follows:

- Noncontact and cooperation rates differ significantly across ESS countries, however it is unclear how much this is a country effect as opposed to a data collection expertise effect;
- Response rates can be improved with changes in fieldwork strategy between rounds;

- Response rates in some countries have declined over time but in others have risen, with a narrowing of the gap between the highest and lowest detected between Rounds 1-3;
- Differences in noncontact and cooperation rates do not necessarily reflect differences in fieldwork efforts across ESS countries;
- Some countries need significantly more fieldwork effort to achieve the same nonresponse outcomes as other countries;
- The best time to make contact with target households / individuals differs significantly across countries reflecting different 'at home patterns';
- The ESS rules that state all sample units must receive at least 4 contact attempts including 1 in the evening and 1 at the weekend, improves response rates considerably in most, but not all, countries;
- Refusal conversion, at least as far as it has been implemented in the ESS to date, does increase the response rate overall but appears to do little to minimise nonresponse bias, possibly even making this worse;
- Applying a range of techniques to try and detect and correct for nonresponse bias, little evidence of its existence was found on the ESS. However, further research on the ESS and other cross-national surveys is now required in order to replicate and further develop that testing.

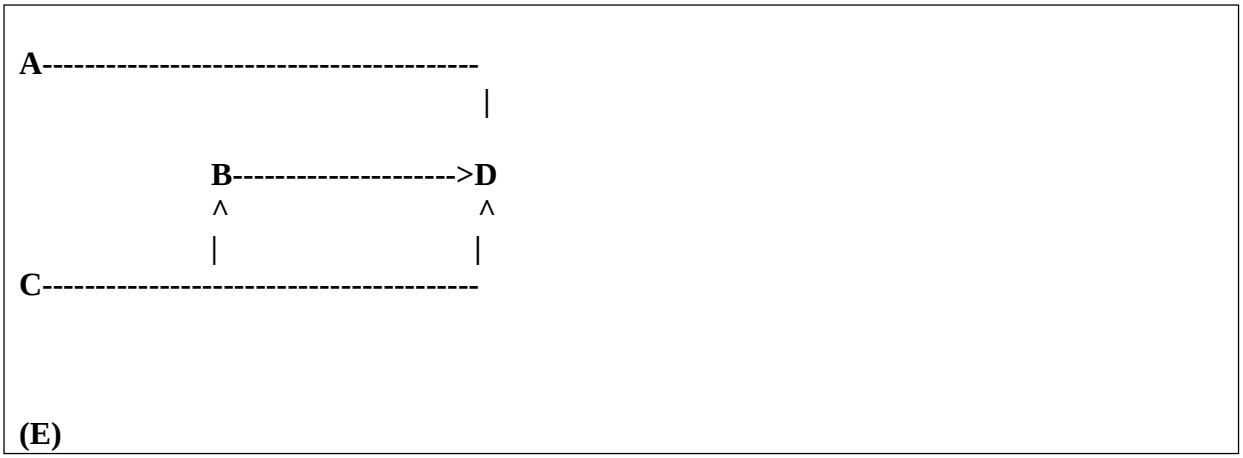
Data from the ESS contact forms has also been used to look for 'traces of bias' and compare these cross-nationally, by comparing those who were reluctant to respond to those who responded more immediately (Billiet et al. 2007). Rather than looking just at comparisons of means or proportions, explanatory models have been used to test the effect of the kind of respondent on those models. Differences that had been observed with simple statistics largely disappeared, although a few remained. In the two countries (Netherlands and Germany) where multivariate analysis was possible due to sufficient sample size, easy-to-convert refusals did not have a significant effect and hard-to-convert refusals mattered on just two of the six measures. Furthermore the remaining (significant) effects were small and did not have serious implications for the parameters of the substantial explanatory variables. However, caution remains advisable because of the (very real) possibility that final refusals may differ from converted refusals (ibid).

The harmonised approach to response measurement used on the ESS and applied in very different contexts, has facilitated direct comparisons between countries in terms of nonresponse which was largely impossible in the past. The availability of this data has not only highlighted differing nonresponse patterns in Europe, but has also led to improvements being implemented over time in ESS countries. The ESS approach of setting nonresponse targets, specifying minimum efforts, along with a workable way to measure outcomes, has furthered understanding of this source of error in a cross-national perspective and improved practice in survey research. And whilst a centrally determined 'responsive design' during fieldwork remains outside the organisational possibilities of the current ESS, the ESS approach to non response measurement has facilitated a responsive design between rounds (Stoop et al. 2010). This illustration from the ESS shows how a source of TSE is magnified in a cross-national study and highlights the interaction between sources of error in that multi-national environment.

Chapter 3: 'The challenge of Cross-national Questionnaire design'

The previous chapter examined the issue of non-response bias as a form of comparison error. This chapter will demonstrate efforts made to reduce the impact of 'specification error' in (cross-national) surveys, by developing improved questionnaire design documentation for the European Social Survey. Specification error is a major source of TSE (Biemer and Lyberg 2003; Biemer 2010), and can seriously threaten the utility of a survey.

Figure 3.1 Model showing specification error



Consider the model in Figure 3.1 in which A, B, C and E predict D. The model would be misspecified if:

(1) E was absent from the theoretical model and omitted from the questionnaire,

(2) E was in the questionnaire, but failed to match or adequately measure the theoretical concept or

(3) E was in the questionnaire and suitable, but not used in the analysis.

These are types of specification error but there may be others. This chapter will discuss efforts to minimise specification error in the ESS, in particular addressing (2) but also (1), using a specially created questionnaire development template which was created to ensure a conceptually driven approach to design (Fitzgerald 2007; Fitzgerald 2015a). The chapter will demonstrate two examples of how this template has influenced the design of two ESS rotating modules: 'Trust in the Police and Courts' (Jackson et al. 2011); and 'Understandings and Evaluations of Democracy' (Winstone, Widdop, and Fitzgerald 2016). The chapter will also discuss a related project that aims to develop a questionnaire design database, based on this

ESS template, with links to a translation and question variable databank (Prestage, Knut, and Fitzgerald 2015) .

Specification Error

Specification error is a very serious source of TSE and one that receives insufficient attention, perhaps because it is more challenging to identify and measure than other sources, for example nonresponse error (see Chapter 2). It occurs when '....the concept implied by the survey question and the concept that should be measured in the survey differ" (Biemer and Lyberg 2003, 38) and is "... often caused by poor communication between the researcher (or subject-matter expert) and the questionnaire designer" (Biemer and Lyberg 2003, 822). Gideon sees specification error as when the theoretical concepts, constructs and variables are not aligned, leading to serious measurement problems. It can even mean the complete exclusion of key concepts from the questionnaire. This often stems from "...an ill-defined research proposal or the lack of a research proposal at all" (Gideon 2012, 40).

Whilst Gideon (ibid) assigns this lack of specificity primarily to junior researchers, my own experience is that this remains an issue for more experienced scholars too. It was in fact experience of questionnaire design on the ESS, working with senior social scientists, which provided the inspiration for developing the ESS questionnaire design template, rather than inspiration from the theory of TSE.

I joined the ESS Core Scientific Team (CST) as the Round 2 questionnaire was being finalised in 2003. It was just prior to the ESS source questionnaire being sent to national teams for translation. Each round of the ESS has 'guest' rotating modules designed by cross-national teams of leading academics who are substantive experts in their field and who *may* also have methodological survey expertise (Fitzgerald and Jowell 2010). Faced with only a short handover period when I joined the ESS before I had to assume responsibility for finalising the Round 2 questionnaire, I was struck with the lack of documentation about the design of the questionnaire. The only documentation apparently available was a string of e-mails from multiple authors, spanning an 18 month period. Whilst that was perhaps not a problem for those who had been involved with the design throughout, for a new researcher it posed a serious challenge when there was little time to try and retrace the development of specific concepts or questions. Furthermore, it quickly became clear that secondary data users, for whom the ESS was established, would not have access to these e-mails and would only have had the original

proposal and final questionnaire available to them. They would then have had to try and link the final items to the original concepts proposed.

During experience of questionnaire design in ESS Round 3, I became aware of another deficiency in the process. When discussing the module design without members of the questionnaire design team present, the question would often be posed to me 'But what are they actually trying to measure?' a question I was often unable to answer. So whilst it might have been clear that a question was related to a specific concept, the more detailed reasons for operationalising a concept with a specific item or set of items was rarely apparent and not clearly documented.

So at the 2007 meeting of the Comparative Survey Design and Implementation workshop (CSDI), I first presented my ideas for a questionnaire design template aimed at structuring and documenting questionnaire design on the ESS (Fitzgerald 2007). The idea was well received, both from a documentation and questionnaire structuring perspective, and following further development was implemented for the first time when designing the ESS Round 4 rotating modules. The template was a response to calls for documentation and transparency on cross-national surveys (Mohler, Pennell, and Hubbard 2012), and the requirement to have a theoretically grounded research design behind survey questions (Gideon 2012).

Documentation of the Questionnaire Design Process

The development of the ESS rotating modules from Rounds 4-7 is now recorded in the ESS Questionnaire design template, which are made available from the ESS website as the data is released. To my knowledge, thus far, the ESS is the only cross-national social survey programme to present (and probably even record) such detailed information about the design and development of its questionnaire. Each module is developed over an 18-24 month period and there is an iterative process of design, expert review, pre-testing, multi-national input, piloting, advance translation and reliability prediction (see Chapter 4), all of which is included in the template. The 'discussion' between the question module design team and the methodological experts of the Core Scientific Team is recorded, and the process of design is itself structured by the document. The design of the template and examples of how it structures the process of design have been published (Fitzgerald 2015a).

Structuring the Design Process

Documenting the design process helps to make it transparent for end users and allows them to evaluate quality. This is particularly important in a cross-national survey where mis-specification may interact with comparison error in wording. For example, if the concept is not clear it would increase the likelihood of a translation error. However simply recording the process does not necessarily help to prevent specification error. The rationale behind asking specific questions was not always made clear in the past, leading to errors in the operationalisation. This problem can be compounded in a cross-national survey, where translation teams need to understand the nuances of items that are clear in the source language but remain context specific for non-native speakers. The ESS questionnaire design template ensures that the failure to have a specified research design (Gideon 2012) is avoided (or exposed), since substantive specialists are required to clearly outline their theoretical approach or model. They are then asked to list the individual concepts they wish to measure, breaking these down into those that are simple concepts and can be measured directly with a single item, and more complex concepts that reflect either latent or multifaceted concepts that require multiple measures. The next step is to describe how the concept will actually be measured within the space available. This part of the specification often proves the most difficult and can lead to the concept itself being redefined. It requires both a specification of the areas to be covered and then the drafting of an item or set of items that tap this concept. Ideally the item drafting takes place only after the conceptual structure and descriptions are clear, although often the process is rather more parallel. This step by step approach to questionnaire design helps to ensure that the process of design itself is conceptually structured, avoiding a focus purely on the question items themselves.

Examples of the influence of the template on ESS module design

The template has led to new levels of transparency in cross-national questionnaire design. Jackson and colleagues, writing about the ESS Trust in Justice module, comment that the template requires module designers to outline the "...substantive measurement aims of the module, its theoretical framework and (to) identify the key concepts and dimensions. This template is then used throughout the questionnaire design process to document decisions and to ensure that the process is continually informed by the agreed measurement aims of the module" (Jackson et al. 2011, 271). The paper "Developing European indicators of trust in justice" (not submitted) describes the aims and development of the ESS 'Trust in Justice' module and in part reflects the structure of the template, for example making a distinction

between simple (level 1) and complex (level 2) concepts (ibid). At the formal launch of ESS ERIC, Mike Hough, a key member of the QDT, commented that the ESS process of questionnaire design had caused him to think harder about questionnaire design and noted that the ESS procedures were more robust than any other similar process he had undergone before.

This chapter now moves on to discuss another example of the influence of the template on the design of an ESS module examining understandings and evaluations of democracy (Winstone, Widdop, and Fitzgerald 2016). In this case the template was able to help minimise several types of TSE including 'Content' and 'Design' issues. The chapter on the measurement aims of the democracy module drew heavily on the questionnaire design template, which provided the only complete record of the design process. The template helped to make the conceptual structure of the 'meanings and evaluations' of democracy module explicit, with a number of concepts to be measured including: accessibility and equality of the judicial system, forms of participation, freedom of press, viable opposition, horizontal accountability, a particular minority in society, opportunities for effective participation and type of electoral system, subjects of representation and efficiency. Taking viable opposition as an example, the template records a detailed description of the item. It states: "The opposition must be viable. It must be able to effectively oppose the governing party, to avoid the tyranny of the majority". It then goes on to show the item (E4) that will be measured to use this, enabling a face validity check on whether the item taps the underlying concept.

CARD 37 Using this card, please tell me how important you think it is for democracy in general...**READ OUT...**

		Not at all important for democracy in general										**Extremely important for democracy in general**
E4	...that opposition parties are free to criticise the government?	00	01	02	03	04	05	06	07	08	09	10

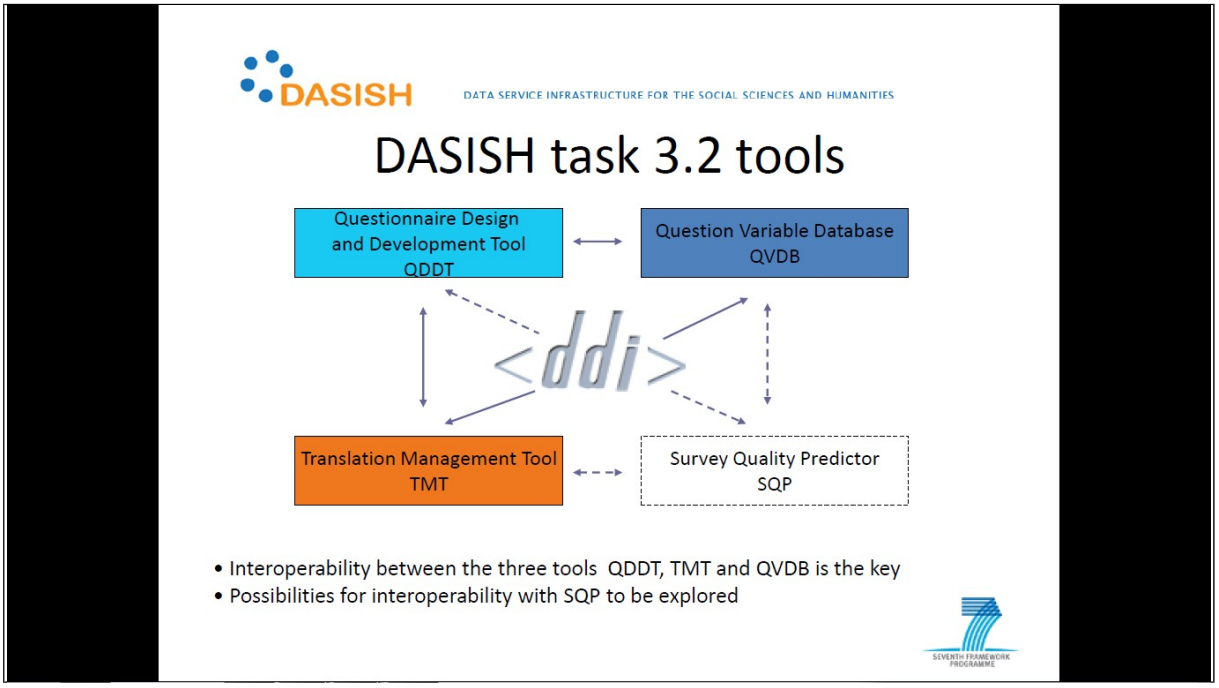
The template also kept a record of key design decisions and linked these back to the original measurement aims. These included the choice to use the word 'importance' when evaluating democracy, the choice of whether to use list wise or pair wise administration of items on the same concept but measuring different dimensions (importance for democracy and evaluation for democracy) and methods to overcome social desirability, amongst other decisions. In addition the template notes that some concepts could not be measured and provides reasons for this. For example, an item aiming to evaluate the ease of immigrants obtaining voting rights was dropped due to low awareness of this issue amongst the general population.

Developing the template into a database

One of the major barriers to using the questionnaire design template efficiently in the ESS, and of its uptake by other survey programmes, has been the paper based nature of the template (Fitzgerald 2015a). However efforts have been underway to transform the template into a searchable tool and database that records the information electronically and allows more flexible use of the information during the development process. This work was funded as part of an EU funded social-science humanities cluster project 'Data Infrastructure of the Humanities and Social Sciences' (DASISH), with further work being funded through the follow-up social science cluster project 'Synergies for Europe's Research Infrastructures in the Social Sciences' (SERISS). The tool will eventually form one part of a suite of tools for cross-national social surveys, incorporating a Questionnaire Design Documentation Tool (QDDT), a Translation Management Tool (TMT) and a Question Variable Data Base (QVDB). It is hoped that together these tools will finally allow the "...ultimate prize for large-scale cross-national surveys...(enabling) the end data user (being able) to trace the development of a questionnaire item from the original design right through to the (translated) finally fielded item" (Fitzgerald 2015a, 13) to be realised.

The aim of the QDDT is to create a multi-language questionnaire development tool to facilitate the development, translation and documentation of the concepts and questions used in cross-national surveys whilst also producing searchable metadata for the whole process. Figure 3.1 illustrates the aim of linking the various databanks, with links to the three new tools, the priority and possible links to the SQP databank to be explored later.

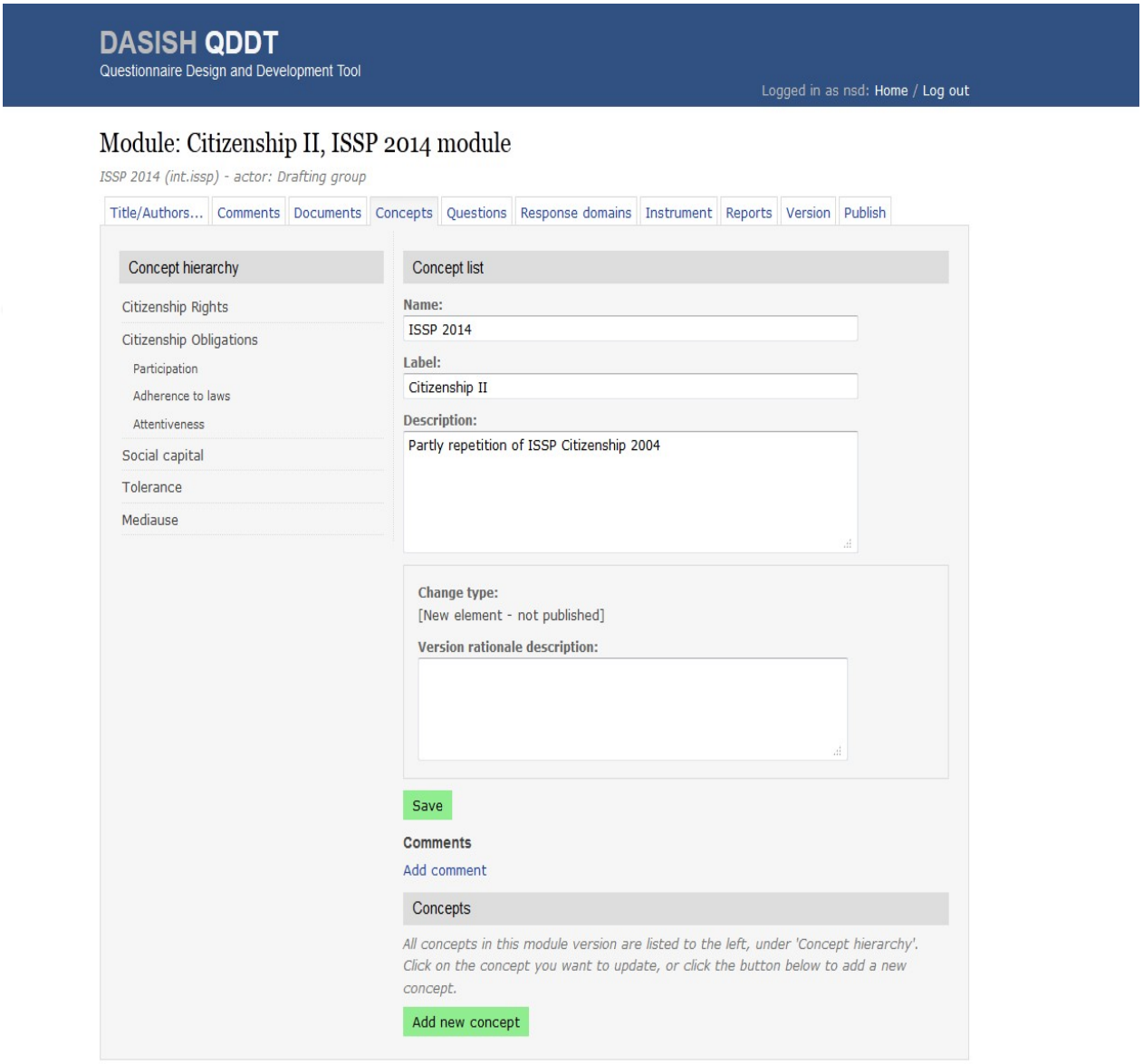
Figure 3.1 Slide from CSDI conference presentation showing links between tools (Prestage, Knut, and Fitzgerald 2015)



In the centre of Figure 3.1 the letters DDI (Data Documentation Initiative) appear. This highlights how the DASISH project aimed to design the tools according to this near universally accepted language for documenting surveys. Whilst this will help to ensure the tools have metadata interoperability, it has considerably extended the development process time compared to developing survey specific tools. At the same time using DDI will help the three tools to 'communicate' with one another, in turn creating a linked suite of tools.

Figure 3.2 shows how the database mirrors the ESS questionnaire design template. Populated in this example with data from the ISSP, one can see the conceptual basis of the design with links to proposed questionnaire items. Versioning control will allow the development of a specific item to be quickly resurrected, whilst field functionality will allow a PAPI version of the questionnaire to be produced.

Figure 3.2 Slide from CSDI Workshop presentation showing example of concept description (Prestage, Knut, and Fitzgerald 2015)



In addition to the QDDT it is hoped that in the next few years we will create a searchable database which researchers can utilise to access final questions fielded in all original languages, the concepts and the variables (QVDB). This will also be linked to a translation tool, building on work conducted here by the Survey of Health Ageing and Retirement in Europe (SHARE) with their Translation Management Tool (TMT). That database facilitates the translation in each country and then feeds it into the single CAPI programme for use in all countries.

Through the SERISS project (www.seriss.eu) it is hoped that other survey programmes will utilise these tools and in particular undertake activities to reduce specification error and make the specification that is undertaken more explicit and transparent.

The ESS questionnaire design template is a novel and relatively effective mechanism for trying to prevent specification error. On its own it cannot prevent concepts being described incorrectly or misunderstandings occurring. However, by making the process transparent and providing a single source of communication, it greatly enhances the possibility of effective co-working between substantive researchers and methodologists. It also helps as a guide to assessing the outputs from pre-testing and provides a space for triangulating pre-testing findings. This will be discussed further in Chapter 4.

The ESS also has mechanisms for evaluating the quality of its questions based on its SQP database, which provides a quality indicator for individual items and evaluations of the quality of the conceptual structure and cross-national equivalence (see www.europeansocialsurvey.org). Discussion of these processes is outside the scope of this thesis but they point to a holistic approach to quality that starts with specification and ends with quality assessment. This is part of the call for a 'defence of rigour' in cross-national survey research (Jowell et al. 2007, 4) that has informed the design and development of the ESS.

In 1986 Converse and Presser complained that with pre-testing '...(the) background behind certain concepts, (information) about why certain new questions took the form they did, about why certain well tried questions from other surveys were preferred to others (was) rarely made clear' (Converse and Presser 1986, 51). It is now made clear by the European Social Survey.

Chapter 4 Instrument and Respondent components in TSE

The previous chapter outlined ways to ensure that the measurement aims for a study have been specified adequately and at a sufficiently detailed level to prevent some of the key elements of specification error. The next task is to design an instrument that can effectively capture data that meets those aims. A poor questionnaire instrument, which impairs the ability of respondents to process it effectively or which does not meet the specified measurement aims, can be a very serious source of error. The ability of a standardized instrument to measure with reliability and validity is essential in order to capture comparable data across all respondents. Arguably, whilst error stemming from deficiencies in the representatives of the achieved sample can be corrected *posthoc,* it is far more challenging and extremely rare with measurement error[4]. This is because the true value the survey is trying to capture is rarely known (hence the need for the survey!) and individual scores for respondents are almost always unavailable, making reliability checks impossible. One clear exception is political opinion polls which are used to predict the results of elections and which highlight clearly whether the outcome has been accurately measured, if conducted immediately before polling takes place. In any event it is clearly critical to ensure that the questionnaire can deliver on the measurement aims of the study and that sufficient resources are devoted to this element of the surveys' design.

Most contemporary cross-national studies tend to use closed questions in their instruments (Heath, Fisher, and Smith 2005) and therefore discussion here is limited to this style of questioning. Smith breaks the instrument component of TSE down into content, wording, response scale and context, whilst noting that respondents need to be able to deal with the instrument in both cognitive and presentational terms (Smith 2011). All of these components need to be considered together when designing a questionnaire instrument. With interviewer administered surveys these components can also interact with the role of the interviewer.

In order to reduce the likelihood of instrument and related respondent error, it is considered best practice to thoroughly pre-test survey instruments, including directly with a sample of (test) respondents, prior to undertaking the actual data collection (Converse and Presser 1986). A range of pre-testing resources are available to survey researchers including structured and

[4] The ESS does in fact attempt this through its MTMM experiments and related SQP programme. See www.europeansocialsurvey.org

unstructured expert review, cognitive interviewing, reliability and validity prediction, behavior coding, test-re-test studies, eye tracking, large-scale quantitative piloting (with analysis) and respondent and interviewer debriefs. Each method offers something different in terms of helping the researcher to assess the likely reliability and validity of the instrument and to tackle different components of instrument and respondent error. For instance, eye tracking can identify potential presentational challenges for the respondent when they are navigating a self-completion questionnaire, whilst a large scale quantitative pilot can produce data allowing the effectiveness of the answer scales to be evaluated. The focus of this chapter is one of these methods - cognitive interviewing.

Approaches to instrument design

There are two main approaches to cross-national instrument design (Harkness et al. 2010). The first is to design an instrument in a single language source version and once that version is finalised use it as the basis for translation into the various target language versions (sequential model). The second is to design instruments simultaneously in all the target languages. The first model is the norm in cross-national surveys, in part due to the cost and logistical demands of the second model. In this chapter the first model is considered since this is the approach used by the European Social Survey (Harkness 2007).

The second key consideration is whether to use an Ask the Same Question (ASQ) or Ask a Different Question (ADQ) approach (Harkness et al. 2010). This essentially boils down to whether to use the same questions in every version – adapted only as absolutely required to make a meaningful translation – or whether to ask different questions in the various language versions and countries, whilst still attempting to tap the same concept. This can range from including completely different questions in every country, with *posthoc harmonisation*, through to adding some additional country specific questions to measure dimensions of a concept that are not universally present (Heath, Fisher, and Smith 2005). Most cross-national survey programmes use an ASQ approach as the default in order to provide data analysts with the same data structure for all countries. Notable exceptions apply to concepts that cannot reasonably be measured with an ASQ approach, such as measuring the highest level of education. The ASQ approach will be considered here, the method predominantly used on the ESS.

In terms of developing and pre-testing a source questionnaire the aim is to produce an

instrument that can effectively serve a dual function: it should be an effective instrument in the source language and context whilst also providing a guide for translation into every target language version. So in the ESS the final source questionnaire must be ready to field in the UK (as it is developed in British English) and then guide translation into 25 or more different language versions. It must also try and prevent every source of measurement error related to the instrument and the respondent in each target version. The development of a high quality source questionnaire is therefore a very challenging task and considerably more burdensome the more countries, languages and cultures it has to function in.

In early rounds of the ESS the source questionnaire was developed and tested using expert review (including from researchers in every country), predictions of reliability and validity using SQP and a 2-nation pilot (Fitzgerald and Jowell 2010). This Chapter now moves on to discuss one key innovation – the addition of cognitive interviewing to that process.

Cognitive interviewing

Beaty and Willis define cognitive interviewing as "...the administration of draft survey questions while collecting additional verbal information about the survey responses, which is used to evaluate the quality of the response or to help determine whether the question is generating the information that its author intends" (Beatty and Willis 2007, 287). This definition stresses that the method tests whether the information being generated by the question meets the authors measurement aims and therefore it is *critical* that there is a clear specification of these aims before any cognitive testing is undertaken (see Chapter 3).

Despite being fairly well established in national studies, cognitive interviewing had not really made its mark effectively on comparative survey work in the past (Smith 2004) probably due to cost and logistical constraints but also reflecting the lack of methodological work on how to adapt this method for cross-national implementation. In addition previous attempts to apply this method cross-nationally had often been blighted by a lack of coordination and harmonization, leading to concerns about the interpretation of data collected in the studies (Miller 2007). In particular it was often very difficult to trace reported findings from national researchers back to the data that had been collected (ibid). A further consideration was that there was little clarity about how cognitive interviewing might be used to influence the development of a source questionnaire, rather than simply being used to test multiple language versions.

In order to try and reduce the error associated with instrument design and respondents' interaction with it, I was personally committed to seeing this method introduced on the ESS. Around the same time I was introduced to Kristen Miller from the National Centre for Health Statistics[5], Washington DC, USA, who was looking for ways to improve the use of cognitive interviewing when used on multi-national and multi-lingual surveys. The idea for a collaboration to develop an effective way to use cognitive interviewing on cross-national was then established.

As part of that collaboration I applied to the ESRC small grants scheme for funds to support the British part of the project and the application was successful. The abstract for the project is shown below.

Recently there has been significant progress towards achieving greater equivalence in cross-national surveys. However, a key area that lags behind is the achievement of construct equivalence. Data collected from the European Social Survey (ESS) has shown mixed success in measuring constructs equivalently and it is therefore essential to tackle this. …(The projects)…key aim is to promote greater equivalence so questionnaires can more easily cross linguistic, cultural and national boundaries. The researchers involved will address this by examining the benefits of including cognitive interviewing in cross-national questionnaire design. A collaboration between the ESS and the Budapest Initiative (a UN / Eurostat / WHO taskforce), the findings from the project will be considered by the International Workshop on Comparative Survey Design and Implementation who are developing best practice guidelines in this area. Issues to address include coordination, consistency of interviewing technique and optimal analysis procedures.

The collaborative project attempted to identify a framework that better ensures equivalent (but not necessarily identical) constructs across nations. Aware that quite small changes in question wording could lead to large differences in answers to survey questionnaires (Groves et al. 2004), the project used cognitive interviewing at the design stage to see if and how critical wording differences between language versions could be minimised in advance.

[5] The mission of the National Center for Health Statistics (NCHS) is to provide statistical information that will guide actions and policies to improve the health of the American people. As the Nation's principal health statistics agency, NCHS aims to lead the way with accurate, relevant and timely data.

The key aim of the project was not simply to 'do' cognitive interviewing in a number of countries. Rather, building on an input harmonized cross-national project design, its primary task was to draft, test and refine a set of guidelines detailing best practice in cross-national cognitive interviewing.

An important task of the project was to use cognitive interviewing methods to distinguish between translation and source questionnaire design issues. As noted earlier most cross-national surveys like the ESS, work on an 'Ask the Same Question' sequential design model (Harkness 2007, 80), in which the various language versions of the questionnaire are derived from translations of a single language 'source questionnaire'. The success of this approach depends upon the suitability of the source questionnaire content and formulation as well as on the quality of the translations. Cross-national cognitive testing is therefore equally likely to be an invaluable tool for assessing and improving translated questions, as well as a tool to improve the source questionnaire. This joint project used an unusual new approach, effectively 'decentring' the questionnaire design (Harkness et al. 2010) by testing questions simultaneously in a range of languages but with the explicit intention of using that to improve the source questionnaire. Care also had to be taken to remember that not all eventual language versions were being tested and therefore the findings had to be considered in that context.

The project had a number of key components:

• Building on the CSDI guidelines, the researchers involved drafted, discussed and revised a set of best practice guidelines for conducting cross-national cognitive interviewing and analysis. The final decisions regarding the adaptation of these for this project were made at the first meeting of the group held at City University, London[6]. The guidelines were then tested as the project progressed. Attention was given not only to issues of coordination in a multinational study, but also to the sort of research design that is sensitive enough to identify country-specific and cross-national issues. The following issues, amongst others, were addressed in the protocols:

a) Sampling and recruitment strategies – specifically how to ensure cross-national equivalence, and issues of sample size and coverage.

[6] The meeting was chaired by myself and Kristen Miller.

b) Interviewer recruitment and training – ensuring harmonisation of training.

c) Developing interviewing protocols and considering the deployment of different techniques (such as probing, think aloud, paraphrasing and unscripted probes). Consideration was given as to whether or not behaviour coding by interviewers should be included, and whether or not issues that arise during interviewing in one country should then be tested in other countries. Regular, pre-arranged conference calls were used to facilitate these discussions.

d) Analysis and interpretation issues – specifically guidance on a range of analysis techniques, format of analysis materials and differentiating between sources of evidence.

The questions tested included 10 Budapest initiative health questions and 10 ESS questions from the two rotating modules for Round 4. The ESS modules covered attitudes to welfare provision and age discrimination respectively. Both modules were still at their design stage and therefore benefited substantially from cognitive interviewing in a cross-national context. For example, the ESS welfare module stems from a proposal based to some extent on a Scandinavian model of welfare, while the ageism questions included many abstract concepts previously designed for a UK survey.

Analysis was then conducted according to a pre-agreed structure, based on various approaches contained in the literature (Collins 2007). Each country used 'charting', which facilitated analyses within-country, across-country and within-subgroups. The starting point for content analysis was the Framework programme, a tool that allows a systematic approach.

Each participating country then summarised their interview data using a pre-agreed matrix, the columns being the test questions and the rows being the individual cases. The templates were compiled in English allowing cross-national comparisons. The matrix also included a summary of respondent demographic characteristics. For each question, a summary was made of respondents' understanding of it, the judgements they made in formulating an answer, any problems they had in answering, and of course the answers themselves. This approach allowed the data to be read both horizontally (as a complete case record for each individual), and vertically (by question across all cases).

The final stage was a team analysis conducted during a three-day intensive meeting[7], thus mitigating the problem of a centred approach where a single team based in one country do all the analysis (Miller et al. 2005).

In many ways the design and conduct of this project reflected similar issues to the development of the main ESS survey. Previously cognitive interviewing in cross-national studies had been blighted by a lack of a common approach being followed across countries, just as many cross-national studies had been in their main stage implementation. In cognitive interviewing projects conducted cross-nationally sample designs often differed, interviewing styles varied, analysis techniques were not uniform and there was a lack of transparency, making comparative analysis unreliable. Translation processes between countries frequently differed or were not as rigorous as those used in the main stage, making the testing itself rather questionable. It was therefore difficult to know if there was sufficient quality across countries. In particular differences discovered between countries at the analysis stage could either reflect real differences in how the questions were being interpreted or simply differences in how the cognitive interviewing itself was being implemented.

This joint project between the ESS and the Budapest initiative led to a new harmonized approach being developed and successfully implemented. By agreeing on an approach to follow based on known best practice, compromising where differences of opinion existed on how to proceed and working closely across international borders, we successfully minimized differences between countries. The approach used by the ESS was published as a working paper (Fitzgerald et al. 2009) and the joint work later appeared in Field Methods (Miller et al. 2011). The approach has also been referenced in the CSDI cross-national survey guidelines (http://ccsg.isr.umich.edu/pdf/11PretestingFeb2012.pdf), implemented in all subsequent rounds of the ESS and further developed by Kristen Miller in her work.

There were deficiencies in the process worth noting linked to differences in levels of experience with the technique cross-nationally, and a tendency for the more experienced to be reluctant to change their way of working to promote equivalence. However, cross-national survey researchers now have a method that can better tackle instrument and related respondent error

[7] Held in Washington DC, USA, in January 2008. I chaired the ESS section and the Budapest initiative section was chaired by Kristen Miller.

using cognitive interviewing to help develop cross-national questionnaires. It was also of note that the analysis phase of the project for the ESS questions was greatly enhanced by having a clear specification of the individual item aims, facilitated by their specification in the questionnaire design template (see Chapter 3). The next chapter discusses an additional component which arose from this project: a Cross National Error Source Typology.

Chapter 5: New Insights in cross-national pre-testing

In this Chapter I discuss how my work has led to a pre-testing approach being developed and implemented on the ESS which is an exemplar within the field. I will also outline how the cross-national cognitive interviewing project discussed in Chapter 4 provided the opportunity for me to reflect on pretesting analysis in a multi country study and develop and test a typology for analysis. I discuss attempts to apply that typology to data produced by pre-testing methods other than cognitive interviewing. I conclude the chapter by discussing the pre-testing cycle that I have implemented on the ESS and highlight the particular benefits of triangulation for cross-national surveys.

Cross-National Error Source Typology

The most widely known theory regarding the survey response process is probably that of Tourangeou (1984) who breaks the response process down into four stages. Following an information request to the respondent via a closed question the steps are: comprehension, recall, judgement and finally response. Campanelli describes cognitive interviewing as "A type of in-depth or intensive interview that pays explicit attention to the mental processes respondents use to answer survey questions and uses specialized techniques, such as thinking aloud" (2012, 198). Cognitive interviewing is therefore clearly well-placed to examine each of these phases. In a cross-national survey the additional challenge is that there are multiple (target language) instruments to test, each seeking equivalent measurement but through a different language. Furthermore, in cross-national survey programmes pre-testing is predominantly focused on developing the source questionnaire, although it can of course be used to test instruments in the final target language versions too (Willis 2015). In addition the issue of differing contexts has to be considered. Smith (2011) notes that context can be a source of error and in particular a source of comparison error.

In a recent example from ESS Round 7, respondents were asked to report on the amount of alcohol consumed in the last 7 days. However because the type and typical amount of alcohol consumed in a serving differs so markedly across Europe, the approach to measuring this in a self-reported survey has to take account of those contextual differences. Or take the inclusion of the first non Judeo-Christian country in the ESS which led to difficulties using the existing ESS religion measures in Turkey (Fitzgerald and Jowell 2010).

Therefore, when it came to the analysis phase of the cross-national cognitive interviewing project (see chapter 4), I was very concerned to ensure that the team were properly prepared to differentiate between sources of error that would not apply in single-nation, mono-lingual surveys and the usual range of design flaws that cause problems in the response process. Therefore a typology was developed on the basis of "...experience of cross-national questionnaire design, translation assessment, feedback from data users and quantitative assessments of question quality" (Fitzgerald 2007, 273). Whilst it transpired later that there were similar typologies that had been developed as part of cognitive pre-testing and behaviour coding projects (see Fitzgerald et al. 2011 for a discussion of these), the CNEST made important additional distinctions, especially in relation to errors arising from the translation process. In addition the cultural category was more applicable to cross-national studies, where the contextual challenges are often greater than with multi-cultural surveys conducted within a single country. With multi-country surveys this could include instances where it is impossible to ask about a concept that does not exist at all in a particular country or group of countries (ibid).

Table 5.1 outlines how CNEST comprises four categories. What is particularly beneficial about CNEST is that its structure points those applying it in the direction of potential solutions to the problem. The first category is *poor source question design* and incorporates all of the usual types of error resulting from poorly designed and particularly challenging questions. Where these difficulties are found the solution is to design a better question or perhaps to drop the question entirely. The second and third categories relate to error arising from the translation process. These difficulties would not be found in a single language survey. The solution to a *translator error* is to improve quality control procedures in future and to brief translators in the main stage about the nature of the problem. However, human error is always likely to persist to some extent. This category is particularly useful as it reassures the researcher that the source questionnaire itself does not need to be changed. The solution to the other translation problem - resulting from source question design – is to acknowledge that in a single country, single language study the question would probably have worked effectively. However the source question is difficult to translate and will either need to be amended to make it easier to render in other languages or translators will need to be given more guidance. My supervisor, Dr Tom Smith, raised the question as to whether translation error can ever be completely eradicated, which is clearly an area that should be investigated further. In the new Horizon 2020 SERISSS project that I am coordinating (www.seriss.eu), we will be investigating what impact greater

adaptation in translation has on data quality. The last category - *cultural portability* - refers to cases where either the concept being measured does *not* exist in all countries, or perhaps it exists but cannot be measured with the current approach. For example, it may be that an ASQ approach is not possible. In this case a new question will be required or an alternative measurement will need to be developed.

Applying CNEST to data from other pre-testing sources and the benefits of triangulation

Chapter 4 outlined the various pre-testing methods, each of which can help to test specific parts of the response process. It was therefore of interest to see whether CNEST could be applied to data from other forms of pre-testing other than just cognitive interviewing. More specifically an attempt was made to see whether CNEST could be applied to findings arising from a combination of different pre-testing sources (Fitzgerald, Winstone, and Prestage 2014a). The testing took place as part of development of the questionnaire for Round 6 of the ESS, and covered the two rotating modules 'Measuring personal and social well-being' and 'Europeans' understanding and evaluations of democracy' (see www.europeansocialsurvey.org). Towards the end of the development process the findings from expert review by National Coordinators, a pilot in 2 countries with interviewer feedback and advance translation were considered together. The FORS working paper (ibid) outlines how the typology was once again comprehensive in mapping the response problems identified through the various sources.

Table 5.1 The Cross National Error Source Typology (CNEST)

		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

Reproduced from Fitzgerald et al (2011, 570)

Upon later reflection it became apparent to me that there are some important issues about the different pre-testing methods available to the cross-national researcher. Whilst cognitive interviewing is an extremely effective tool in helping to identify and explain response problems, the likely prevalence or scale of the problem is of course not identified. It is still a serious scientific error in my view to 'count' responses and findings in cognitive interviewing studies, a trend which persists despite having no clear scientific basis and which reflects poorly

on our field. This is particularly the case where sample sizes are small. The reason for concerns about this are firstly related to the sample design, which is almost always purposive and not random. Secondly they relate to the sample size which is normally so small that differences between groups would almost never be statistically significantly different. Third they relate to the tendency to extrapolate statistically from a qualitative interview. By their very nature the process of a cognitive interview is not standardised like a field interview and therefore the basis for creating statistics from the data is not present. The norm in the USA, however, is often not to conduct more than 10 cognitive interviews and yet counting remains a common practice and sometimes one even sees findings presented as percentages. However, there is evidence that sample sizes in cross-national studies are increasing to help reach the 'saturation point' when new findings no longer emerge (Willis 2015). This reminds us that the reliability and validity of cognitive interviewing is still in need of detailed investigation and perhaps suggests it is a technique best used in combination with other methods.

Quantitative findings from the ESS pilot clearly give an indication of the prevalence or size of problems, however, without findings from cognitive interviewers, feedback from pilot interviewers or findings arising during advance translations, the reason for the error is often difficult, if not impossible, to determine. Take the example of a scale where the pilot identifies that a latent variable which is measured effectively in one country based on 3 items, is not replicated in the second country. The reasons for this lack of equivalence can statistically be linked to the specific items that measure the scale but beyond that there is no clear evidence for why the scale has not worked in both countries. Additional qualitative feedback of some form or some additional quantitative data is therefore required to complement the statistics.

The benefits of triangulation, a process that generates findings about questionnaire problems based upon combining evidence from multiple pre-testing sources, has been demonstrated in the work discussed here (Fitzgerald, Winstone, and Prestage 2014a). This is a theme I hope to explore more in future research. An example, where different methods suggested alternative problems or differing reasons for the same problem, suggest not only that triangulation is critical when developing cross-national survey instruments, but also serves as an important reminder that pre-testing itself is subject to error.

A new cycle for cross-national questionnaire design and pre-testing

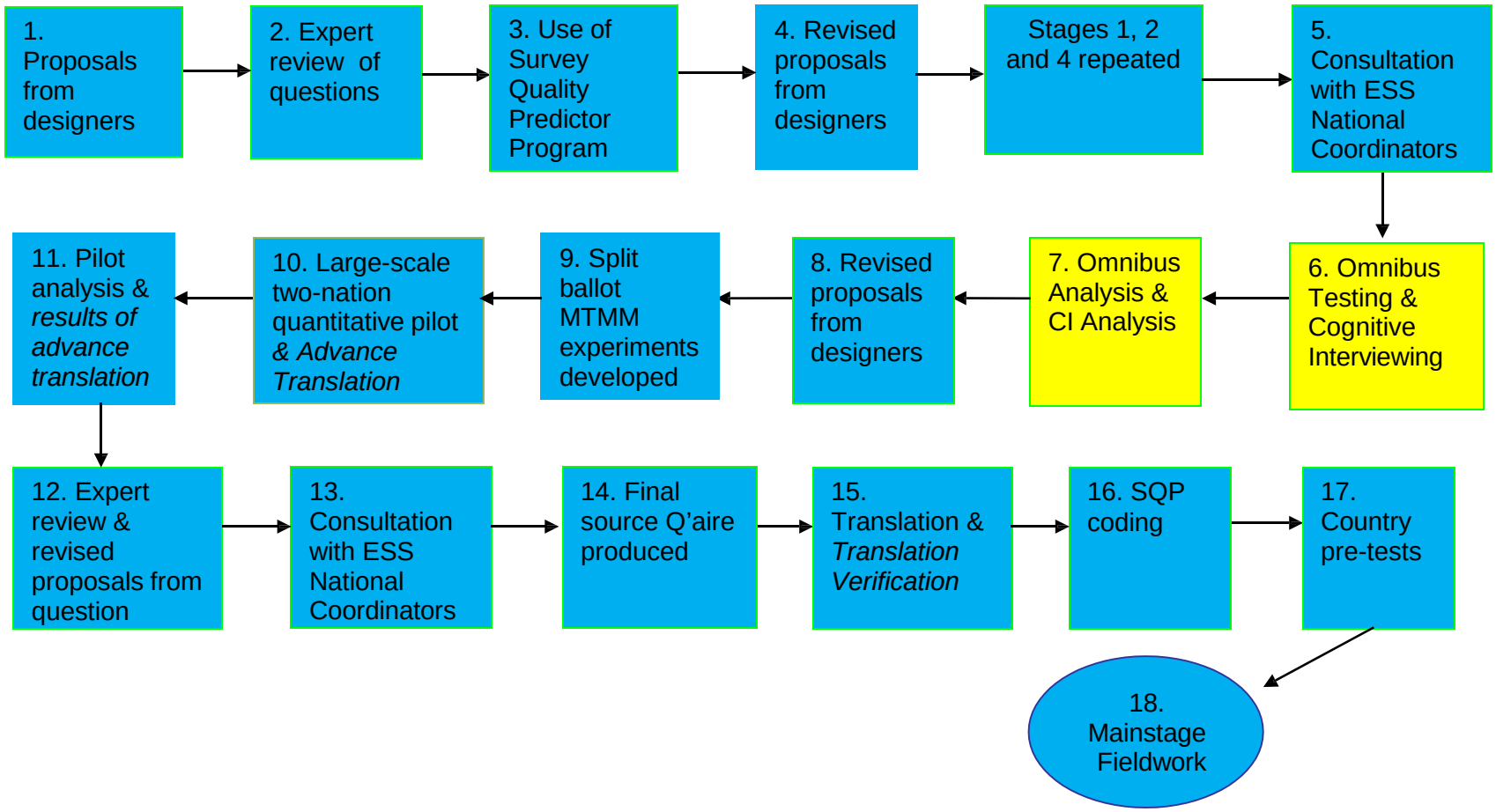
In early rounds of the ESS (1-3) pre-testing was limited to expert papers by substantive specialists, expert review (including with National Coordinators), use of the Survey Quality Predictor program, a large-scale 2-nation pilot and Multi Trait Multi Method experiments (Fitzgerald and Jowell 2010). However, early assessments of data collected on the ESS suggested mixed success in developing cross-national questionnaires that measure attitudinal constructs equivalently (Saris and Gallhofer 2007, 71). The early design meant that there was only a single quantitative test in two countries, which was quite late in the process. In turn, this meant that there was no opportunity to test the impact of any changes made in the pilot. Furthermore, apart from comments from NCs in all countries, detailed pre-testing work was limited to the UK and one other country. This often led to major changes being requested *after* the source questionnaire had been finalised, once national teams started work on their instruments. These requests sometimes results in chaotic last minute changes of having to ignore less urgent issues to avoid causing uneven implementation of the changes.

I therefore redesigned the questionnaire construction and pre-testing schedule for later waves of the ESS. Figure 5.1 shows the overall timetable for Round 8, showing the enhanced programme used in more recent rounds. A key addition has been the inclusion of earlier quantitative testing, using commercial omnibus surveys in three countries (stages 6 and 7). This allows more obvious examples of quantitative error to be identified, such as high item non-response, skewed data and a lack of scalability, as well as facilitating an initial look at the relationships between variables and concepts. It also results in an early test of the translation, which is conducted according to the ESS committee approach. Not all items are tested but the design team identifies those most in need of initial quantification. Around the same time cross-national cognitive interviewing is conducted in 3-4 countries with the tests focused on around 5 items from each of the rotating modules. This allows those items where there are concerns about the response process to be tested (also stages 6 and 7). Finally, in addition to the 2-nation pilot, I supported the translation team to introduce advance translation (stage 10 in italics), something the late Janet Harkness had repeatedly argued for in surveys like the ESS (Harkness 2007). This intensive process asks translation teams to document, in detail, the difficulties they have in rendering the source questionnaire in their own language. Finally to help quality control, the final translation phase, translation verification (Stage 15 – in italics), was also added (Dorer, Widdop, and Fitzgerald 2013). This involves national teams sending their

translated questionnaire to an external agency who compare the source and target versions and raise queries with the national teams where they feel the rendering has not been effective.

At two key points in the design process the relevant information from these different pre-testing methods is then triangulated (Fitzgerald, Winstone, and Prestage 2014a) to strengthen the analysis and also to help question the individual pre-testing method conclusions themselves. The ESS is well financed and is therefore able to allocate funds to these tasks. However having developed a method for applying cognitive interviewing cross-nationally and a typology for classifying pre-testing findings and having demonstrated the benefits of triangulation, it was easier to make the case to colleagues for investing resources in tackling this element of TSE.

Figure 5.1 Source Questionnaire development and pre-testing Round 8



The CNEST typology is a useful tool for the survey researcher when developing and pre-testing cross-national questionnaire instruments. It aids them when trying to ascertain the source of error in order to work out if it is a standard component of 'instrument' or 'respondent' TSE, or instead something specific to comparison error and cross-national implementation. Knowing the source of error in turn helps in finding a potential solution to the problem.

Upon reflection, there is probably a category missing from the typology which refers to failures in the pre-testing method itself, meaning that the source of the error cannot be determined. There were instances where the cognitive interviewing data collected did not allow a conclusion to be confirmed across all countries (Fitzgerald et al. 2009). This category should probably be added since pre-testing itself can of course be error-prone. A discussion of 'total pre-testing error' is beyond scope of this dissertation, however it reminds us that pre-testing feedback itself can be problematic. Having more than one source of evidence to support a conclusion is therefore particularly helpful especially in a cross-national context.

My contribution to improving pre-testing for cross-national surveys is clear, developing significantly better methods and interpretative tools than those available previously. In future I plan to do work which assesses the impact of different pre-testing tools in terms of improving data quality and to what extent they help instruments to meet the measurement aim specification. There are few studies that look at the reliability and validity of particular methods and fewer still which assess their utility versus specified measurement aims, particularly in cross-national perspective. To a large extent that remains unchartered territory.

Chapter 6: Error in the results derived from a social survey

My work on the ESS has naturally been focused on the design, implementation and further development of the survey and its associated infrastructure programme. However, I have had a few opportunities to analyse the surveys data for substantive purposes, stimulated by invitations to write for the British Social Attitudes series and a desire to contribute a paper to the ESS international conference in Cyprus (2012). Such opportunities are important, not only for personal satisfaction and social inquiry, but also because methodologists can arguably contribute more effectively to their field when they themselves have worked with the data collected for substantive purposes. Such exposure makes one better able to reflect on possible TSE impacts on the analysis and to decide where the priorities should lie when directing a social survey programme.

This chapter's focus is not on the substantive conclusions *perse*, but rather on the potential for error in the results derived from the survey. Most TSE scholars rather oddly do not include the analysis step in their models. Instead they stop with the data processing stage including coding, data handling and perhaps weighting (for example see Groves and Lyberg 2010; Biemer 2010). However, this omission is rather unfortunate since TSE is about understanding how the statistics generated from social survey data may not accurately reflect the intended measures. It is therefore this is left out, since the potential for error feeding into, or being generated during, the analysis stage, is rather large, especially with cross-national analysis which is significantly more complex than with a single nation study.

Smith however explicitly includes analysis error as an observational component of TSE. In particular he identifies conceptual, statistical and presentational error (Smith 2011), each of which pose a serious threat to the reliability and validity of the data. Even the most professionally and rigorously designed and implemented survey could end up being used to produce inaccurate findings, if the analysis that is performed is inaccurate in some way. In his seminal paper 'How comparative is comparative research' Jowell outlines '10 golden rules for comparative research', many of which apply to analysis, in addition to those focused on the design and implementation of the survey. In particular he argued *against* including too many countries in analysis to prevent over complication and data overload; *for* ensuring aggregate level variables are used to complement the individual level survey data; *for* being as open to the limitations of the data as much as its explanatory power; and *for* checking carefully whether

differences discovered are methodological artefacts, before celebrating the discovery of a new social trend (Jowell 1998). This approach suggests ways to reduce the possibility that 'comparison analysis error' will damage the findings from the study.

In this chapter I present and critically evaluate two of my own contributions from the perspective of the error that may still be present in the findings. I discuss the analysis error that may be contributing to deficiencies in my findings. In addition I identify errors from parts of the survey design and implementation process that might also have an impact on this analysis.

Acculturation and attitudes towards homosexuality

Attitudes towards homosexuality have been shifting rapidly towards greater tolerance in much of the western world, with Europe arguably leading the way. However, there is a rather clear divide even in Europe on this issue, with Northern and Western Europe showing increasing tolerance towards homosexuals, whilst Eastern and parts of Southern Europe remain hostile or have even become more hostile in recent years (Fitzgerald, Winstone, and Prestage 2014). There has also been significant migration within Europe, with migrants in Eastern Europe choosing to move to Western Europe. This provides a natural experiment, allowing one to compare those who migrated to those who remain in the country of origin and to those in the destination country. In other words one can try to isolate the impact of moving from a less tolerant to a more tolerant environment in terms of attitudes towards homosexuals. We were then able to pose the question in our paper as to whether attitudes became more permissive, less permissive or there was no change as a result of moving westward. Data from the ESS was used to answer these questions (ibid).

One of the distinctive features of the ESS is that it includes all residents aged 15+ in each participating country, and not only those holding citizenship. By combining multiple waves of the study we were therefore able to produce a dataset that had sufficient numbers of migrants who had moved from Eastern to Western Europe, and to have enough cases where the respondent had moved at different historical intervals. This allowed us to isolate the impact of the elapsed time since migration on such attitudes and provided the basis for assessing possible acculturation. It is worth pausing here however to consider a possible source of error, and in particular comparison error. The ESS allows each country to use the best approach to probability sampling available in each country and then increases the sample size where possible to compensate for deviations from a simple random sample (Häder and Lynn 2007).

One key difference is that some countries use population registers as their sampling frame whilst others use household sampling frames or even area based samples. One clear difference here might be the inclusion or otherwise of migrants and immigrants in the samples drawn. Recent migrants might be excluded from population registers for a sustained period prior to registration whilst they might more easily be sampled by interviewers where there is an area or household based sample. On the other hand there have been reports from ESS National Coordinators that interviewers are confused about whether those who are working outside of the country should be included in the sample frame being as they are unsure about how to apply rules regarding eligibility. It is impossible to be clear about the impact these differences might have on the estimates in our paper. This paper did not make comparisons between individual countries, however when combining the data across countries and weighting the data from each country by its population size, immigrants moving to certain countries may have been over- and under-represented. If the impact of particular destination countries on acculturation differs, this may have had an impact on the conclusions drawn or at least on the size of the effects. When using the ESS dataset to compare the whole population between countries this would be unlikely to have an impact on the findings, but using the data to look at this small group of respondents (migrants) leaves open a much bigger possibility of error. So frame and selection error may have occurred as sources of TSE.

Another distinctive feature of the ESS is the very rigorous translation procedures, which set new standards in the field (Harkness 2007). By rejecting back translation in favour of a committee approach, standards were raised (Harkness, Villar, and Edwards 2010). However, translation is an intensive process and it was therefore agreed to require language versions *only* for languages spoken as a first language by 5% or more of the population. At the same time, to prevent deviations from delivering a standardised interview to all respondents, it was agreed to forbid live translations for any languages spoken by less than 5%. Again, as noted earlier, the impact on estimates for the general population in each country is likely to be small (see Chapter 2), however when examining the views of migrants the issue is clearly more salient. Unfortunately we cannot estimate the scale of the problem because when people are unable to take part in the study because they cannot speak the required language, no further information is collected about them. It is very likely that many of them were migrants to the country and this is clearly a potential source of error. In fact we were concerned enough about this to mention it explicitly in the paper. We concluded that: “The results of this study, however, cannot be generalized to immigrants who are unable to speak the language(s) spoken by >5%

of the population in the destination Western European countries" (Fitzgerald, Winstone, and Prestage 2014b, 16).

One of the particular challenges of cross-national analysis is assessing whether concepts have been measured equivalently across countries. Braun and Johnson (2010) outline a number of approaches to checking this, including: checking distributions and means, examining correlations, exploratory factor analysis, multi-group confirmatory factor analysis using structural equation modelling, latent class analysis and many more. In this study the dependent variable was a single agree / disagree item:

"Gay men and lesbians should be free to live their own lives as they wish"

As this is an ordinal single item, most advanced statistical procedures for checking conceptual equivalence are difficult or impossible to apply with statistical certainty (Braun and Johnson 2010). In addition, some of the independent variables that we included in our analysis (education and religious attendance) are difficult to assess for equivalence of measurement, as they are behavioural and nominal measures respectively. It was not therefore feasible for this paper to ascertain equivalence for the dependent and most of the independent measures. It was therefore fortuitous that we did not seek to compare means but were rather interested in the relationships *between* variables. However, we have to accept that there may have been some analysis error at the conceptual and statistical level in our reporting.

A final but important limitation in terms of analysis error is that we did not include post-stratification weights in our analysis. At the time that the paper was produced design and population weights were not available and the ESS, with post stratification weights only available from 2014, after the paper was published. This means that unequal response bias between countries was not being corrected, although there may be some debate about whether post-stratification weights were appropriate for this analysis as it involved a small subgroup – immigrants (ESS 2014). Considering the variation in response rates and possible non response bias amongst the origin and destination countries, this may be a cause for concern (see Chapter 2). However, the possible scale of any bias is not known.

Finally, it is noteworthy that we combined rounds of data in the paper which included a slightly different set of countries in each round with more Central and Eastern European countries

included in later rounds. This means that whilst some destination and origin countries were included from every round this was not the case for all countries, potentially over- or under-emphasising any acculturation effect where this differs. In addition there were some changes in fieldwork agency between rounds in particular countries, possibly introducing house effects. One final consideration is that there were changes in the performance of countries even when the same fieldwork agency was used, which again were not controlled for.

Despite these possible limitations the work arguably still makes a valuable contribution to the field (Fitzgerald, Winstone, and Prestage 2014b). For instance, it is rare in comparative surveys to have samples of immigrants derived from random sampling methods. In addition there were reassuring parts to the analysis such as finding the expected relationships between all of the independent and dependent variables. In reality few published social science papers are really scrutinised from a TSE perspective and whilst the analysis failings are sometimes listed, many of the data collection problems are not mentioned.

Traditional norms in European societies

The second substantive contribution discussed in this chapter involved a cross-national examination of the differences in disapproval for the 'transgression' of traditional norms, related to marriage and children. The chapter 'A chorus of disapproval? European attitudes to non-traditional family patterns' appeared in the 2010 British Social Attitudes series (Harrison and Fitzgerald 2010). The ESS Round 5 'Timing of life' module asked people how much they would approve or disapprove if someone transgressed social norms. Historically these norms stemmed from the European Christian tradition. In a split ballot questionnaire design half the sample were asked about a man, the other half a woman, enabling comparisons dependent on the gender of the transgressor. Specifically respondents were asked the following:

"How much do you approve or disapprove if a man/woman …

… chooses never to have children?
… lives with a partner without being married to her/him?
… has a child with a partner he/she lives with but is not married to?
… has a full-time job while he/she has children aged under 3?
… gets divorced while he/she has children aged under 12?"

In addition to mapping differences between countries in terms disapproval of those transgressing these norms, we also wanted to see whether such views were driven by levels of modernity and traditionalism in each country. This made a cross-national analysis particularly appealing. For the modelling analysis we therefore created an index that included (GDP) per capita by purchasing power, the proportion of females in employment, the divorce rate (number of divorces per 100 marriages) and the country's religiosity (the proportion of individuals claiming to belong to some religious denomination). In order to avoid a simplistic model we tested for the presence of a latent variable and then generated a factor score. This is an example of avoiding analysis error by developing a factor score which effectively weights the individual components. However, we also generated an additive index score for the number of items respondents said they disapproved of. "This index combined attitudes to each of the five norms, with respondents awarded one point for each item of which they disapproved" (Harrison and Fitzgerald 2010, 151). By combining 'strongly disapprove' and 'disapprove,' the relative weighting respondents may have given to transgressions on a particular item were ignored and this may of course have differed across countries. This latter approach to analysing scales is frequently used as an alternative to generating factor scores for analytical ease, but also because it is easier to report to the reader.

In line with the recommendations from Jowell (1998), my co-author and I decided *not* to include all possible countries in our analysis. This certainly made interpretation easier and also made things more straightforward in terms of presentation for the reader. However, it may have been that our 'exemplar' countries were not representative of the regions / welfare state regime types on this particular topic. However, our analysis and text implied this was indeed the case and this *may* have been a form of analysis error.

During our work on acculturation (Fitzgerald, Winstone, and Prestage 2014b) we did not have access to post stratification weights. Therefore it is possible that our analysis hid differences in response bias between countries and regional groupings. In addition we did not check for functional equivalence for the items in our analysis.

Once again it is worth stressing that despite these possible sources of analysis error the chapter still makes an interesting contribution to the field. In terms of face validity, there is no doubt one would have expected distributions suggesting more disapproval of the transgression of these norms in Eastern Europe, with Scandinavia being far more tolerant. This was indeed the

case. Additionally the traditionalism-modernity index we generated suggested a clear and robust relationship between higher levels of traditionalism and greater disapproval of norm transgression. However, an assessment that considers possible TSE impacts on the findings suggests that whilst the overall patterns are robust, it might be wise to treat the exact size of differences with some caution.

The analysis phase of the survey life cycle can be fraught with difficulty, complexity and error, particularly when dealing with cross-national data. It is unlikely in the near future that there will be complete, operational statistical methods for measuring sources of TSE during the data collection phase, or guaranteed methods for avoiding it during the analysis phase. The aim of this chapter of the exegesis was to be self-critical of my own substantive work, in order to highlight the possible impact of TSE before and during the analysis phase. This process is useful for signalling where caution may be required and highlighting areas for improvement in future.

Chapter 7: Conclusion

The last century and the early part of the 21st century have seen substantial and positive developments in the scope and quality of cross-national social surveys. At the same time, the development of the TSE framework has highlighted how complex and challenging it is to design and implement such studies to a standard that will produce reliable, valid and comparable data. It is therefore important that tools are developed to reduce error in cross-national studies in turn making them better placed to deal with the challenges societies face due to globalisation (Heath, Fisher, and Smith 2005). This exegesis has attempted to outline how my own work has contributed to that task, whilst demonstrating an awareness of the remaining impact of TSE on analysis of comparative survey data.

There are particular ways in which my own work has sought to tackle TSE and comparison error. By identifying protocol error as a key driver of comparison error in the past, my work with others on the ESS has led to a full specification of requirements aiming to maximise compliance with the harmonised requirements of the survey (Fitzgerald and Jowell 2010). Through the collection and analysis of harmonised paradata in regard to fieldwork on the ESS, non response error has been examined in greater depth than before, including in publications I co-authored with colleagues (Stoop et al. 2010). Through the introduction of the questionnaire design template which I developed (Fitzgerald 2007) there is now clear, documented specification in the questionnaire design stage of ESS, minimising 'specification error'.

The development of a new methodology for applying cognitive interviewing cross-nationally which I developed with Kristen Miller (Miller et al. 2011), and a typology for interpreting the additional sources of error in a cross-national study which I developed (Fitzgerald et al. 2009) both set new standards of rigour in the field. Together they add two new components to the more intensive development and pre-testing cycle which I introduced on the ESS. In turn this has facilitated more detailed triangulation of pre-testing results to help reduce TSE and its comparative complications in regard to instrument design.

The final theme addressed in this exegisis relates to the analysis phase of the survey lifecycle. It highlights both the additional error analysis itself can introduce and how TSE during the design and implantation phases can 'feed forward' and negatively impact the conclusions drawn. This is an important theme and should remind all methodologists that error is truly

problematic in terms of its potential impact on this final part of the process. This also suggests that the PI of any cross-national survey should perhaps approach the task of minimising TSE in an environment of limited resources by assuming an analysts mind set. In addition they might remind analysts of the potential for methodological error to compromise their work.

Despite the general focus on error in this thesis, the final substantive chapter highlights the interesting and informative contribution that cross-national survey research can make in helping us understand our own societies, and those of the rest of the world. The many thousands of publications that use the ESS data demonstrate the relevance of the ESS and cross-national research in helping us understand our societies and providing a basis for helping us address key challenges in the future.

My planned future research agenda will build substantially on many of the themes addressed in this thesis. In terms of fieldwork quality and response rates, I plan to look further at the impact of different design decisions on output quality. In addition I also hope to introduce better communication and project management tools to help ensure more efficient communication and management of the decentralised fieldwork process on the ESS. The SERISS (www.seriss,eu) project will provide a perfect platform to support that work.

In terms of specification error the SERISS project should see the completion of the database for questionnaire specification and design based on the ESS design template I developed. It is hoped that this will also enable other survey programmes to implement specification procedures in their life cycles.

In the area of pre-testing my future research agenda will be focused upon more detailed assessment of pre-testing methods, and evaluating the planned switch of early pre-testing from face-to-face commercial omnibus surveys onto the ESS web panel being developed under the SERISS project.

Most immediately my priority will be to publish a series of papers derived from the ESS mixed mode methodology programme, to highlight the serious problems such a method holds and recommending that this approach be avoided for cross-sectional cross-national surveys in Europe for the foreseeable future. This work relates to another key area of TSE related to instrument and mode.

Finally, in terms of substantive research I am planning to use ESS data to compare how the views of those who migrated some time ago (or whose parents did) compare to the native population in assessing views towards immigration. In addition, a recent grant from the Newtown Fund, to facilitate the inclusion of some ESS modules on democracy and health inequalities, will provide an opportunity to make rare comparisons between Europe and South Africa on this topic.

My future research will be informed by the attempts to tackle and better understand TSE and comparison error in my published work. I would hope that my future work will ensure a greater awareness about TSE amongst colleagues whether methodologists or substantive analysts, whilst also directing attention to the areas where we are best placed to minimise that error.

References

Beatty, Paul C., and Gordon B. Willis. 2007. "Research Synthesis: The Practice of Cognitive Interviewing." *Public Opinion Quarterly* 71 (2): 287–311.

Biemer, Paul P. 2010. "Total Survey Error: Design, Implementation, and Evaluation." *Public Opinion Quarterly* 74 (5): 817–48.

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. Holberk New Jersey: John Wiley & Sons.

Billiet, Jaak, Michel Philippens, Rory Fitzgerald, and Ineke Stoop. 2007. "Estimation of Response Bias in the European Social Survey: Using Information from Reluctant Respondents in Round One." *J. Off. Stat* 23 (2): 135–62.

Boersch-Supan, Axel, Hendrik Jürges, Karsten Hank, and Mathis Schröder. 2010. "Longitudinal Data Collection in Continental Europe: Experiences from the Survey of Health, Ageing and Retirement in Europe (SHARE)." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 485–96. New Jersey: Wiley Hoboken, NJ.

Braun, Michael, and Timothy P. Johnson. 2010. "An Illustrative Review of Techniques for Detecting Inequivalences." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 375–94. New Jersey: Wiley Hoboken, NJ..

Bulmer, Martin, 1998. "The Problem of Exporting Social Survey Research". *American Behavioral Scientist* 42 (2): 153–67.

Capanelli, Pam. 2012. "Testing Survey Questions." In *International Handbook of Survey Methodology*, edited by Edith D. de Leeuw, Joop Hox, and Don Dillman. New York: Routledge.

Collins, Debbie. 2007. "Analysing and Interpreting Cognitive Interview Data: A Qualititative Approach." Paper presented at Questionnaire Evaluation Standards Conference. http://wwwn.cdc.gov/qbank/Quest/2007/QUEST%202007%20Proceedings-all%20papers.pdf. Ottawa, Canada.

Converse, Jean M., and Stanley Presser. 1986. *Survey Questions: Handcrafting the Standardized Questionnaire*. Beverly Hills: SAGE Publications, Inc.

Converse, Jean M. Survey Research in the Unites States. Roots and Emergence. 1980-1960. 2009. New Brunswick, NJ: Transaction Publishers.

Couper, Mick P., and Edith D. de Leeuw. 2003. "Nonresponse in Cross-Cultural and Cross-National Surveys." In *Cross-Cultural Survey Methods*, edited by Janet A. Harkness, Fons JR Van de Vijver, and Peter Ph Mohler. Hoboken, NJ: Wiley.

Dorer, Brita, Sally Widdop, and Rory Fitzgerald. 2013. "Translation Verification in the ESS: A Means for Achieving Equivalent Translations in a Cross-National Survey?" Paper presented at 5th ESRA conference. 15-19 July 2013. Ljubljana, Slovenia.

ESS. 2014. "Weighting European Social Survey Data." http://www.europeansocialsurvey.org/docs/methodology/ESS_weighting_data_1.pdf.

ESS. 2015. "Round 8 Survey Specification for ESS ERIC Member, Observer and Guest Countries." European Social Survey ERIC. http://www.europeansocialsurvey.org/docs/round8/ESS8_project_specification.pdf.

European Science Foundation, (ESF) Standing Committee for the Social Sciences. 1999. "The European Social Survey (ESS) –a Research Instrument for the Social Sciences in Europe." Strasbourg. http://www.europeansocialsurvey.org/docs/about/ESS_blueprint.pdf.

Fitzgerald, Rory. 2007. "Improving Documentation of Questionnaire Development." Paper presented at CSDI workshop. March 29th to 31st. Chicago. USA.

Fitzgerald, Rory. 2015a. “Sailing in Uncharetered Waters: Structuring and Documenting Cross-National Questionnaire Design.” GESIS Working paper. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_papers/GESIS-Papers_2015-05.pdf.

Fitzgerald, Rory. 2015b. “The Challenges of Cross-National Surveys: The Example of the European Social Survey (ESS).” Paper presented at Caucasus Research Resource Centers (CRRC) conference. June 26th and 27th 2015. Tblisi, Georgia.

Fitzgerald, Rory, and Roger Jowell. 2010. “Measurement Equivalence in Comparative Surveys: The European Social Survey (ESS) - From Design to Implementation and Beyond.” In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 485–96. New Jersey: Wiley Hoboken, NJ.

Fitzgerald, Rory, Sally Widdop, Debbie Collins, and M. Gray. 2009. “Testing for Equivalence Using Cross-National Cognitive Interviewing.” *Centre for Comparative Social Surveys Working Paper*, no. 01.

Fitzgerald, Rory, Sally Widdop, Michelle Gray, and Debbie Collins. 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): 569–99.

Fitzgerald, Rory, Lizzy Winstone, and Yvette Prestage. 2014a. “A Versatile Tool? Applying the Cross-National Error Source Typology (CNEST) to Triangulated Pre-Test Data.” *FORS Working Paper Series*, no. 2. http://forscenter.ch/wp-content/uploads/2014/08/FORS_WPS_2014-02_Fitzgerald.pdf.

Fitzgerald, Rory, Lizzy Winstone and Yvette Prestage. 2014b. “Searching For Evidence of Acculturation: Attitudes Toward Homosexuality Among Migrants Moving From Eastern to Western Europe.” *International Journal of Public Opinion Research*, July, edu021. doi:10.1093/ijpor/edu021.

Gideon, Lior. 2012. *Handbook of Survey Methodology for the Social Sciences*. New York: Springer.

Groves, Robert M., and Mick P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.

Groves, Robert M., Floyd Fowler, Mick P. Couper, James Lepowski, Eleanor Singer, and Roger Tourangeau. 2004. *Survey Methodology*. Hoboken, New Jersey: John Wiley & Sons.

Groves, Robert M., and Lars Lyberg. 2010. “Total Survey Error: Past, Present, and Future.” *Public Opinion Quarterly* 74 (5): 849–79.

Häder, Sabine, and Janet A. Lynn. 2007. “How Representative Can a Multi-Nation Survey Be?” In *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 33–52. London: Sage.

Harkness, Janet A. 2007. “Improving the Comparability of Translations.” In *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 79–92. London: Sage.

Harkness, Janet A., Brad Edwards, Sue Ellen Hansen, Debra Miller, and Ana Villar. 2010. “Designing Questionnaires for Multipopulation Research.” In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 17–29. New Jersey: Wiley Hoboken, NJ.

Harkness, Janet A., Ana Villar, and Brad Edwards. 2010. “Translation, Adaptation and Design.” In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 117–40. New Jersey: Wiley Hoboken, NJ.

Harrison, Eric, and Rory Fitzgerald. 2010. “A Chorus of Disapproval? European Attitudes to Nontraditional Family Patterns.” In *British Social Attitudes 26th Report*, edited by Alison

Park, John Curtice, Katarina Thomson, Miranda Phillips, Elizabeth Cleary, and Sarah Butt, 135–58. London: Sage.
Heath, Anthony, Stephen Fisher, and Shawna Smith. 2005. “The Globalization of Public Opinion Research.” *Annu. Rev. Polit. Sci.* 8: 297–333.
Jackson, Jonathan, Ben Bradford, Mike Hough, Jouni Kuha, Sally Stares, Sally Widdop, Rory Fitzgerald, Maria Yordanova, and Todor Galev. 2011. “Developing European Indicators of Trust in Justice.” *European Journal of Criminology* 8 (4): 267–85.
Jowell, Roger. 1998. “How Comparative Is Comparative Research?” *American Behavioral Scientist* 42 (2): 168–77.
Jowell, Roger, Max Kaase, Rory Fitzgerald, and Gillian Eva. 2007. “The European Social Survey as a Measurement Model.” In *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 1–31. London: Sage.
Kish, Leslie. 1994. “Multipopulation Survey Designs: Five Types with Seven Shared Aspects.” *International Statistical Review / Revue Internationale de Statistique* 62 (2): 167–86.
Miller, K., R. Fitzgerald, J.-L. Padilla, S. Willson, S. Widdop, R. Caspar, M. Dimov, et al. 2011. “Design and Analysis of Cognitive Interviews for Comparative Multinational Testing.” *Field Methods* 23 (4): 379–96.
Miller, Kristen. 2007. “Design and Analysis of Cognitive Interviews for Cross National Testing.” Paper presented at the ESRA conference. Prague.
Miller, Kristen, Gordon B. Willis, Connie Eason, Lisa Moses, and Beth Canfield. 2005. “The Use of Cognitive Interviewing to Evaluate Translated Survey Questions: Lessons Learned.” In *Methodological Aspects in Cross-National Research*, edited by Janet A. Harkness and Jürgen Hoffmeyer-Zlotnik. ZUMA.
Mohler, Peter Ph, and Timothy P. Johnson. 2010. “Equivalence, Comparability, and Methodoligcal Progress.” In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 17–29. New Jersey: Wiley Hoboken, NJ.
Mohler, Peter Ph, Beth-Ellen Pennell, and Frost Hubbard. 2012. “Survey Documentation: Towards Professional Knowledge Management in Sample Surveys.” In *International Handbook of Survey Methodology*, edited by Edith D. de Leeuw, Joop Hox, and Don Dillman. New York: Routledge.
Norris, Pippa. 2009. “The Globalization of Comparative Public Opinion Research.” In *The Sage Handbook of Comparative Politics*, edited by Todd Landmann and Neil Robinson. Thousand Oaks, CA: Sage Publications Ltd.
Prestage, Yvette, Skjåk Knut, and Rory Fitzgerald. 2015. “New Tools for Complex Surveys: The DASISH Questionnaire Design Documentation Tool (QDDT) and Question Variable Data Base (QVDB).” Paper presented at CSDI workshop. 25-28 March 2015. London.
Saris, Willem, and Irmtraud Gallhofer. 2007. “Can Questions Travel Successfully?” In *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 1–31. London: Sage.
Smith, Tom. 2004. “Developing and Evaluating Cross-National Survey Instruments.” In *Methods for Testing and Evaluating Survey Questionnaires*, edited by Stanley Presser, Jenifer Rothgeb, Mick P. Couper, Elizabeth Lessler, Jean Martin, and Elanor Singer. Hoboken New Jersey: Wiley.
Smith, Tom. 2010. “The Globalization of Survey Research.” In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 485–96. New Jersey: Wiley Hoboken, NJ.

Smith, Tom. 2011. “Refining the Total Survey Error Perspective.” *International Journal of Public Opinion Research* 23 (4): 464–84. doi:10.1093/ijpor/edq052.

Smith, Tom, and Yang-Chih Fu. 2014. “The Globalization of Surveys.” GSS Cross-national Report. http://publicdata.norc.org:41000/gss/documents/CNRT/CNR34.pdf.

Stoop, Ineke. 2007. “If It Bleeds It Leads: The Impact of Media-Reported Events.” In *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 1–31. London: Sage.

Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. John Wiley & Sons.

Tourangeau, Roger. 1984. “Cognitive Science and Survey Methods: A Cognitive Perspective.” In *Cognitive Aspects of Survey Methodology: Building a Bridge Between Disciplines*, edited by Thomas Jabine, Miron Straf, Judity Tanur, and Roger Tourangeau. Washington DC: National Academy Press.

Willis, Gordon B. 2015. “The Practice of Cross-Cultural Cognitive Interviewing.” *Public Opinion Quarterly* 79 (S1): 359–95.

Winstone, Lizzy, Sally Widdop, and Rory Fitzgerald. 2016 (forthcoming, in press). “Constructing the Questionnaire: The Challenges of Measuring Attitudes Towards Democracy Across Europe.” In *How Europeans View and Evaluate Democracy?* edited by Kreisi Hanspeter and Mónica Ferrín. Oxford University press.

Portfolio of Published Work

Billiet, Jaak, Michel Philippens, Rory Fitzgerald, and Ineke Stoop. 2007. "Estimation of Response Bias in the European Social Survey: Using Information from Reluctant Respondents in Round One." *J. Off. Stat* 23 (2): 135–62.

Fitzgerald, Rory. 2015a. "Sailing in Uncharetered Waters: Structuring and Documenting Cross- National Questionnaire Design." GESIS Working paper. http://www.gesis.org/fileadmin/upload/forschung/publikationen/gesis_reihen/gesis_papers /GESIS-Papers_2015-05.pdf.

Fitzgerald, Rory, and Roger Jowell. 2010. "Measurement Equivalence in Comparative Surveys: The European Social Survey (ESS) - From Design to Implementation and Beyond." In *Survey Methods in Multinational, Multiregional, and Multicultural Contexts*, edited by Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars Lyberg, Peter Ph Mohler, Beth-Ellen Pennell, and Tom W. Smith, 485–96. New Jersey: Wiley Hoboken, NJ.

Fitzgerald, Rory, Sally Widdop, Michelle Gray, and Debbie Collins. 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): 569–99.

Fitzgerald, Rory, Lizzy Winstone, and Yvette Prestage. 2014a. "A Versatile Tool? Applying the Cross- National Error Source Typology (CNEST) to Triangulated Pre-Test Data." *FORS Working Paper Series*, no. 2. http://forscenter.ch/wp-content/uploads/2014/08/FORS_WPS_2014- 02_Fitzgerald.pdf.

Fitzgerald, Rory, Lizzy Winstone and Yvette Prestage. 2014b. "Searching For Evidence of Acculturation: Attitudes Toward Homosexuality Among Migrants Moving From Eastern to Western Europe." *International Journal of Public Opinion Research*, July, edu021. doi:10.1093/ijpor/edu021.

Harrison, Eric, and Rory Fitzgerald. 2010. "A Chorus of Disapproval? European Attitudes to Nontraditional Family Patterns." In *British Social Attitudes 26th Report*, edited by Alison Park, John Curtice, Katarina Thomson, Miranda Phillips, Elizabeth Cleary, and Sarah Butt, 135–58. London: Sage.

Jowell, Roger, Max Kaase, Rory Fitzgerald, and Gillian Eva. 2007. "The European Social Survey as a Measurement Model." In *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey*, edited by Roger Jowell, Caroline Roberts, Rory Fitzgerald, and Gillian Eva, 1–31. London: Sage.

Miller, K., R. Fitzgerald, J.-L. Padilla, S. Willson, S. Widdop, R. Caspar, M. Dimov, et al. 2011. "Design and Analysis of Cognitive Interviews for Comparative Multinational Testing." *Field Methods* 23 (4): 379–96.

Stoop, Ineke, Jaak Billiet, Achim Koch, and Rory Fitzgerald. 2010. *Improving Survey Response: Lessons Learned from the European Social Survey*. John Wiley & Sons.

Winstone, Lizzy, Sally Widdop, and Rory Fitzgerald. 2016 (forthcoming, in press). "Constructing the Questionnaire: The Challenges of Measuring Attitudes Towards Democracy Across Europe." In *How Europeans View and Evaluate Democracy?*,edited by Kreisi Hanspeter and Mónica Ferrín. Oxford University press.