8

Implementing a Multinational Study of Questionnaire Design

Henning Silber[1], *Tobias H. Stark*[2], *Annelies G. Blom*[3], *and Jon A. Krosnick*[4]

[1] *GESIS – Leibniz Institute for the Social Sciences, Mannheim, Germany*
[2] *ICS, Utrecht University, Utrecht, The Netherlands*
[3] *Department of Political Science and Collaborative Research Center 884 "Political Economy of Reforms", University of Mannheim, Mannheim, Germany*
[4] *Departments of Communication, Political Science, and Psychology, Stanford University, Stanford, CA, USA*

8.1 Introduction

The past decade has seen a rise in online panels for social scientific research. To a large extent, this is driven by the cost and time efficiency of the online mode of data collection [1]. However, the popularity of online panels has been met with criticisms regarding their ability to accurately represent their intended target populations [2, 3]. The reason for this is that most commercial online panels are based on nonprobability samples. Probability samples require that all population members have a known, nonzero probability of selection into the panel. In contrast, nonprobability online panels are typically recruited via a variety of different procedures such as self-selection by registering via the panel website, banner ads on websites, or pop-ups when surfing the web, where the selection probability of panel members remains unknown.

In recent years, in order to provide higher sample quality in online data collections, there has been a rise in online panels based on probability samples that aim to be representative of the general population. These studies typically draw their samples offline via established probability sampling procedures; for example, by sampling persons from population registers or via random digit dialing (see Ref. [4] for an overview). Some of these panels include persons who did not previously have a computer or Internet access at home. The study therefore takes into account coverage error by collecting information about offline panel members and then provides the equipment needed to participate [5, 6].

Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC), First Edition. Edited by Timothy P. Johnson, Beth-Ellen Pennell, Ineke A.L. Stoop, and Brita Dorer.
© 2019 John Wiley & Sons, Inc. Published 2019 by John Wiley & Sons, Inc.

Probability sample surveys are also particularly valuable in light of the growing need for internationally comparable social scientific data, as researchers aspire to test theories that have been verified only in a single country. This need for cross-national data has led to a variety of large-scale social research projects conducted in probability sample face-to-face surveys with hour-long interviews. Such large-scale cross-national projects typically focus on a single broad research topic, such as social attitudes, health, or education. For smaller projects, however, such large, face-to-face undertakings are not viable, because of the cost and time required. Furthermore, experimental research often relies on complex programming of randomization, filters, and editing checks. In such a situation, probability sample online panels might be an attractive alternative.

The multinational study of questionnaire design (MSQD) explored whether the principles of question design derived primarily from American research decades ago still apply in the United States today and, if so, can be generalized to other countries. For this purpose, the project utilized probability sample online panels from around the world, as well as a few other data collection modes.

In this chapter, we lay out the design of the MSQD and the challenges faced when implementing the project across countries. In particular, we elaborate on the sampling and online implementation of the questionnaire, as well as the questionnaire design experiments selected for the study. We discuss challenges faced in the translation of experiments in which question wording plays a central role. We also present a few exemplary results from the study.

8.2 Scope of the MSQD

The MSQD implemented well-tested split-ballot design experiments in single-country contexts in multiple countries to gauge country-specific differences in response behavior, satisficing, and social desirability bias [7].

For nearly a century, experimental methodology has been very helpful in the study of questionnaire design (e.g. Refs. [8–15]). However, the vast majority of this work has been conducted in the United States, which might limit the generalizability of results to other cultural and linguistic contexts. An increasing amount of such work is now being conducted in other countries, but this work is less well documented in the literature [16]. So far, scientists who studied multinational settings mostly implemented research on questionnaire translation and language accuracy, which is of course fundamental for multinational survey projects. However, Yang et al. [16] suggested that culturally founded response behavior might be equally important and should therefore also be investigated.

In fact, there have been remarkably few attempts to conduct the same question design experiment across countries to ascertain whether principles of optimal design can be transferred from the United States to other countries. The few existing multinational survey question studies have tested either very global hypotheses about national differences based on cultural response styles (such as patterns of individualism in Western societies and more collectivism-based response behavior in Eastern societies, e.g. Refs. [16–18]) or reported findings about cultural aspects of response behavior, such as differences in masculinity, power distance, uncertainty avoidance, communication style, cognitive editing, and cultural norms based on data from homogeneous subsamples (e.g. students) not collected using identical questionnaires during the same time period [16, 19]. In addition, the work conducted in the United States identified education as an important moderator of response effects (many such effects occurred more among less educated respondents; see Ref. [20]), presumably because educational attainment correlates with cognitive skills. It would be of theoretical value to ascertain whether the same moderation appears across cultures.

Should survey researchers from other countries expect the same question form, wording, and context effects across countries, or should they expect different data patterns? The answer to this question hinges on the cognitive mechanism(s) that explain why the effects occurred in the United States. If some question design effects occur because of culture-specific response behaviors (e.g. the tendency to defer to seemingly higher-status researchers; the tendency to express opinions regardless of confidence in them), then we might expect to see the same effects in countries whose cultures work according to the same social norms and might not see those effects in countries with different social norms. If some question design effects occur because of strategies that all respondents in all countries implement when they lack motivation and/or ability to answer questions optimally (e.g. Ref. [21]), then we might expect to see the same effects across all countries.

Theoretically, the ideal process of responding to a survey question is often thought of as entailing four cognitive steps: interpreting the question, retrieving information from memory, integrating the information, and reporting the answer (e.g. Refs. [14, 15]). The satisficing theory [21] describes how, why, and when respondents may process information differently while answering survey questions. Specifically, respondents may truncate or skip one or more of the four fundamental cognitive steps, thus producing response effects [21]. The likelihood of satisficing is believed to depend on three factors: respondents' ability to optimize and carefully think about their answers, respondents' motivation to optimize, and the cognitive difficulty of optimizing inherent to the survey question.

If some question design experiments produce similar results across countries, differences across respondents can then perhaps be explained by the theory of

survey satisficing, and the relevant guidelines for optimal questionnaire design that have been developed based on experiments conducted in the United States would appear to be applicable in other countries. If some results fail to appear in some countries, this would suggest that (i) explanations for the effects may not hold across countries and (ii) principles of optimal questionnaire design might need to vary across countries.

Some past research anticipates some response effects that differ across countries, because response styles that affect the determinants of survey satisficing appear to vary across countries and cultural regions [16–18]. Culture-specific communication styles and cultural differences in cognitive editing may also be related to differential perception of the difficulty of a task. In this respect, questionnaire translation plays a key role, if different substantively equivalent translations of specific words are related to different communication styles and cognitive editing. The subtleties in the language that trigger response effects might thus be generated by or lost in translation. Finally, levels of education differ greatly across countries and may thereby predict differences in cognitive skills of respondents and therefore also differences in the magnitudes or presence of some response effects.

To explore these issues and gauge the extent to which principles of questionnaire design generalize across nations, we conducted a series of experiments in 14 countries: the United States, Canada, Denmark, France, Germany, Japan, Iceland, the Netherlands, Norway, Portugal, Spain, Sweden, Taiwan, and the United Kingdom. The source questionnaire was written in English, and we aimed at achieving functionally equivalent translations in the various different languages (see Section 8.5 for a description of the translation process).

8.3 Design of the MSQD

The MSQD project was coordinated by a core project team consisting of the authors of this chapter. In addition, a global project team included researchers from each of the participating countries amounting to more than 20 researchers in total.

The core project team wrote specifications for sampling, translation, fieldwork procedures, and sample sizes. Each sample had to be a randomly selected probability sample of the general population of all adults living in the particular country, with little or no noncoverage. Each sample had to include at least 1000 respondents. Samples of specific subpopulations (e.g. students) were not acceptable, nor were data collected from nonprobability samples.

We preferred collecting data from adult respondents who belonged to a probability sampled online panel. This means that every adult resident of the country should have had a known, nonzero probability of being invited to join the panel and that the individuals invited to join the panel were selected via probability

sampling from the population. This also entails that people with and people without Internet access should be included in the population (e.g. Refs. [6, 22]).

The questionnaire to be implemented in each country included programming instructions for randomizations, filters, and edit checks. The questionnaire was provided in American English, and each country's project team translated the questionnaire into their national language(s) by means of TRAPD translation procedures (see Refs. [23–26]).

Members of the global project team were responsible for implementing the MSQD in their country according to these specifications. Typically, the MSQD was implemented as part of a larger survey data collection, for example, by adding the questions to a wave of data collection from an existing panel or as an add-on to a cross-sectional survey.

Recruiting the MSQD global project team was challenging. The project started on a small scale with only four countries. In these countries, the core project team submitted a proposal to open calls for questionnaire proposals for ongoing panel studies. In addition, we spread the word about the project at workshops, conferences, and through relevant mailing lists. In addition to the researchers who eventually joined the global project team of the MSQD, we were also contacted by researchers from Chile, China, Columbia, Costa Rica, Finland, Estonia, India, Russia, Slovenia, and Switzerland. Unfortunately, these countries were ultimately not able to join the project due to a lack of funding.

Details regarding the members of the global project team and the MSQD implementation in each country are shown in Table 8.1.

8.4 Experiments Implemented in the MSQD

The aim of the MSQD was to conduct well-cited question design experiments originally conducted in the United States and assess whether similar results would be observed decades later in the United States and in other countries. When selecting the experiments to be implemented, we applied the following criteria:

1) In their seminal book, Schuman and Presser [27] reported many tests of response effects. Their results are still widely cited, and many best practice guidelines for questionnaire design are based on these experiments and findings. Accordingly, most of the experiments implemented in the MSQD are experiments reported by Schuman and Presser [27].
2) Of the eligible experiments in Schuman and Presser's [27] book, some involve issues that are not relevant today. One example is this:

> Looking back, do you think our government did too much to help the South Vietnamese government in the war, about the right amount, or not enough to help the South Vietnamese government?

Table 8.1 The MSQD implementation across participating organizations.

Country	Organization/panel	Project team	*N*	Mode	Survey type	Sample type
Canada	University of Saskatchewan	Karen Lawson	1317	O	Specifically recruited sample or existing online panel	General population without offliners
Denmark	University of Aalborg	Sanne Lund Clement, Ditte Shamshiri-Petersen	1325	O, M, T	Part of another data collection	General population
France	ELIPSS Panel, Sciences Po	Anne Cornilleau, Anne-Sophie Cousteaux, core team	835	O	Existing online panel	General population
Germany 1	German Internet Panel, University of Mannheim	Annelies Blom	1137	O	Existing online panel	General population
Germany 2	GESIS Panel, GESIS – Leibniz Institute for the Social Sciences	Michael Bosnjak, core team	4221	O, M	Existing mixed-mode panel	General population
Japan	National Institute for Environmental Studies	Midori Aoyagi	1548	F	Part of another data collection	General population
Iceland	University of Iceland	Guðbjörg Andrea Jónsdóttir	3141	O	Existing online panel	General population
Netherlands	LISS Panel, CentERdata	Core team	2257	O	Existing online panel	General population
Norway	Citizen Panel, University of Bergen	Endre Tvinnereim, core team	5489	O	Existing online panel	General population without offliners
Portugal	University Institute of Lisbon	Ana Belchior	1204	O, T	Part of another data collection	General population with telephone
Spain	Centro de Investigaciones Sociologicas	Mónica Méndez Lago	NA	O	Specifically recruited sample	General population without offliners

Sweden	Citizen Panel, University of Gothenburg	Johan Martinsson	1770	O	Existing online panel	General population without offliners
Taiwan	Academia Sinica	Ruoh-rong Yu, Pei-shan Liao, Su-hao Tu	790	O	Follow-up study to another data collection	General population without offliners
United Kingdom	Understanding Society Innovation Panel, University of Essex	Peter Lynn, core team	2262	O, F	Existing mixed-mode panel	General population
United States 1	Knowledge Panel, GfK	Core team	1029	O	Existing online panel	General population
United States 2	Gallup Panel	Core team	2012	O	Existing online panel	General population without offliners

F, Face-to-face; M, Mail; NA, Not yet available; O, Online; T, Telephone.

3) We sought to select experiments whose question wordings would be meaningful outside of the United States. Some of Schuman and Presser's experiments could be implemented in the United States now, but would not have the same meaning if currently asked in another country now. An example is:

> Would you favor a law which would require a person to obtain a police permit before he could buy a gun?

This question only makes sense in a country where guns can be purchased without police permits. That is, the word "would" in the question implies that the question proposes a hypothetical condition for respondents to evaluate. It would not make sense to ask this question in a country that already requires police permission to buy a gun.

4) A statistically significant effect of the experimental variation had previously been found. Thus, we only selected experiments that yielded a statistically significant effect in a prior implementation.
5) The experimental manipulations were diverse. We tried to include experiments with as many manipulations as possible and aimed to include multiple experiments of every manipulation type.

Even though Schuman and Presser [27] reported numerous experiments investigating the impact of question and questionnaire design on response behavior, our criteria yielded a relatively small selection of experiments for our study. To augment this small pool, we incorporated four additional experiments that were not reported by Schuman and Presser [27]. These experiments were selected following the selection criteria 2–5 above. Three of the additional experiments investigated response order effects, and one investigated question order effects. They included, for example, a response order experiment from a Stanford University survey about global warming comparing the following two questions:

> Form A: As far as you know, would you say that average temperatures around the world have been higher in the last three years than before that, lower, or about the same? (Response Categories: Higher, Lower, About the same)
>
> Form B: As far as you know, would you say that average temperatures around the world have been lower in the last three years than before that, higher, or about the same? (Response Categories: Lower, Higher, About the same)

Additional sources for experiment were Stanford University's Face-to-Face Recruited Internet Survey Panel (FFRISP) from 2009 (Krosnick et al., work in preparation. *Combining the Best with the Best for Survey Research: Creating the*

Face-to-Face Recruited Internet Survey Platform. Stanford, CA), and a paper published by Schuman and Ludwig [28]. Table 8.2 lists all experiments implemented as part of the MSQD. The question wordings in English, the experimental groups, and the translated questionnaires as implemented by the global project team can be found at Krosnick [29].

The experiments tested for differences in response behavior produced in the following ways:

1) By altering the order in which response options are presented.
2) By altering the order in which questions are asked.
3) By varying question wording to test for acquiescence response bias (the tendency to agree with a presented statement).

Table 8.2 Overview of the experiments.

Experiment	Source	Version	Manipulations
Oil supply	S&P	4	Response order, some/others
Oil prices	S&P	4	Response order, agree/disagree
Adequate housing	S&P	4	Response order, some/others
Individuals and social conditions	S&P	4	Acquiescence, response order
Jobs	S&P	4	Acquiescence, response order
Women in politics	S&P	4	Acquiescence, response order
Complicated	S&P	2	Acquiescence
Free speech	S&P	2	Question wording
Global warming	SGWP	2	Response order
Courts	S&P	4	No opinion
Leaders smart	S&P	4	No opinion, response order
Leaders crooked	S&P	4	No opinion
Fuel shortage	S&P	4	Question balance, response order, counterargument
Unions	S&P	4	Question balance, counterargument
Abortion	S&P	2	Question order
Unions and businesses	S&L	2	Question order
Trust	FFRISP	2	Response order
Inequality	FFRISP	2	Response order

FFRISP = Stanford University's Face-to-Face Recruited Internet Survey Panel 2009; SGWP = 2012 Stanford Global Warming Survey; S&L = Schuman and Ludwig [28]; S&P = Schuman and Presser [27].

4) By varying the presence or absence of various no opinion response options (option 1: not enough information; option 2: no opinion; option 3: don't know).
5) By using the mentioning of "some people" and "other people" in an effort to balance a question (e.g. "Some people feel the government should see to it that all people have adequate housing, while others feel each person should provide for his or her own housing. Which comes closest to how you feel about this?").
6) By varying the presence of a counterargument.

Each experiment had either two or four versions of a question or question sequence and up to three manipulations (see Table 8.2).

8.5 Translation Requirements and Procedures

When implementing a cross-national survey project across 14 countries, the source questionnaire must be translated into multiple languages. The core team used existing questions that had previously been fielded in the United States, that is, they had been drafted in American English (see Section 8.4 for a description of the item selection process). This was for two reasons. First, all experiments that we aimed to conduct were originally conducted in American English. Second, English is the most widely spoken language in survey research, which is why large social science projects typically develop their source questionnaires in English, although some projects have created two source questionnaires, for example, English and French, e.g. the Eurobarometer and the Programme for International Student Assessment (PISA) [30].

The goal of a questionnaire translation should be to achieve a functionally equivalent version in the target language [31]. Usually in survey research, this means that one follows an ask-the-same question approach, where the questions are translated such that the same concept is measured on the same measurement scale across languages [25].

To achieve functionally equivalent translations for the MSQD, we followed the translation, review, adjudication, pretesting, and documentation (TRAPD) approach developed by Janet Harkness and colleagues [23–26]. This meant that in every country, at least two translators with a background in survey research separately drafted a full translation of the questions (T). Then, the translators, together with the national project head, reviewed the two drafts (R) to produce one joint translation (A). The resulting translated questionnaires were subsequently implemented in the survey. The translated and programmed questionnaires were carefully proofread, and their randomizations and filters tested by the researchers from the global team and also by researchers from the

core team to ensure that the experiments were correctly programmed and the question wordings and orderings matched the source questionnaire. In addition, most countries conducted a dress rehearsal pretest as part of their usual fieldwork procedures (P). Due to budget constraints, cognitive pretests were not conducted. All translations and screenshots of each question were documented, alongside detected deviations (D).

The questionnaire was translated into 11 languages, including Chinese, Danish, Dutch, French, German, Icelandic, Japanese, Norwegian, Portuguese, Spanish, and Swedish. Teams from countries with shared languages (such as French in France and Canada) were encouraged to exchange their translations. However, researchers from different countries did not work on joint translations.

8.6 Findings on Feasibility and Limitations Due to Translations and Required Adaptations

Replication of psychological research involves a new research group repeating an existing experiment using the same methods with different subjects. In a survey research, this means that exactly the same questions should be asked to a new group of respondents. When translating survey questions into different languages, however, a strict replication is not feasible and cannot be the goal of a cross-national study, because every translation introduces changes in meaning, even though they might be subtle. Therefore, the MSQD investigated whether the questionnaire design principles replicated in the United States only. When implementing the questionnaire in other countries, however, we aimed to assess whether the questionnaire design effects reappear in other contexts, i.e. whether they can be generalized across countries.

Nonetheless, the question wording played a pivotal role for our test of generalizability in many MSQD experiments. In particular, when investigating acquiescence, question balance, and counterargument effects, testing the generalizability is only meaningful when key formulations in the question are functionally equivalent to the source version. For example, to test for acquiescence effects, cross-national generalizability can only be evaluated if close translations of the words “agree” and “disagree” are used in the target questionnaire. A translation in the gist of “I think so” and “I don’t think so” or “I believe” and “I don’t believe” will not be a true test of an acquiescence effect. For such key formulations, the translation has to stay close to the source to ensure functional equivalence of the whole question with regard to the questionnaire design effect under investigation.

To achieve this, we annotated the MSQD questionnaire for the country teams, marking words that had to be translated as closely as possible to the

source words.[1] Unfortunately, we became aware of the need for exact translations of key formulations only during the data collection phase, when many translations and data collections had already been implemented. As a consequence, we revisited the translations for all countries together with the country teams after the data collection to cross-check whether the translations were functionally equivalent for our purpose. This process revealed that for some experiments in some countries, the translations were not functionally equivalent to test the generalizability of a questionnaire design principle. These country/experiment combinations are excluded from our analyses.

In addition, the translation process taught us that it can be difficult for researchers to have complete confidence in translations into unfamiliar languages, especially if communication with country teams in English is challenging. In the case of two countries that participated in the MSQD, Japan and Taiwan, the translated questionnaires are written in ideograph languages unfamiliar to all members of the core project team. To enable a closer evaluation of the translated questionnaires, a company specialized in survey translations was therefore hired to evaluate the translated Japanese and Taiwanese questionnaires. The translation evaluators were instructed to give special attention and report deviations for key formulations where we had asked the translating teams to produce translations that should be as close as possible.

In addition to the translation issues that were specific to our methodological research aims, we also encountered queries that were of a rather topical nature, because survey translations for substantive research always entail some need for adaptation to the national contexts. Researchers in several countries, including France, the Netherlands, Norway, Spain, Taiwan, and Sweden, suggested country-specific adaptations of the experiments. This concerned, for instance, a question balance experiment about attitudes toward fuel consumption for heating homes:

> If there is a serious fuel shortage this winter, do you think there should be a law requiring people to lower the heat in their homes, or do you oppose such a law? (Response Categories: Should be a law, Oppose such a law)

As we learned, people in Norway heat their homes with hydropower instead of fuel. Because the word "fuel" was not a key formulation of the question wording experiment on "fuel shortage," we allowed an adaptation in this case and used the Norwegian term "energi," which literally translates to "energy" instead of "fuel." In Taiwan, the homes are rarely heated because it is a subtropical

1 The instruction given to the translators was: "Please translate the highlighted words and expressions as close to the English wording as possible."

country with generally high temperatures all year. Therefore, we also employed an adaptation of the question by using air conditioners instead of heating.

Another instance of an adaptation occurred with an experiment comparing these two questions:

> Form A: In general, do you think the courts in this area deal too harshly or not harshly enough with criminals, or do you not have an opinion on that? (Response Categories: Too harshly, Not harshly enough, No opinion)
> Form B: In general, do you think the courts in this area deal too harshly or not harshly enough with criminals? (Response Categories: Too harshly, Not harshly enough)

The German court system is organized differently than the US court system, in a way that made the reference "in this area" nonsensical in Germany. Therefore, this phrase was dropped in the German translation.

In other situations, country teams called for adaptations or even for leaving out questions, because of a lack of societal relevance. A particularly contested example was the following questions measuring attitudes toward abortion:

> Do you think it should be possible for a pregnant woman to obtain a legal abortion if she is married and does not want any more children? (Response Categories: Yes, No)
> Do you think it should be possible for a pregnant woman to obtain a legal abortion if there is a strong chance of serious defect in the baby? (Response Categories: Yes, No)

A number of country teams claimed that these questions were unsuitable in their country context, because they expected little variation in opinions and therefore little variation in survey responses. We addressed this concern by presenting recent findings from the World Values Survey (WVS) (2005–2009), which showed sufficient variation in public opinion on this issue in all participating countries. In addition, differences in agreement rates were of interest to our research into the generalizability of this question order effect. In the end, most country teams agreed to also implement the abortion questions.

8.7 Example Results

The abortion experiment yielded very interesting results. Schuman and Presser [27] demonstrated that support for abortion by a married woman dropped considerably when that question was preceded by the question about a birth defect. Two explanations have been considered for the effect: perceptual contrast and subtraction. According to the first explanation, a birth defect seems like a much

better reason for an abortion than the desire for no more children, and considering the strong reason (birth defect) before considering the weaker reason (no more children) makes the weaker reason seem even weaker. According to the subtraction explanation, respondents who are asked about the married woman question first might assume that one reason she might not want more children is because the baby might have a high risk of a birth defect. That is, the birth defect reason might be encompassed within the married woman's situation, thus justifying her desire. But if the birth defect question is asked before the married woman question, respondents might assume that the second question is not meant to include the reason already asked about (birth defect), thus making the married woman's situation less compelling.

Both of these hypotheses might lead to the expectation that the more support a country expresses for abortion by the married woman, the more likely a question order effect is to occur, because there is more room for approval to drop as the result of considering the birth defect first. Across countries, there was considerable variation in the degree of support of abortion by a married woman. Support ranged from 56.3% in the United States (TESS) to 93.4% in Sweden (see Table 8.3).

As shown in Table 8.3, even in countries with very high levels of support, such as in Sweden and Denmark, statistically significant question order effects appeared in the expected direction. Support for the married woman's right to abortion dropped when respondents were first asked whether it should be possible to obtain a legal abortion in the case of a high risk of a serious defect in the baby. In Denmark, for instance, support for legal abortion for a married woman dropped from 91.5 to 81.6% (χ^2 (1) = 27.35, $p < 0.001$) when this question was asked after the birth defect question. This question order effect appeared in all but one country, namely, Japan (see Table 8.3).

The size of the question order effect (shown in column 3 of Table 8.3) was related to the starting point of support for abortion by the married woman (shown in column 1 of Table 8.3). Specifically, the more people supported the married woman's right to an abortion when asked that question first, the larger the order effect was ($r = 0.37$). This was not consistently true, though. Among the countries manifesting the highest levels of initial support were Sweden (which manifested one of the smallest question order effects, three percentage points) and Iceland (which manifested a moderate question order effect of 10 percentage points). But in general, the higher the country started, the farther it tended to fall due to the question order manipulation.

The lack of the question order effect in Japan raises the possibility that translation might be the cause. In that country, 41.2% of respondents supported abortion for the married woman when asked that question first, and this percentage remained unchanged when the birth defect question preceded it (χ^2 (1) = 1.7, $p = 0.43$, Table 8.3). This null result is unexpected. To attempt to understand this effect, we first consulted a native Japanese woman

Table 8.3 Question order experiment on attitudes toward abortion for a married woman.

Country	% Yes when asked first	% Yes when asked second	Difference	χ^2	N^a
US S&P 1981[b]	60.7	48.1	−12.6	9.52**	293
US Gallup	65.0	54.8	−10.2	20.96***	1963
US TESS	56.3	50.8	−5.5	3.06	1015
Canada	80.1	72.1	−8.0	11.48**	1309
Sweden	93.4	90.0	−3.4	6.64*	1718
Denmark	91.5	81.6	−9.9	27.35***	1308
Norway	85.9	75.2	−10.7	28.66***	1584
Iceland	86.2	75.8	−10.4	51.99***	2984
Germany GIP	80.2	59.6	−20.6	53.55***	1048
Germany GESIS	77.7	56.9	−20.8	205.44***	4188
The Netherlands	72.7	60.8	−5.9	35.72***	2243
UK total	76.8	64.1	−12.7	42.84***	2183
UK online	76.6	67.8	−8.8	6.86**	705
Portugal	66.4	52.0	−14.4	25.92***	1204
Taiwan	77.3	66.3	−11.0	11.74**	789
Japan	41.2	42.9	1.7	0.43	1471

[a] N refers to the number of cases without missing values. In Norway, only a subset of respondents was asked these questions.
[b] US S&P 1981 refers to the original results presented by Schuman and Presser [27].
***$p < 0.001$; **$p < 0.01$; *$p < 0.05$.

who is thoroughly fluent in both Japanese and English (though not a professional translator). She explained various ways that the Japanese translation could be interpreted, and some of these interpretations may be inconsistent with the intended meaning of the question in English. These deviations from intended meaning would undermine the question order effect. We therefore have commissioned a professional review of the Japanese translation to help us resolve this issue. If this review reveals that the abortion questions were translated into Japanese with clearly unintended and problematic meanings, this sample will be removed from the analyses.[2]

The data from this experiment (and all other experiments we conducted) allow for the testing of Schuman and Presser's [27] "form-resistant correlation"

2 Unfortunately, the results of this review were not available when finalizing this chapter.

hypothesis. Schuman and Presser found that correlations between measurements were remarkably consistent regardless of changes in the forming, wording, or ordering of the questions. So, for example, answers to a question measuring an opinion correlated with the age of the respondent similarly, regardless of whether the opinion question was in one form or another.

We explored this issue with the present data by exploring whether the rank ordering and spacing between countries in terms of their opinions were maintained regardless of the order in which the questions were asked. The results were indeed strikingly powerful. Treating country as the unit of analysis, the correlation between answers to the married woman question when asked first versus second (column 2 vs. column 3 in Table 8.3) was 0.91. Thus, researchers interested in studying cross-national differences in opinions on this issue would reach nearly identical conclusions regardless of which question order was used to make the measurements.

8.8 Conclusion

Planning and organizing a cross-national questionnaire design project like the MSQD is challenging, because the potential for implementation errors and thus blind spots in cross-national inference are multiplied [32]. Our project started out on a small scale and grew tremendously over time. The most successful strategy to secure data collection in a country was submitting our project proposal to open calls for data collection in existing random probability online panels. We first submitted the project to two open calls and later to many more. To date, the MSQD has been accepted by open calls for proposals for survey research in seven countries, including Germany, France, the Netherlands, Norway, Sweden, the United Kingdom, and the United States. Finally, we reached additional research groups by a call through various mailing lists. We were overwhelmed by the number of responses we got to the invitation, and to date, researchers from seven countries joined our project this way.

When starting out, we did not anticipate the large-scale project that the MSQD eventually became. As a consequence, we had to adapt our strategies and materials multiple times throughout the data collection phase to cater for the many requirements and queries by the country teams. This is not ideal, as it would be best to keep everything stable throughout the duration of the project. In addition, the very different organizational structures, local project groups, and data collection methods used required a lot of flexibility from our side.

We still consider the MSQD a small-scale project, which is not comparable in organization and resources to established cross-national surveys like the European Social Survey (ESS), the WVS, the European Values Study (EVS), and the Survey of Health, Aging and Retirement in Europe (SHARE). Despite the smaller scale of the MSQD, however, the most important lesson we learned in terms of project organization was about the importance of securing funding for

central project coordination and for country implementations well in advance to allow for a more structured approach. This is particularly important when many collaborators collect data at the same time or when it is necessary to follow up with data collection companies to evaluate their data collection strategies.

In this chapter, we also provide a first teaser of the results of the MSQD project to illustrate the value of testing the generalizability of question design principles. Despite reservations across several countries about the high levels of acceptance of legal abortions, we found that respondents across countries evaluated the two questions differently and that the order, in which the questions were posed, was important despite the high acceptance rates in the Nordic countries. This first result suggests that general guidelines developed in the United States (such as the question order effect, as shown by Schuman and Presser) may apply to other countries – at least for this particular type of question design effect. We will continue to investigate whether this finding holds for other types of question design issues as well.

The findings discussed in this chapter thus demonstrate that a cross-cultural and a cross-national view on findings and theories that have been well established in one country can be beneficial for expanding questionnaire design insights across countries.

Another finding described in this chapter relates to the fact that the rise of online panels employing probability samples in recent years opened new avenues for researchers interested in small-scale substantive or methodological research. Those changes in the survey research infrastructure enable small teams of researchers to organize large-scale international survey projects, even when financial resources are limited.

Acknowledgments

The authors thank Midori Aoyagi (Japan), Ana Belchior (Portugal), Michael Bosnjak (Germany), Anne Cornilleau (France), Sophie Cousteaux (France), Franziska Gebhard (Germany), Melvin John (Germany), Guðbjörg Andrea Jónsdóttir (Iceland), Karen Lawson (Canada), Pei-shan Liao (Taiwan), Sanne Lund Clement (Denmark), Peter Lynn (UK), Jenny Marlar (United States), Johan Martinsson (Sweden), Mónica Méndez Lago (Spain), Ditte Shamhiri-Petersen (Denmark), Su-hao Tu (Taiwan), Endre Tvinnereim (Norway), and Ruoh-rong Yu (Taiwan) for their contributions to the multinational study of questionnaire design.

References

1 Dillman, D.A. and Bowker, D.K. (2001). The web questionnaire challenge to survey methodologists. In: *Dimensions of Internet Science* (ed. U.-D. Reips and M. Bosnjak), 159–178. Lengerich, Germany: Pabst Science Publishers.

2 Bethlehem, J. and Stoop, I. (2007). Online panels – a paradigm theft? In: *The Challenges of a Changing World* (ed. M. Trotman et al.), 113–131. Southampton, UK: ASC Proceedings of the Fifth International Conference of the Association for Survey Computing. University of Southampton (12–14 September).
3 Couper, M.P. (2000). Review: web surveys: a review of issues and approaches. *Public Opinion Quarterly* 64 (4): 464–494.
4 Blom, A.G., Bosnjak, M., Cornilleau, A. et al. (2016). A comparison of four probability sample online and mixed-mode panels in Europe. *Social Science Computer Review* 34 (1): 8–25.
5 Leenheer, J. and Scherpenzeel, A. (2013). Does it pay off to include non-internet households in an internet panel? *International Journal of Internet Science* 8: 17–29.
6 Blom, A.G., Herzing, J.M.E., Cornesse, C. et al. (2016). Does the recruitment of offline households increase the sample representativeness of probability sample online panels? Evidence from the German Internet Panel. *Social Science Computer Review* doi: 10.1177/0894439316651584.
7 De Maio, T.D. (1984). Social desirability and survey measurement: a review. In: *Surveying Subjective Phenomena*, vol. 2 (ed. C.E. Turner and E. Martin), 257–282. New York: Russell Sage.
8 Cantril, H. (1940). Problems and techniques experiments in the wording of questions. *Public Opinion Quarterly* 4 (2): 330–332.
9 Dillman, D.A. (2000). *Mail and Internet Surveys. The Tailored Design Method*. Hoboken, NJ: Wiley.
10 Gallup, G.H. (1941). Question wording in public opinion polls. *Sociometry* 4: 259–268.
11 Gallup, G.H. (1947). The quintamensional plan of question design. *Public Opinion Quarterly* 11: 385–393.
12 Rugg, D. and Cantril, H. (1944). The wording of questions. In: *Gauging Public Opinion* (ed. H. Cantril), 23–50. Princeton, NJ: Princeton University Press.
13 Saris, W.E. and Gallhofer, I.N. (2007). *Design, Evaluation and Analysis of Questionnaires for Survey Research*. Hoboken, NJ: Wiley.
14 Schwarz, N. and Sudman, S. (1996). *Answering Questions. Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco, CA: Jossey-Bass.
15 Tourangeau, R.L., Rips, J., and Rasinski, K. (2000). *The Psychology of Survey Response*. New York: Cambridge University Press.
16 Yang, Y., Harkness, J.A., Chin, T.-Y., and Villar, A. (2010). Response styles and culture. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (ed. J.A. Harkness, M. Braun, B. Edwards, et al.), 203–223. Hoboken, NJ: Wiley.
17 Schwarz, N., Oyserman, D., and Peytcheva, E. (2010). Cognition, communication, and culture: implications for the survey response process. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (ed. J.A. Harkness, M. Braun, B. Edwards, et al.), 177–190. Hoboken, NJ: Wiley.

18 Uskul, A.K., Oyserman, D., and Schwarz, N. (2010). Cultural emphasis on honor, modesty, or self-enhancement: implications for the survey-response process. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (ed. J.A. Harkness, M. Braun, B. Edwards, et al.), 191–201. Hoboken, NJ: Wiley.

19 Johnson, T.P., Kulesa, P., Cho, Y.I., and Shavitt, S. (2005). The relation between culture and response style: evidence from 19 countries. *Journal of Cross-Cultural Psychology* 36: 264–277.

20 Narayan, S. and Krosnick, J.A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly* 60 (1): 58–88.

21 Krosnick, J.A. (1991). Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 50: 537–567.

22 Scherpenzeel, A. (2011). Data collection in a probability sample internet panel: how the LISS panel was built and how it can be used. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique* 109 (1): 56–61.

23 Harkness, J.A. (2003). Questionnaire translation. In: *Cross-Cultural Survey Methods* (ed. J.A. Harkness, F.J. van de Vijver and P.P. Mohler), 35–56. Hoboken, NJ: Wiley.

24 Harkness, J.A. (2007). Improving the comparability of translations. In: *Measuring Attitudes Cross-Nationally: Lessons from the European Social Survey* (ed. R. Jowell, C. Roberts, R. Fitzgerald and G. Eva), 79–95. London: Sage.

25 Harkness, J.A., Villar, A., and Edwards, B. (2010). Translation, adaptation, and design. In: *Survey Methods in Multinational, Multiregional, and Multicultural Contexts* (ed. J.A. Harkness, M. Braun, B. Edwards, et al.), 117–140. Hoboken, NJ: Wiley.

26 Survey Research Center (2016). *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/ (accessed 5 March 2018).

27 Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Surveys*. San Diego, CA: Academic Press.

28 Schuman, H. and Ludwig, J. (1983). The norm of even-handedness in surveys as in life. *American Sociological Review* 48 (1): 112–120.

29 Krosnick, J. (2017). Current research projects. https://pprg.stanford.edu/krosnick-research-projects/ (accessed 5 March 2018).

30 Behr, D. and Scholz, E. (2011). Questionnaire translation in cross-national survey research. *Methods Data Analyses* 5 (2): 157–179.

31 Harkness, J., Pennell, B.-E., and Schoua-Glusberg, A. (2004). Survey questionnaire translation and assessment. In: *Methods for Testing and Evaluating Survey Questionnaires* (ed. S. Presser, J.M. Rothgeb, M.P. Couper, et al.), 453–473. San Diego, CA: Academic Press.

32 Lynn, P. (2003). Developing quality standards for cross-national survey research: five approaches. *International Journal of Social Research Methodology* 6: 323–336.