SCUOLA DI DOTTORATO

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Department of Sociology and Social Research

PhD program in Applied Sociology and Methodology of Social Research

Cycle XXXI

# CAN WE TRUST DATA COLLECTED USING WEB SURVEYS? ASSESSING THE QUALITY OF AN ITALIAN NON-PROBABILITY ONLINE PANEL

Surname RESPI Name CHIARA

Registration number 062190

Tutor: professor EMANUELA MARIA SALA

Co-tutor: professor KATJA LOZAR MANFREDA

Coordinator: professor CARMEN LECCARDI

**ACADEMIC YEAR 2017/2018**

# Contents

# Introduction

In this Introduction I will provide i) a short overview of the history of sample surveys and of the technological innovations that spearheaded the development of the different modes of data collection, ii) rationale, aim and contributions of my study, and iii) contents of the thesis.

This thesis sits in the survey-methodology field and explores the challenging concept of data quality when using online panels as a sample source in the survey industry. Sample surveys are one of the most important data collection methods in many disciplines. Over time, such surveys implemented different techniques to overcome various challenges: mainly face-to-face interviews until the 1970s, telephone interviews in the 1980s, and web surveys from the 1990s on. In particular, high costs, need for timely data delivery, and undercoverage (in the case of telephone surveys) are the main drivers of the development of online panels in survey research. Online panels are large pools of registered people who agreed to take part in web-based research in exchange for some form of incentive. They have the advantages of i) reducing both costs and time devoted to data collection and delivery, and ii) providing a sampling frame of registered individuals who consented to participate in surveys. The main objective of this thesis is to empirically assess the quality of primary data collected from an Italian non-probability online-panel survey. More specifically, the overall aim is to compare the estimates obtained from this web survey to those obtained from a probability-based reference survey conducted by the Italian National Institute of Statistics (ISTAT), used as benchmarks.

Sample surveys are the main data collection method for quantitative research; they represent "(relatively) systematic, (mostly) standardized approaches to collecting information on individuals, households, organizations, or larger organized entities through questioning systematically identified samples" (Marsden and Wright, 2010, p. 3). Sample surveys are now used (in Italy and elsewhere) in several research areas: market research, political and opinion polls, and social research. Modern sample surveys were firstly developed in the United States in the 1820s. Only later, in the 1930s, did they spread across Europe and then to Latin America, Asia and Africa (Gobo and Mauceri, 2014). Their origins date far back in time and a prominent early reason for the surveys was to contribute to the understanding of a social problem (Groves *et al.*, 2009). Several scholars (e.g. Marsden and Wright, 2010) identify Charles Booth's research on the London working class conditions of 1890 as the first modern sample survey. However, its distinctive features were already present in previous research. For example, the application of statistical analysis to empirical observations dates back to 1662, when John Graunt published the seminal work *Observations upon the Bills of Mortality*. The probability theory and the statistical inference appear in the works of Bayes, Gauss and La Place between the 18th and the 19th century, and Le Play probably invented in the late 1840s the first prototype of the questionnaire, used to collect information about family budgets (Gobo and Mauceri, 2014). The protocols for standardised data collection were already employed in the 1790 United States Census, in the 1801 British Census and in the research on working conditions and education of the British population from the Manchester Statistical Society in 1834. At the beginning of the 20th century, Neyman (1934) formulated the sampling theory in its modern form. Previously, all investigations followed the census model, which envisaged interviewing all the members of a population. Between 1920 and 1935, the census model was abandoned in favour of psychological experiments and political polls. In contrast to studies of social phenomena, market research grew to use surveys to discover people's moods and voting intentions, and to predict the results of upcoming elections (Gobo and Mauceri, 2014; Groves *et al.*, 2009).

A decisive turn came in 1936, when the comparison of data from two pre-election polls made it clear that small but representative samples provided more reliable estimates than large but partial or biased samples, such as samples that exclude some members of the population or include individuals with different characteristics from the excluded ones. Around this time, new sampling techniques were introduced, such as the probabilistic multistage area sampling technique, which allows the selection of a sample with known probability from a large and geographically widespread population. Sample survey research started to be adopted in universities, where scholars introduced longitudinal analysis (i.e. repeated interviews administered to the same respondents over long periods of time) and multivariate statistical analysis. Also, the number of topics covered by surveys increased and the questionnaire became increasingly sophisticated.

During World War II, the United States government systematically carried out surveys to monitor the public opinion on war and allegiances, hence enhancing the creation of the *National Opinion Research Centre (NORC),* which, together with the *Survey Research Center*

*(SRC)* (established after the War) became one of the most advanced social sciences research centres in the United States. During the post-war period, survey research was re-discovered and also widely applied in Europe. In Britain, surveys of morale were important and the governmental survey work was intense, focussing on such matters as consumer needs and shortages of consumer goods, food and nutrition, publicity and information activities (Ayrton, 2017). The first major example in Italy is the constitutional referendum in 1946, yet the first forecast of election outcomes came later with the pre-election political poll of 1948.

Since the 1960s, sample surveys have become the main tool for social science research, used by the academic community and largely employed in both the public and the private sectors (Marsden and Wright, 2010). Even though survey methodology was still not recognised as an autonomous discipline within the social sciences during the 20th century, it developed several features which later characterised survey methodology as a separate field of social science. The most telling sign of this evolution is the plurality of associations and scientific journals which combine the study of surveys in different fields in an academic environment. The most remarkable reference is the *American Association for Public Opinion Research* (*AAPOR*), operating in the United States since 1947. This is a professional association which, thanks to the journal *Public Opinion Quarterly*, combines the survey research conducted by market research companies and research centres specialised in political polls with survey research developed in the academic environment. Another example in Europe is constituted by the *European Survey Research Association* (*ESRA)*, founded in 2005 in order to coordinate survey research in Europe; its official journal, *Survey Research Methods*, was first published in 2007. In Italy, still not much attention is paid to survey methodology. The number of scientific journals devoted to the topic is small, and their impact on the European research environment is limited. As an example, no Italian scholar based in Italy has published papers on *Survey Research Methods* since 2014.

As survey techniques developed, new modes of data collection were adopted for interviewing nationally representative samples of the general population. Until the 1970s, research practices focused on face-to-face interviews, with the interviewer asking questions and recording the answers on a prepared answer sheet. This mode had the advantages of yielding high response rates (thanks to well-trained interviewers who persuade reluctant persons to take part in the survey), permitting the development of good interactions between interviewers and interviewees, making use of visual aids that facilitated the answer process

(Chang and Krosnick, 2009), and assisting respondents in giving relevant answers to the questions, often resulting in better data (Bethlehem and Biffignandi, 2012). Since the 1970s, the costs of face-to-face administration increased, due to the costs in time and money spent for the human resources involved in training and conducting face-to-face interviews, travelling from one respondent to the next. As a consequence, researchers were pushed into exploring alternative modes, such as telephone interviewing, self-administered paper-and-pencil mail questionnaires, audio computer-assisted self-interviewing (ACASI), telephone audio computer-assisted self-interviewing (T-ACASI), and Interactive Voice Response (IVR) surveys (Dillman, 1978; Dillman, 2000).

Since the 1970s, the diffusion of landline telephones increased: almost 90% of the North American and European populations had a telephone (Tucker and Lepkowski, 2008). This allowed telephone surveys to be a quicker and cheaper data collection mode than face-to-face interviews and also, in the case of Random Digit Dialling (RDD) samples, comparable in terms of data quality to probability-based face-to-face survey samples (e.g. Groves and Kahn, 1979). Today, the majority of public opinion surveys and election polls are still conducted exclusively by telephone, using RDD samples or voter registration lists (Dillman *et al.* 2014) or also landline phone registers. Nonetheless, telephone interviewing has recently posed new challenges that deal with the sampling frame and the response rate. The first challenge is the availability of incomplete sampling frames from which to select respondents. Landline directories may be affected by undercoverage because more and more people prefer their phone numbers to be unlisted, and many have now replaced their landlines with mobile phones (see Bethlehem and Biffignandi, 2012, Häder *et al.*, 2012, and Lavrakas *et al.*, 2017 for US; for Europe; Sala and Lillini, 2015 for the Italian case)[1]. These tendencies may lead to biased survey estimates due to coverage error, e.g., households with landlines have different characteristics to those with no phones or only mobile phones, both in Europe (Mohorko *et al.*, 2013) and in the United States (Blumberg and Luke, 2017). The second challenge is the problem of nonresponse. Declining response rates in both cross-sectional and panel surveys all over the world are well-documented by many scholars (Bethlehem *et al.*, 2011; Brick and Williams, 2013; Tourangeau and Plewes, 2013); for example, the Pew Research Center documented a dramatic decline in telephone survey response rates from 36% in 1997 to 9% in 2012 (Buskirk and Dutwin, 2016; Dutwin and Lavrakas, 2016). In addition, response rates are

[1] The directories of many countries do not list mobile phone numbers (Bethlehem and Biffignandi, 2012).

ever more difficult to maintain. The costs needed to achieve response rates as high as those achieved thirty years ago are greater (Holbrook *et al.*, 2007; Lavrakas, 1997). Lastly, nonresponse bias may occur, as respondents have characteristics that are different from those of nonrespondents (Lipps *et al.* 2015).

In the 1980s, technological innovations made computer-assisted forms of data collection possible (de Leeuw and Hox, 2015). With the advent of the Personal Computer, researchers realised the potential to decrease the human resources necessary to conduct surveys. This became possible thanks to the newly automated survey conduction process and the reduction of time devoted to data collection. By the end of the 1980s, the widespread use of landline phones was a key driver for the uptake of Computer-Assisted Telephone Interviewing (CATI). This mode of data collection, often used in many surveys with demoscopic aims (Natale, 2004) and also election polls, offers a number of advantages: - the process of data gathering is faster, cheaper and easier and data quality increases.

Since the 1990s, market and scientific research reduced the adoption of face-to-face interviews in sample surveys, with the notable exception of governments (Groves *et al.*, 2009) or official research institutes. Personal computers have been widely and increasingly adopted, and are now used in every research step, from data collection to data analysis. In the middle of the 1990s, the use of the Internet exploded when technological advances enabled the World Wide Web to become widely available. This technological innovation had implications for survey research, such as the adoption of new techniques of data collection, referred to as “web surveys”. These allow for quick, simple and cheap access to large groups of potential respondents. They include different types of self-administered, automated approaches to conducting surveys online (for an overview, see Couper, 2000) and are particularly appropriate when a complete list of contacts is available (e.g. e-mail addresses) and the target population is technologically savvy. Nonetheless, web surveys have three serious limitations: i) the differing spread of Internet coverage and access, ii) the inadequate sampling frame to survey the general population, and iii) the ethical issues posed by sending e-mail to people never contacted before.

The first challenge they have to face is Internet coverage: in 2017 the segment of the population that had access to Internet access ranged from 85.2% in Europe to 95.0% in North America (Internet World Stats, 2018). Even if penetration rates tend to become higher over time, the Internet population still differs from the general one, as Internet access is unevenly distributed over the population- highly educated and younger people more often have and use an Internet connection (e.g. Bethlehem and Biffignandi, 2012; Couper *et al.*, 2007; Rookey *et al.*, 2008; Sterrett *et al.*, 2017; Valliant and Dever, 2018). This differing coverage across countries and demographics may result in biased estimates on variables of interest in a specific study (Mohorko *et al.*, 2013). In addition, there is no adequate general population sampling frame (i.e. a comprehensive list of all or even most known members of the population) and no simple procedure is available for drawing samples in which individuals have a known, nonzero chance of being included, like the development of an algorithm for random selection from all the addresses. Moreover, ethical issues and cultural barriers are posed by contacting randomly generated e-mail addresses or sending e-mails to populations with which the surveyor has no pre-existing relationship (Dillman *et al.*, 2014; Smyth and Pearson, 2011).

To overcome coverage and nonresponse problems of online research, and still enjoy the advantages of web surveys (their cost-effectiveness, mainly), researchers implement mixed-mode designs with web surveys as one of data collection methods in the mix (de Leeuw and Berzelak, 2016; de Leeuw and Toepoel, 2018; Tourangeau, 2017). From a cost saving perspective, data collection usually starts with the most cost-effective method and then follows up with more expensive modes. The main disadvantage in using multiple methods within a survey is the potential occurrence of unwanted mode-measurement effects that researchers should try to estimate and adjust (De Leeuw, 2018). Several statistical procedures have been proposed, but are still under development (for an overview, see Hox *et al.*, 2017).

Another way to address coverage and sampling limitations is the implementation, since the mid-1990s, of Internet panels (Dillman *et al.*, 2014; Smyth and Pearson, 2011), and in particular of non-probability online panels (also called volunteer, opt-in or access panels). This thesis focuses on one of these last types of online panel, whose members sign up on a voluntary basis (Callegaro *et al.*, 2015). The main advantage of online panels is the accessibility of large databases of potential survey respondents: as some web panels contain tens and even hundreds of thousands of active panelists at a time, the challenges of finding and contacting web survey respondents are circumscribed (Dillman *et al.*, 2014). Another advantage of panels is the availability of key demographic information about each member (Hewson *et al.*, 2016). A potential drawback of online panels is that most of them use a variety of non-probability methods (e.g. self-selection, links and banners on web sites, or snowballing) to recruit their members (Couper, 2017; Vonk, van Ossenbruggen and Willems, 2008). This leads to two methodological flaws: undercoverage and self-selection. In terms of coverage, online-panel survey results do not apply to the general population, but only to the Internet population (Scherpenzeel and Bethlehem, 2011). Moreover, the volunteer mechanism by which people decide to join the panel results in self-selection (i.e. the selection probabilities of each member are unknown to the researcher) and maybe selection bias (Scherpenzeel and Bethlehem, 2011). Despite these problems, there still remains scores of panels across most countries with sufficiently large Internet populations, increasingly used by academics and other researchers (Callegaro, Baker, Bethlehem, Göritz, and Lavrakas, 2014; Schonlau and Couper, 2017). In recent years, the quality of access panels gave rise to concern (Faasse, 2005). Nonetheless, there is still relatively little knowledge on panel data quality and wide variability in results from non-probability online panels (see e.g. Craig *et al.*, 2013; Erens *et al.*, 2014; Vonk, van Ossenbruggen and Willems, 2008; Yeager *et al.*, 2011). Moreover, in Italy there are no studies focussing on opt-in panels, even if they are fairly widespread and used in mixed-mode surveys by several market research institutes. Previous research found some evidence for bias both in sample representativeness and in quality of data collected using non-probability online panels (Dutwin and Buskirk, 2017; Kennedy *et al.*, 2016; Yeager *et al.*, 2011), but findings are still inconclusive. Further studies are needed in order to assess the effectiveness of panel data in survey research.

This study sits in the challenging context described above. The overall aim is to address issues regarding the quality of data collected involving a sample from an Italian non-probability online panel and their effectiveness in representing the general population. In particular, I focused on i) undercoverage and self-selection bias in different samples (i.e., the Internet population, panel members, and web survey respondents), ii) nonresponse bias, comparing the characteristics of web survey respondents to those of nonrespondents, iii) data quality, comparing the estimates obtained from the web survey to those obtained from a probability-based reference survey.

I used three data sources. The first is data of all the members of the Italian non-probability online panel, recorded in the panel archive. This is a valuable source because companies usually do not provide panelists' data to their clients. As a consequence, researchers are not allowed to perform specific analyses on panel members as a whole. However, I was able to study i) the panel members' representativeness in comparison to both the Internet and the general population, ii) the web survey respondents' representativeness in comparison to the panel members, and iii) the differences between the characteristics of web survey respondents and those of nonrespondents. The second dataset comes from my primary data collection, conducted on a sample of Italian adults selected from the panel. The last source is the dataset from a probability-based reference survey, provided by ISTAT. The availability of microdata allowed me to apply a more specific weighting strategy to correct the estimates obtained from the web survey.

The impact of my study is relevant to both the Italian and the international survey research fields. In particular, this is the first study in Italy that i) uses panel survey data, and ii) compares the panel survey estimates to those from a probability-based reference survey, to assess the quality of data collected from a non-probability online panel. Results from the study contribute to i) filling the knowledge gap in the Italian context, and ii) providing empirical evidence from Italy to include in future comparative research.

The relevance of my study for international research on non-probability online panels consists in broadening the current findings on three under-researched topics. Firstly, to the best of my knowledge, there is only one publication that has assessed the differences in the demographic and socio-economic characteristics of panel members, the Internet population and/or the general population (Alvarez *et al.*, 2003). Secondly, there is only one European study on the assessment of the differences between opt-in panel-study respondents and nonrespondents (Pedersen and Nielsen, 2016). Lastly, there are only a few recent studies that use a gold standard to assess data quality in non-probability online panels (e.g. Dutwin and Buskirk, 2017; Erens *et al.*, 2014; Heen *et al.*, 2014; Yeager *et al.*, 2011).

This thesis has five chapters. Chapter 1 starts with a definition of an online panel and an overview of the spread of the opt-in panels across the world. Then, a systematic review of online panels as a method of data collection is conducted. I also reviewed a selection of published papers to detect the characteristics and usage of online panels in survey methodology. Chapter 2 offers a detailed description of the study design. Here, I formulate the research questions, describe primary and benchmark data used, and explain my methods of analysis. Chapter 3, 4, and 5 report the results of the study, following the main distinctions made in total survey error framework (Biemer, 2010). I pay attention to non-sampling errors and focus on three specific components of non-sampling error: coverage, nonresponse, and measurement. In particular, Chapter 3 addresses coverage and self-selection bias in different samples: Internet population, panel members and panel survey respondents. I also applied a statistical adjustment technique to minimise distortion. Chapter 4 provides an overview of the nonresponse process as occurs in the panel survey, and assesses nonresponse bias, comparing the socio-demographic characteristics of respondents to those of non-respondents. In Chapter 5 I address data quality and, in particular, I focus on measurement error, estimating bias on specific survey questions and trying to reduce it through weighting. The thesis concludes with some remarks and discussion about the use of non-probability online panels in social research.

# 1

# Online panels: what, when, and why

This chapter is divided into two main parts. I start with a definition of online panels and a brief history of their development (sections 1.1 and 1.2). Then, in the second part (sections 1.3-1.6), I present research design, methodology and results from a systematic review on usage of online panels in survey research. Section 1.7 concludes the chapter.

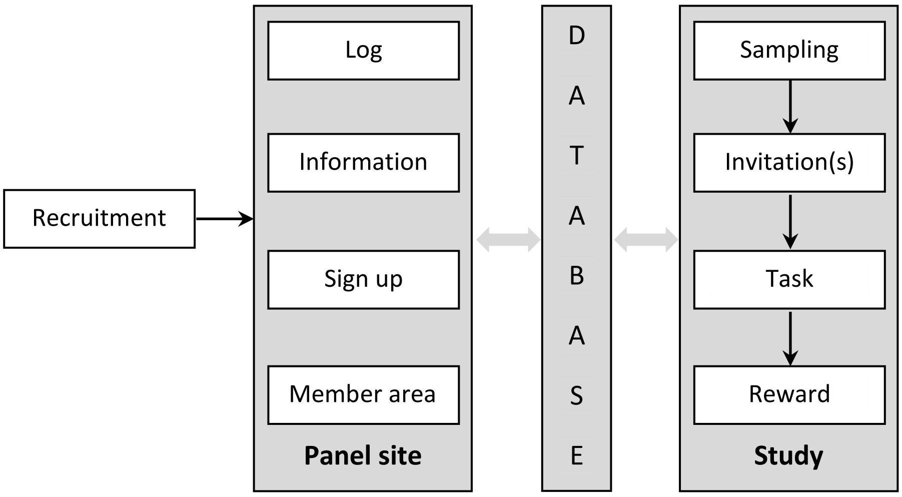
## 1.1 Definition of online panel

The International Organization for Standardization (ISO 26362) defines *online panel* as "a sample database of potential respondents who declare that they will cooperate with future (online) data collection if selected" (International Organization for Standardization, 2009). In practice, that means having a set of pre-recruited respondents who are ready to be contacted to participate in online research, most often by answering web surveys, although involvement in other types of research (e.g. qualitative data collection or measuring online behaviours passively from respondents, as in so-called Internet ratings panels) also is possible (Callegaro *et al.*, 2015).

In this chapter, I am interested exclusively in panels in which panel members are invited to answer web questionnaires on a regular basis, in which they voluntarily answer questions on various topics, often in return for some incentive (Couper, 2017; Göritz and Luthe, 2013b).

The implementation of an online panel goes through five stages: recruitment, joining procedures, profiling, sampling for specific studies, and panel maintenance (Baker *et al.*, 2010b; Callegaro, Baker, Bethlehem, Göritz, and Lavrakas, 2014), as shown in Figure 1 (Göritz, 2009).

Figure 1. Structure of an online panel.



The recruitment activity can include both offline (e.g. face-to-face or telephone interviews) and online (e.g. links or posts on different websites, banners, and e-mail messages) strategies, and can adopt probability methods (such as address-based sampling, random-digit dialing, and area probability sampling) or non-probability methods (such as selection from incomplete lists of contacts or self-selection of volunteers). At the joining stage, most reputable research companies follow a “double opt-in” process: after signing up, each potential member receives an e-mail and must take action indicating intent to join the panel (Baker *et al.*, 2010b). Upon confirming membership, some companies ask the new panelist to complete a profiling survey that collects a wide variety of background, demographic, psychographic, attitudinal, experiential, and behavioural information, that are refreshed regularly and also combined with answers provided to individual surveys (Baker *et al.*, 2010b; Callegaro, Baker, Bethlehem, Göritz, and Lavrakas, 2014). Profile data are useful at the sampling stage in selecting panelists for specific studies. Simple random sampling is rarely used because of the tendency to obtain samples that are skewed toward certain demographic characteristics (e.g. younger people are more likely to go online and respond to a web survey), whereas quota sampling is the most commonly used technique to maximize sample representativeness (Baker *et al.*, 2010b; Callegaro, Baker, Bethlehem, Göritz, and Lavrakas, 2014). When the sample is drawn, the selected members usually receive an e-mail invitation that as a minimum contains a hyperlink to the survey and a description of the incentive. A reward of some form (e.g. money, gift cards or points that can be redeemed for various goods or services) is offered to panelists in order to increase survey completion rates and combat panel attrition (Baker *et al.*, 2010b). Other activities focused on panel maintenance include keeping members active by inviting them to surveys and to do “cleaning” procedures on e-mail addresses, panelists who do not respond or provide bad data, and duplicate members (Baker *et al.*, 2010b).

## 1.2 Brief history and current spread of online panels

The idea of involving computer-assisted self-interviewees as respondents goes back to the late 1970s in England, France and Canada. The first pioneer probability-based telepanel was launched in the Netherlands in 1986, when a thousand households were selected to receive a PC and a modem and to complete questionnaires downloaded to the PC every weekend (Callegaro, Baker, Bethlehem, Göritz, and Lavrakas, 2014). The non-probability volunteer online panel concept has its origins in the earlier mail panels developed by a number of market research companies at least fifty years ago (Baker *et al.*, 2010a). Since the late 1990s, numerous opt-in web panels have been recruited to complete online questionnaires (Couper and Bosnjak, 2010; Postoaca, 2006), mainly in the United States, but with a few early adopters also in Europe, for example in the Netherlands (Comley, 2007; Faasse, 2005).

Nowadays, probability (general population-based) online panels are relatively rare because of the high investment in terms of costs and time to build and maintain them (Callegaro *et al.*, 2015), as well as the cultural and technological challenges posed by the digital divide. To the best of my knowledge, only twenty probability online panels are now active worldwide: eight of them involve populations from North America (i.e. US and Canada), eleven are based on populations from various European countries (i.e. the Netherlands, Germany, France, Denmark, Finland, and Iceland), and one involves Australian panelists. On the contrary, non-probability online panels are very popular around the world because they are easier and cheaper to build than probability ones, as they do not attempt to have an exhaustive sampling frame of the general population from which to recruit their members. Unfortunately, it is difficult to know how many non-probability online panels are currently in use, as a comprehensive list is not published. Nonetheless, many years ago certain scholars provided an estimate of their widespread use: in the Autumn of 1999, Göritz and Moser (2000) identified 64 online panels in a web search, and as of December 2004, Batinic and Moser (2005) estimated their number at 650-750 worldwide. I also tried to update these counts, checking the the Survey Police website (www.surveypolice.com), which reports more than 250 top-rated survey panels spread across the world. Looking at the panel coverage by members’ country, I calculated that the countries with the highest number of panels including a sample from their populations are as follows: United States (175 panels), United Kingdom (111 panels), Canada (100 panels), Australia (82 panels), and Germany (81 panels)[2]. The main geographic areas represented in the top-10 list of countries are North America, Europe, and Oceania. After this, I focused on non-probability online panels that also gather Italian members. Starting from the list of Italian panels published by Survey Police, I deepened my search online. I found 72 companies, only 11 of which are Italian, that own at least one access panel and I estimated a total of at least 87 active panels.

## 1.3 Introduction to systematic review

Increasingly low response levels when using offline methods and shrinking data-collection budgets - combined with lower costs and faster data collection with online surveys, more widespread Internet access, and opportunities in questionnaire design on the web - are the main reasons for the increasing use of online panels in the survey industry. Specifically, online panels have become a popular solution to the problems of coverage and recruitment. When there is no usable and complete list of e-mail addresses (or other contact information) as a sample frame for a particular target population, survey researchers select respondents using panelists (Baker *et al.*, 2010b), which are often considered the second-best option. This is in line with the observed shift from cross-sectional surveys to continuous survey-data collection, which has been solving problems with recruiting members. An initial investment to recruit panel members eliminates such expenses in subsequent surveys since the panel is already set up and on call (Couper, 2005).

In practice, we can use different types of online panels based on the kind of data they provide, membership composition, and panel-recruitment method (Callegaro *et al.*, 2015). I have already stated that this chapter focusses only on online panels that collect survey data, while I have no limitations regarding membership composition (e.g. general population, specialty groups, and proprietary panels) and recruitment methodology. Regarding the latter, non-probability online panels raise concerns about data quality. In addition to their limited (or non-existent) possibilities on statistical inference, data-quality issues related to the problem of "professional" respondents are often raised. Empirical studies in market research have found evidence that some panel members are completing large numbers of surveys and/or answering

[2] These numbers refer to panels that include people from the listed countries, but are not necessarily owned by companies from that country. Moreover, each panel can cover different nationalities, so it can be counted more than once, depending on the number of the members' nationalities.

questions in ways that maximize their chances of qualifying, while others are performing high levels of satisficing (Baker *et al.*, 2010b).

Nowadays, online panels are used often, especially in social and marketing research, as a sample source for more substantive research (e.g. Beierlein, Kuntz, and Davidov, 2016; Nedelec, 2018; Simons *et al.*, 2017). In addition, they are used in the survey-methodology field, either as 1) an object of research itself (e.g. Blom, Gathmann, and Krieger, 2015; Matthijsse, de Leeuw, and Hox, 2015; Smith *et al.*, 2016) or as 2) a sample source for various experimental studies on survey-data quality (e.g. Couper *et al.*, 2013; Schonlau, 2015; Struminskaya, Weyandt, and Bosnjak, 2015). These two approaches have two distinct objectives. In the first approach, the objective is to study the characteristics and quality of online panels themselves. In the second approach, the objectives are empirical studies that address various issues on data quality in survey methodology, in which online panelists are used as sample sources.

In this chapter, I focus on the usage of online panels in the survey-methodology field, looking at either the first or second approach, in which a systematic review of their usage, as well as reflections on their quality, have not yet been studied.

## 1.4 Research questions

I present a systematic review of the usage of online panels in the survey-methodology field, which aims to assess i) the characteristics of online panels used in survey methodology, ii) the quality issues of these online panels, iii) the characteristics of individual panel studies, and iv) the usage of online panels as a sample source for research in survey methodology.

After presenting the methodology of my research and describing the characteristics of online panels and of online panel studies, I address the research questions as follows:

RQ1: What are the characteristics of online panels used in survey methodology? I explore the different online panels used for the two approaches: for studies on online panels themselves and for studies of online panels as sample sources. Probability panels are more likely to be used to study panels themselves than non-probability panels.

RQ2: What dimensions of online-panel quality are addressed by survey methodologists? This question deals with survey methodologists' interest in the quality of online panels. More specifically, I get insights into the main concerns that survey methodologists have regarding the quality of online panels. In this case, I do not perform my own empirical assessment of online-panel quality, but rather review quality issues that have been raised by survey methodologists.

RQ3: What are the characteristics of individual panel studies, i.e. studies using samples from online panels?

RQ4: What are the research questions addressed by individual studies that use online panels as a sample source in the survey-methodology field? Here, the focus is on specific experimental research based on samples of panelists who complete web surveys. By this, I get insights into how valuable online panels are as sample sources for survey methodologists studying data quality.

## 1.5 Method

The research questions are answered using a systematic review of bibliographic references (papers, book chapters, etc.) dealing with 1) online panels as the object of methodological research and/or 2) survey data quality issues studied using online panels as sample sources. For this purpose, I systematically searched through and reviewed relevant bibliographic references, coded relevant data and built three separate datasets referring to three units of analysis: reference, online panel, and individual online panel study (a study using both online panels themselves and panels as sample sources).

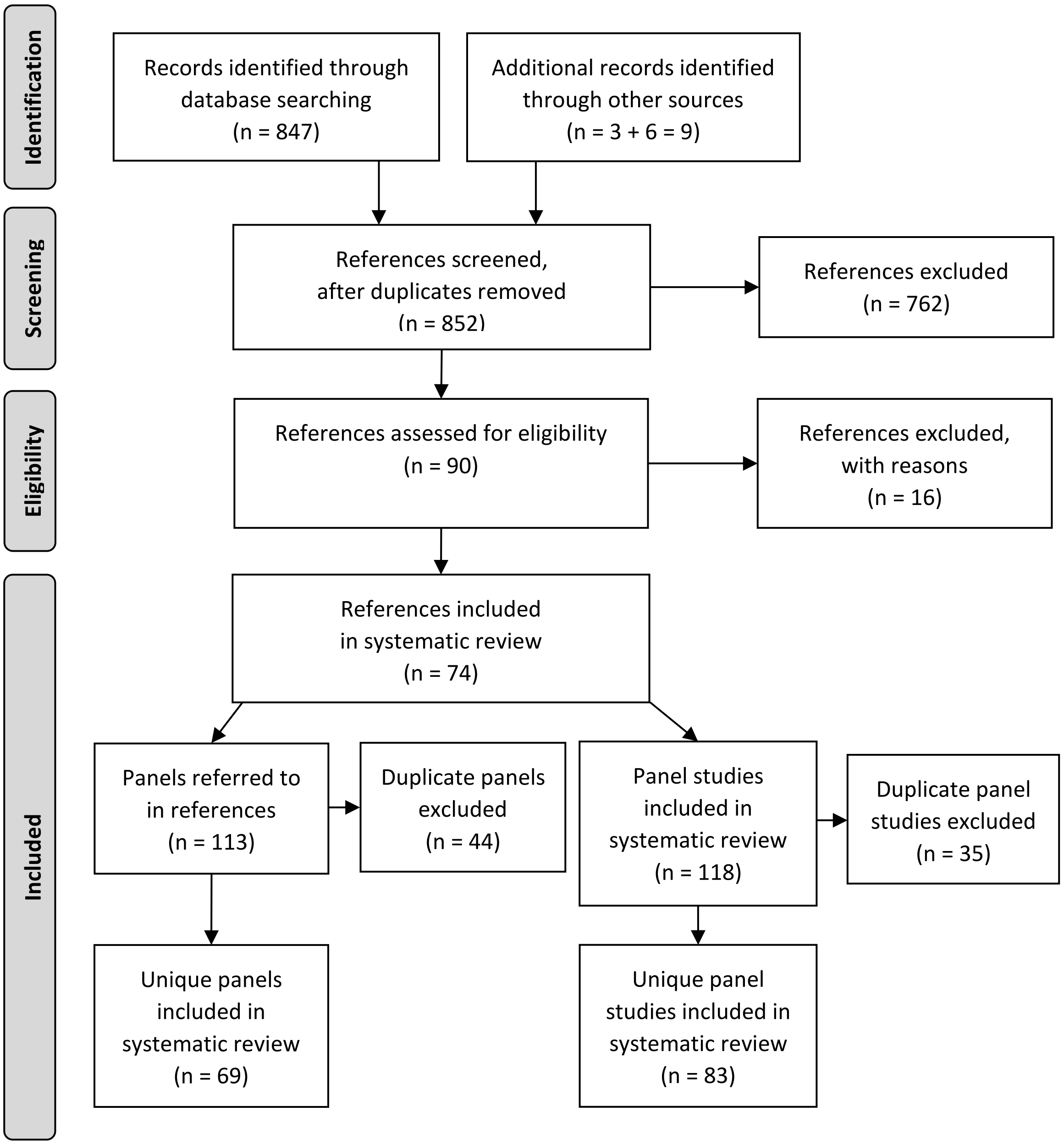
I drew on the three types of units of analysis to organize my analytical approach. As shown in Table 1, to describe the characteristics of online panels used for survey methodology research (RQ1) I planned an analysis at the level of the unique online panels extracted by my review. If a particular online panel was mentioned by several references, it is counted here only once. To answer my second research question, about the scholars' interest in the quality of online panels themselves, I looked at the level of references to state what dimensions of data quality are addressed by selected references. For the characteristics of individual online panel studies (RQ3), I focused on individual panel studies that used both online panels themselves and panels as sample sources. It is possible that an individual panel study used samples from different online panels, but when the same study design was used, I counted this as an individual panel study. Finally, an analysis at the level of individual panel studies was conducted to respond to my fourth research question (RQ4), concerning the usage of online panels as sample sources for research in survey methodology.

Here, I further illustrate the methodology of my systematic review, looking at the selection process of relevant references, the coding procedure I applied to the units of analysis, and the analytical approach I adopted.

### 1.5.1 Selection process

To describe the selection process, I follow a four-step structure suggested by Moher et al. (2009), presented with a PRISMA flow diagram in Figure 2.

Figure 2. PRISMA flow diagram describing the selection process of references.


Identification
Screening
Eligibility
Included
Records identified through database searching (n = 847)
Additional records identified through other sources (n = 3 + 6 = 9)
References screened, after duplicates removed (n = 852)
References excluded (n = 762)
References assessed for eligibility (n = 90)
References excluded, with reasons (n = 16)
References included in systematic review (n = 74)
Panels referred to in references (n = 113)
Duplicate panels excluded (n = 44)
Panel studies included in systematic review (n = 118)
Duplicate panel studies excluded (n = 35)
Unique panels included in systematic review (n = 69)
Unique panel studies included in systematic review (n = 83)

*Step 1 – Identification*. The aim of my search strategy was to find references that deal with the usage of online panels in the survey-methodology field. I used a bibliographic database by WebSM (http://www.websm.org), which is a relevant information source on web-survey methodology (Callegaro *et al.*, 2015; Hewson *et al.*, 2016, p. 151; Lozar Manfreda and Vehovar, 2007). It includes a vast and up-to-date database of bibliographic references related to web-survey methodology from published scientific papers, books, and book chapters through white papers, dissertations, theses, and unpublished conference presentations. Established and maintained by the Centre of Social Informatics' Faculty of Social Sciences at the University of Ljubljana since 1998, it has 4,000 to 8,000 visitors monthly. In June 2016, when the search was performed, it turned up 7,889 references related to web-survey methodology.

The preliminary identification of relevant references from this database followed three general inclusion criteria (not yet directly related to online panels): type of resource, language of the full text, and year of publication. I selected journal papers and book chapters only, published in English between January 2012 and June 2016. In addition, this included some journal articles published by the end of 2016.

In this way, I obtained a collection of the most recent and relevant references (relevant in the sense that they were published) in the web-survey methodology field, which included 847 references. In addition to this total, I included nine more references in the initial database. Regarding the year of publication, I made one exception and included six relevant chapters of a book published in 2011 (Das, Ester, and Kaczmirek, 2011) because it is one of the three volumes (the other two are Callegaro, Baker, Bethlehem, Göritz, and Krosnick, 2014, and Engel *et al.*, 2015) that focus explicitly on online panels. Three additional published papers not stored in the WebSM archive were identified at the end of the selection process, after sifting through the bibliography reported by the eligible references and surfing the Internet. To summarize, Step 1 yielded 856 results (references) on web-survey methodology, published in English in 2012-2016, including six chapters of the book published by Das and colleagues in 2011.

*Step 2 – Screening*. In this phase, I removed four duplicated records from the WebSM database selection, leaving 852 unique references. Then I applied a specific inclusion criterion: seven keywords to search for within the title text of the references obtained in Step 1, using the phrase "panel OR probability OR non-probability (both with and without dash) OR weight OR score OR representativeness." I used "panel," the most explicit word, to define our research interest. "Probability", "non-probability", and "nonprobability" refer to the typology of online panels. I also included three keywords: “weight”, “score,” and “representativeness,” referring to oft-mentioned aspects of data quality in online panel studies, dealing with the issue of representativeness of samples from online panels. A total of 90 references were retained, with 762 excluded, as they did not meet the keyword-inclusion criteria.

*Step 3 – Eligibility.* The 90 screened references were checked for eligibility through abstract and full-text appraisal. I first read each abstract in detail and tried to determine whether the study corresponded with our topic, i.e. the usage of online panels in the survey-methodology field. If it was clear that a resource was ineligible, I excluded it from the database. Otherwise, I moved to the second step and examined the full text to test eligibility further and decide which to include and exclude. The main reasons for ineligibility were web surveys conducted without panel members, usage of offline panels, and theoretical papers that examined online panels without presenting empirical evidence on their usage and quality. This led to the exclusion of 16 more references.

*Step 4 - Included references.* At the end of the selection process, 74 references (the first type of the units of analysis) met the criteria and were included in my systematic review (see Appendix 1). At this point, I further identified two other types of units for our analysis. First, the included references mentioned 113 online panels, but some references might have mentioned the same panels (e.g. two papers by Eckman, 2016, and Lugtig, 2014 both refer to the same panel). Overall, 69 unique online panels[3] (the second type of analysis units) were identified as reported by references in the survey-methodology field. Second, references were reported on 83 empirical studies using those online panels (further referred to as panel studies, the third type of analysis units). Here, I counted as unique panel-study cases in which a study was applied on one panel (e.g. a paper by Bosnjak *et al.*, 2013) and also when it was applied on different panels, but conducted using the same design and indicators of data quality (e.g. a paper by Binswanger, Schunk, and Toepoel, 2013).

### 1.5.2 Results from selection process: Overview of selected references

Two thirds of the 74 selected references are journal articles, while the other third consists of book chapters. From those I reviewed (from the beginning of 2012 to June 2016), the most prolific year for publications about online panels was 2014, when nine papers and two books

[3] In some references, the name of the panel is not reported, but the authors refer to it using generic expressions, such as “panel vendor,” “web-panel firm,” “commercial panel,” and “volunteer web panel.” Since I could not know exactly which panels they were, I counted them, as they were different.

on this topic were published. The journals that published such papers[4] the most frequently are *Social Science Computer Review* (13 references); *Public Opinion Quarterly* and *Field Methods* (four references each); and *International Journal of Public Opinion Research*, *International Journal of Market Research*, *Survey Research Methods*, and *Methods, Data, Analyses* (three references each).

As for geographical distribution of published studies in the selected references, the Netherlands led the pack (27 references), followed by the U.S. (18 references) and Germany (16 references). Most references refer to only one panel (62 references) and one panel study (68 references), but some examples focused on different panels and presented different panel studies: Twelve references report on data from 2 to 19 panels, and six references report on data from 2 to 4 panel studies.

The target population of the online panels reported is mainly the general or the adult (people ages 18 and older) population (in 48 references), meaning I focused on general-population panels. Other references reported on studies interested in specific subgroups from these general-population panels. More specifically, Internet users and visitors of popular websites are populations of interest in eight references, while thirteen references targeted specific groups, such as smartphone owners, people from different walks of life, residents in a particular geographic area, grocery shoppers, or road users (e.g. car drivers and motorcyclists). In addition, specialty panels - which are built by recruiting specific types of people because of some specific demographic or other characteristics (Callegaro *et al.*, 2015) - are referred to in one reference. I also found two examples of proprietary panels (Callegaro *et al.*, 2015) in which members would participate in research for a particular company (e.g. a Switzerland market research company's clients and a Google application's [app] users).

### 1.5.3 Coding procedure

The coding procedure followed the needs of my four research questions. For each specific purpose, I identified a group of categories and created appropriate variables to measure them. In the first two columns in Table 1 in Appendix 2, I report on the coding variables, grouped across my aims and research questions.

For RQ1, on the characteristics of online panels, the coding variables refer to these characteristics. The unit of analysis is online panels.

[4] Here I refer to the 49 journal articles. The other 25 references are book chapters.

For RQ2, on the dimensions of quality in online panels, the coding variables refer to these dimensions and are listed from the ones referred to by the largest number of references, to the ones referred to by the smallest number of references. The unit of analysis here is references, while the variables are the quality dimensions referred to in the references.

For RQ3, on the characteristics of individual panel studies, the variables refer to these characteristics. The unit of analysis is individual panel studies.

For RQ4, on the purposes of studies using online panels as a sample source in the survey-methodology field, the variables refer to these purposes. The unit of analysis here is individual panel studies, while the variables are the purposes.

I provide definitions of variables used to answer these research questions in Appendix 2.

## 1.6 Results

The references selected for this systematic review report on 69 unique online panels, which are used for survey-methodology research, either to study the quality of online panels themselves or as a sample source for other methodological research. Some panels are cited more often and therefore used more frequently in the methodological research in this field. The most emblematic case is represented by the first established online panel (2007) that was freely available to researchers to perform their own research, i.e. the Dutch Longitudinal Internet Studies for the Social Sciences (LISS) panel, which is mentioned 21 times, both for studying the quality of this panel itself and as a data source for the analysis.

When looking at the level of individual panel studies, 118 of these studies refer to 83 unique individual panel studies. As I have already pointed out, the same study can be applied to many panels, and different studies can be conducted on the same panel.

In the following subsections, I describe the characteristics of online panels (section 1.6.1) and the dimensions of quality of online panels addressed (section 1.6.2). Then I focus on the characteristics of the empirical studies for survey methodology research, conducted involving panel members (section 1.6.3), and on the purposes of the usage of online panels as a sample source for research on survey methodology (section 1.6.4)[5].

[5] A second coder applied the coding scheme to a sample of references and validated my results.

### 1.6.1 Characteristics of online panels used in survey methodology

To define the type of online panel, I draw on Callegaro and colleagues's (2015) classifications of online panels regarding i) the membership composition of the panel, and ii) the recruitment strategy of panel members. The first classification distinguishes online panels that attempt to include every possible type of respondent (general-population panels) from panels built by companies or public institutions by recruiting specific types of people or particular population groups (specialty and proprietary panels). Specialty panels recruit people with specific characteristics, e.g. young or employed people, while proprietary panels involve members who participate in a survey for a particular company (Callegaro *et al.*, 2015). The online panels described in my review are almost solely (66 out of 69) general-population panels. The second classification considers recruitment methodology: Probability online panels recruit members using a probability sampling mechanism that provides each member of the target population with known and non-zero probability of selection for the panel, while non-probability online panels are those involving a self-selection process by the people who want to join the panel. The latter are also referred to as volunteer, opt-in, or access panels. The online panels in my sample are mainly non-probability (53 out of 69). If I cross-check data from the two classifications, I find that most of the general-population panels (51 out of 66) adopt a non-probability recruitment strategy, with the two proprietary panels based on non-probability sampling and the specialty panel based on probability sampling.

Table 1. Membership composition of online panels by recruitment methodology (N = 69 online panels).

| Membership composition | Recruitment methodology | |
|---|---|---|
| | probability | non-probability |
| general population | 15 | 51 |
| specialty | 1 | 0 |
| proprietary | 0 | 2 |
| Total | 16 | 53 |

Finally, another feature that characterizes a panel is its size, i.e. the number of members who joined the panel. The selected references report this information only for one third of the online panels to which they refer. Sizes vary, from 1,000 to 490,000 members. These panels are mainly small (from 300 to 3,200 members) and medium (from 3,201 to 10,000) size: half of the panels have a maximum of 10,000 members.

Furthermore, most of the online panels are commercial (57) and rarely academic (9) or research/non-commercial (3) panels. Their geographical coverage is mainly national (85.5%), while international panels make up only 14.5%.

### 1.6.2 Survey methodologists' interest in the quality of online panels

Almost all references (89.2%) address at least one issue about the data quality of the panel itself. This means that survey methodologists pay significant attention to the problem of the quality of online panels. Table 2 shows all the dimensions addressed by scholars. Those especially relevant are nonresponse issues, respondents' behaviour, a comparison of point estimates from online panel surveys with a gold standard or with other modes of data collection/study designs, weighting techniques, loyalty to the panel, measurement error, and recruitment strategies for setting up an online panel.

Table 2. Dimensions of quality of online panels addressed by survey methodologists.

| Dimensions | References | |
|---|---|---|
| | N | % (N=74) |
| Nonresponse issues | 25 | 33.8 |
| Respondents' behaviour (speeders', fraudulents' and professional respondents' behaviour, and panel conditioning) | 18 | 24.3 |
| Comparison with a gold standard | 17 | 23.0 |
| Weighting techniques | 15 | 20.3 |
| Loyalty to the panel | 15 | 20.3 |
| Measurement error | 15 | 20.3 |
| Comparison with other modes of data collection/study designs | 11 | 14.9 |
| Recruitment strategies | 8 | 10.8 |
| Maintenance of the panel | 4 | 5.4 |
| Questionnaire design | 1 | 1.4 |

Nonresponse error is assessed with 25 references. In three cases, the problem is only mentioned, while in all others it is measured through various indicators of the response process. Here, the scholars distinguish between indicators of nonresponse applied at the recruitment stage (recruitment and profile rates) and at specific study stages (starting/participation/completion rate, screening/eligibility rate, break-off rate, and cumulative response rate). Some of the indicators I mentioned are also used to study sample composition. Socio-demographics and other characteristics of respondents' experiences on panels or in surveys are included in bivariate analyses or in regression models to predict, for example, response propensity. These analyses can result in some indications on how to increase the willingness of people to join an online panel, to stay in it for a long time, and to answer survey questionnaires.

Thus, scholars gathered empirical evidence on the other two key topics: recruitment strategies for setting up online panels (eight references) and members' loyalty to the panel (15 references). Here, strategies to motivate people to become panel members were evaluated, with some indication of their usefulness given (either through results of an experiment or evaluated in some other way). The main tools addressed by survey methodologists are monetary incentives (prepaid vs promised) of varying amounts, types of reminders (i.e. via letter, SMS, or e-mail), and multi-mode contacts (i.e. face-to-face, phone, and mail). Moreover, researchers reported offering gifts (e.g. tablet PCs, 3G Internet, high-quality chocolate) as rewards for participating in panels or experiments, or for evaluating the content of advance letters. Furthermore, researchers asked respondents to help them find other respondents, applying the respondent-driven sampling method, and experimenting on/evaluating the effect of the sponsors' authority. Finally, one reference in my review examines the reasons for joining the online panel and provides a set of material items to explain the choice to become a panel member.

After making this decision, a panelist can exercise different options, such as i) staying active on the panel/continuing to answer web surveys for a long time, ii) remaining on the panel, but reducing participation, or iii) dropping out of the panel after some time. Loyalty to the panel, defined by these three scenarios, is addressed in different ways in the references reviewed. First, attrition or retention rate is reported, but only in six out of 74 references, and three references distinguish between response rates of long-stay and of newer panelists. Second, it is possible to categorize references into two categories based on which factors influence attrition by focussing on background profile (i.e. socio-demographics, homeownership, urbanicity, voting behaviour, and Internet access) and panel-specific factors (i.e. saliency of the topic, incentives, reminders, recruitment mode, number of surveys assigned and completed, and panel-member tenure). Finally, one reference analyzes the “frailty effect”, i.e. the tendency of respondent attrition to be contagious within the respondent’s family, and another reference studies the relationship between response quality (as measured by satisficing indicators) and attrition propensity.

Roughly 20.3% of references deal with the problem of measurement error. In my systematic review, I found various operational definitions of this source of error: satisficing behaviour (i.e. acquiescence, extreme responses, non-differentiation or straight-lining, neutral middle, “don’t know” answers, item nonresponse, junk responses to open-ended questions, discordant answers, low-probability screening questions, little mapping effort in Public Participation Geographic Information Systems surveys [PPGIS]), quality estimates obtained by adopting a multi-trait/multi-method (MTMM) matrix, bias between true values and estimates, social desirability bias, and strong axiom of revealed preference (SARP) violations. Some studies computed correlations or regressions between the indicators of measurement error, socio-demographic variables, and the respondent’s behaviour during his or her stay on the online panel (i.e. number of completed surveys, membership on one or more panels, and attrition).

The issues described so far (i.e. nonresponse error, recruitment strategies, loyalty to the panel, and measurement error) concern a sort of “internal” quality of online panel data. In the literature, I also found evidence of a sort of “external” quality based on the comparison between online panel survey data and data from a gold standard (17 references), or from surveys conducted using various other modes of data collection and study designs (11 references). The gold-standard data come from face-to-face national surveys (e.g. the European Social Survey (ESS) and the German General Social Survey (ALLBUS), and official national statistics (e.g. U.S. Census, Statistics Netherlands, and the U.K. Population Census). The variables used for the comparison are socio-demographic characteristics, attitudinal and behavioural variables, urbanicity, Internet access and use, and other residual aspects (i.e. gender identity, religious confession, and health status). Moreover, online panel data are compared with data collected using other survey modes and study designs, such as random digit-dialing phone surveys, CATI surveys, mail surveys, and face-to-face surveys. Different study designs are also implemented comparing online panel data with other web-survey data collections using various samples: i) members belonging to another online panel, ii) a self-selected sample, iii) a random household sample or iv) an on-site recruitment sample. A wide range of variables on respondents' characteristics (i.e. socio-demographics, behaviours, attitudes, urbanicity, Internet access, co-morbidities, and religious affiliation) and their experiences on the panel (i.e. length of stay on panel, mean number of panels joined, and average number of surveys completed per week) is used in comparisons between different survey modes and study designs. It is worth noting that scholars also include some indicators of data quality: data usability in public participation GIS surveys, response issues (i.e. response time and likelihood of completing the survey), and satisficing (i.e. item nonresponse, speeding, straightlining, inattentive behaviours, cheating, little mapping effort in public-participation GIS surveys). This is a clear attempt to assess the quality of online panels as methods to obtain data for social research.

In addition to the attempt to check "external" quality, I found 15 references that apply weighting techniques to try and improve that quality. In my review, panel data are frequently adjusted by post-stratification weights, design weights, and propensity scores, but a combination of these different weights also is adopted. The purpose of weighting (in six out of 15 references) mainly is to compensate at the same time for various sources of error, such as coverage, sampling, and nonresponse errors.

Except for maintenance of the panel and for questionnaire design, which are marginal issues, results from the remaining dimensions of quality show that many references address other specific respondents' behaviours (18 references), in addition to the loyalty to the panel I examined above, particularly dealing with speeding, panel conditioning, and fraudulent and professional respondents.

### 1.6.3 Characteristics of individual online panel studies

As an online panel is established, many individual surveys are designed involving its members. The way in which these surveys are conducted provides insights on how the online panels are used in survey methodology. To describe the individual studies reported in my review, I focus on these characteristics: sampling method, sample size, and questionnaire length.

When looking at the level of individual studies (the third type of analysis units), the sampling is carried out from a sample frame consisting of panel members that may be obtained by a probability or non-probability method. Here, I look at the sampling design for an individual survey, regardless of the type of sampling used for panel recruitment. In this respect, in most cases, probability sampling (32 out of 81 individual studies) is used. Non-probability sampling characterizes 21 studies. Usually, the type of non-probability sampling is not specified in the papers, but there are nine individual panel studies that report adopting quota-sampling methodologies. However, we should keep in mind that the list of panel members may be obtained in a non-probability way; thus, the final sample is not based on probability, even if probability selection was used for the particular study. I also found that in 30 data-collection designs, there's no selection of any respondents' samples from online panels, as all panelists are involved. They can be all members, all active members, all online members, all panelists who use smartphones or mobile Internet, or all respondents to some previous waves.

The size of the samples obtained with any mentioned sampling method ranges from 312 to 153,758 units. Samples that are more frequently used are quite small: 53% of all samples contain 300 to 3,200 individuals. Nonetheless, another 21.2% of studies involves a larger sample that includes 6,000 to 10,000 people.

Regarding the questionnaire length in individual online panel studies, I need to consider two indicators: number of questions and time (in minutes) taken to fill out the questionnaire, as different references report on different indicators. Data on the first are available only for 15 studies out of 83. The number of questions ranges from two to 130, and the two modal values are 24 and 26. The thirty surveys for which I have the information on the completion time (either estimated time or actual time calculated) last from 1.3 to 30 minutes each. The most frequent duration range is 10-15 minutes.

In addition to these characteristics, it would be beneficial to know about the representativeness of the initial or final samples to the target population from these studies. However, only a small group (25 individual studies) actually addresses this issue, and from the information available, it is not possible to conclude how many of the samples can be considered sufficiently representative.

### 1.6.4 Usage of online panels as a sample source for research in survey methodology

My review shows that the usage of online panels as a sample source for research in survey methodology is less frequent than studying the quality of the online panel itself. Indeed, as I already have mentioned, this approach involves 37.8% (28 out of 74) of the references, compared with 89.2% of the references that report studies about panel data quality[6]. The references report 34 (out of 83) unique panel studies addressing the usage of online panels as a sample source for research in survey methodology. Thus, this part of the analysis includes 41% of the unique studies.

Table 3 indicates the issues addressed by unique studies using online panels as a sample source for research in survey methodology.

Table 3. Issues addressed using online panels as a sample source for research in survey methodology.

| Issues | Unique studies | |
|---|---|---|
| | N | % (N=83) |
| Measurement error | 27 | 32.5 |
| Response process | 23 | 27.7 |
| Questionnaire design | 13 | 15.7 |

When focussing on unique, individual panel studies, they are mostly used as a sample source to study indicators of measurement error (27 studies). Scholars use online panels to investigate a wide range of such indicators. The first indicator is satisficing behaviours, both in closed questions (measured by primacy and recency effects, midpoint selection, straight-lining, mental coin-flipping, number of answers in check-all-that-apply, and non-substantive responses) and open-ended questions (counting the number of completed open questions, number of characters/words of an open-ended question, and frequency of avoiding the open-ended "Other" option). The second indicator deals with time needed in survey-data collection, in which different definitions/measures of time are used. Specifically, these include the latency period between survey invitation and first survey access, time taken to answer specific questions, response time on the respective survey page, and duration of the questionnaire. Other indicators of measurement error used in the panel studies I considered are mode effect, conspicuous response behaviour in grid statements, and socially undesirable responses.

Studies using online panels as a sample source and dealing with the response process (27.7%) can be divided into two macro-categories: studies on indicators of response/nonresponse and

[6] Recall that the sum does not add up to 100% because 27% (20 out of 74) of the references report unique studies addressing both purposes: panel-data quality and use of panels as sample sources for survey methodology research.

study measures to increase response rates. The issues included in the first category are survey-outcome rates, number of contacts to obtain panel or survey participation, response rate for specific questions, initial panel-recruitment refusals, respondent reluctance, number of days for the respondent to complete the survey, response speed, and imputation as a way to estimate nonresponse bias. The second group gathers experimental designs that test some measures to increase response rates. Much empirical evidence focuses on the use of various types of incentives, such as lotteries (offering cash prizes, vouchers, and/or surprise gifts), monetary donations, and text appeals (altruistic vs. ego-oriented). Scholars also paid attention to invitations to fill in questionnaires and compared different e-mail texts (with or without the attribute "survey topic") and different modes of contact (text message vs e-mail). Finally, I found an example in which a combination of lotteries and offering study results is used to boost survey participation.

Questionnaire design features are addressed by a small number (13) of studies using online panels as a sample source, but covering a wide range of experiments. Comparing question-layout choices and interactive and visual features are the most researched topics. Other studies evaluated items per screen design, questionnaire versions/layouts, number and order of answer options, question-order effect, and different question wordings. Some authors tested offering closed-ended vs open-ended questions, while others asked respondents for their opinions about different questionnaire features (i.e. orientation, color, design, and usability).

## 1.7 Conclusions

In this systematic review, I dealt with the objective of using online panels in the survey-methodology field. In line with the shift from cross-sectional surveys to continuous survey-data collection (Couper, 2005), which in the long term spares resources for survey-data collection, online panels are often considered the future of survey research. The fact that for some potential populations of interest almost all members are online (e.g. college students and business executives) makes online panels a useful method of data collection. The main reason for their prominent use is that online panels have become a popular solution to the sampling frame problem, when a complete list of e-mail addresses for the target population is not available or usable (Baker *et al.*, 2010a). Having a group of members who are willing to cooperate in online research is a useful sample that does not need to be built for each survey. This condition saves the researcher's time and money for respondents' selection.

I applied my analyses at the descriptive level, particularly addressing four questions: What are the characteristics of online panels? What dimensions of the quality of online panels are addressed? What are the characteristics of individual panel studies? What are the research questions addressed by individual studies that use online panels in survey-methodology research?

I found that in the period from 2011 to 2016, 72 out of 74 references reported experimental studies conducted on online panels. Compared with the total number of references published on web-survey methodology in the same period (852), this result shows that issues regarding online panels are still under-researched. Given the large number of online panels out there (over 250, according to Survey Police[7]), I can assume that although online panels are not very widespread in the survey-methodology field, they definitely are more often used for substantive research, either in social sciences, health studies, or market research.

The references I selected focused on two purposes of using online panels: quality of the panel itself and use of the panel as a sample source in the survey-methodology field. The main interest of these works is the data quality of the online panel itself (66 references, equal to 62.2%), with 20 of them (27% of the references) also reporting on the usage of online panels as a sample source for survey methodology research. Only eight references (10.8%) reported only on the usage of online panels as a sample source without considering the quality of the online panel itself.

I may conclude that the quality of online panels themselves is an important issue for survey methodologists, but these panels could be used more often for methodological research, especially exploratory studies. The main limitations could be professional respondents and differences from other respondents involved in other more-frequent sample design (i.e. cross-sectional) in survey-data collections.

[7] https://www.surveypolice.com.

# 2

# Study design

## 2.1 Introduction

The popularity of non-probability online panels in market and social research and, at the same time, the shortage of methodological studies on the quality of data collected from these panels are the main drivers of my research project. As I stated in the introduction of this thesis, while the few studies conducted abroad showed a wide variability of (contrasting) results on the quality of data from panel surveys, the Italian context is still in need of research aimed at evaluating the data quality of opt-in panel surveys.

My study sits on this framework and addresses the shortcomings of the field. The overall aim of this work is to assess the quality of non-probability online-panel survey data in Italy. In particular, I explore the impact of undercoverage and nonresponse on sample selectivity, and the impact of nonresponse and measurement error on the quality of data from an Italian opt-in panel. Moreover, I assess the impact of weighting in i) reducing the composition bias and ii) improving the quality of survey data collected using non-probability online panels.

To achieve my aims I used three different datasets: 1) a unique dataset on the panelists of the non-probability online panel *Opinione.net*, 2) a dataset on the responding sample from a web survey conducted on the *Opinione.net* members, and 3) a dataset on a sample of the Italian population from a probability-based reference survey, with no coverage error and high response rates, conducted every year by ISTAT and often used as the gold standard.

To address my research aims, I adopted a variety of methods. First, I compared the panel survey estimates to those obtained with the reference survey, performing bivariate analyses and running logistic regression models. In addition, I computed specific accuracy metrics to assess sample representativeness and data quality. Last, I calculated different indicators of nonresponse and measurement error. When appropriate, I also applied a particular type of statistical adjustment strategy, i.e. quasirandomization weighting, to improve the survey estimates. In my comparative work I used inferential statistics. Although, strictly speaking, these statistics can be calculated only with probability samples, their use with non-probability samples (in particular, quota samples) is generally accepted practice in the survey industry (Duffy *et al.*, 2005).

There are different designs comparing estimates from online panels of non-probability samples to other data sources. They differ with respect to whether the comparison is made against probability-based or non-probability survey samples, the mode of data collection of the comparison study, and the availability of benchmark estimates (Callegaro, Villar, Yeager, and Krosnick, 2014). Drawing on the designs found in literature by these authors (Callegaro, Villar, Yeager, and Krosnick, 2014), I adopted a particular application of "Design 4" (see Table 2.1, p. 25). Whereas this design uses a face-to-face survey with a probability sample as comparison study, my design uses this type of survey as benchmark. In my case, the non-probability survey is a cross-sectional study conducted on a sample of panelists and it does not focus on capturing time attrition and its reasons.

In this chapter I go through my research questions (section 2.2), the data I used (section 2.3), and the methods of the analysis adopted to answer each research question (section 2.4).

## 2.2 Research questions

This study addresses four research questions that are drawn from the reading of the literature on the quality of opt-in panel studies in survey research. In Chapters 3, 4, and 5 I critically review the papers on sample representativeness, nonresponse, and measurement error in non-probability panel survey data. In this section, I provide a very brief overview of the works discussed later with the view of framing the context in which the research questions are set.

Most of the panels used in survey research recruit their members adopting non-probability strategies, raising concerns about the generalizability of the findings. Research has shown that sometimes these panels are used to run experiments; at other times they are used to study specific populations (Schonlau and Couper, 2017). The report of the AAPOR task force on non-probability sampling pointed out that the use of opt-in panels should be avoided when the estimate of the general population values is a key research objective (Baker *et al.*, 2013). Following this caveat, in many studies on opt-in panel survey data the generalizability of survey estimates is not a priority goal. For example, these works focus on questionnaire design, mode of data collection, and strategies to boost survey participation. On the other hand, the AAPOR task force stated that there are times when a non-probability online panel is an appropriate choice. Within this context, a minority of studies explicitly or implicitly addresses the objective of the inference to the general population, using opt-in panel survey samples. However, these studies are primarily concerned about the impact of specific survey features on response and quality of data from panel survey samples, with a view to better design future surveys. In general, these works can be grouped in three research streams, i.e., sample composition, nonresponse to a specific panel survey, and measurement error.

Studies belonging to the first stream have shown that undercoverage and nonresponse occurring at different stages of the implementation of an opt-in panel (i.e. recruitment, joining procedures, profiling, sampling for specific studies, and panel maintenance) may result in sample composition bias (Dutwin and Buskirk, 2017; Eckman, 2016; Erens *et al.*, 2014; Sterrett *et al.*, 2017; Tsuboi *et al.*, 2015). A consistent finding is that panel members and respondents to a given panel survey are usually not representative of the population they claim to mirror (i.e. the Internet population or the general population).

My first research question aims to evaluate the impact of undercoverage and nonresponse on sample composition at different stages of the development (in other words, “the life of the panel”) of the Italian non-probability online panel *Opinione.net*. As discussed in Chapter 1, opt-in panels have the specificity to go through sequential stages of development: i) recruitment, ii) joining procedures, iii) profiling, iv) sampling for specific studies, and v) panel maintenance (Baker *et al.*, 2010b; Callegaro, Baker, Bethlehem, Göritz, and Lavrakas, 2014). At each stage, a sample selection mechanism may be activated and result in different types of non-sampling errors. In particular, I focus on undercoverage occurring at the recruitment stage, and on nonresponse occurring during the whole life of the panel. Given the cross-sectional (and not longitudinal) nature of my survey, I excluded the last stage (panel maintenance).

Focussing on sample composition is a key task, especially when researchers are interested in using the survey estimates obtained with non-probability online panels to draw conclusions on the inference population (e.g. the general population). Samples that are not representative of the population they claim to mirror may produce biased estimates and, thus, undermine the inference goal.

As online panels are built involving people who have Internet access and not everyone has Internet access, undercoverage occurs. Thus, the first step towards the inference is assessing the composition of the Italian population with an Internet connection (recruitment stage). In particular, I assess whether the Italian Internet population can be considered representative of the general population.

Moreover, people should be exposed to the stimulus "become a panelist" and decide to join the panel, but we know that only a portion of the Internet population is exposed to the stimulus and then agrees to subscribe. Thus, the second step (joining and profiling stages) is to assess the representativeness of the Internet population that self-selected into the panel. In particular, I assess whether the *Opinione.net* panelists are representative of both the Internet and the general population.

Lastly, only a subsample of panelists is invited to and takes part in a specific survey. This results in nonresponse. Thus, the final step (sampling for specific study stage) is to assess the representativeness of respondents in comparison with the samples selected during the previous two stages (i.e. the general population, the Internet population, the panelists, and the members invited to the specific survey). In particular, I assess whether the responding (study) sample is representative of the general population, the Internet population, the *Opinione.net* panel and the selected sample.

The second research stream gathers papers that focus on survey participation to a specific study, and describe the nonresponse process, documenting the quality of this process. Three main findings stand out: i) the types of response/nonresponse are not reported and are implicitly drawn from the response metrics, e.g., participation rate, and break-off rate; ii) despite the availability of guidelines, there is little agreement on the terminology used to define these metrics (Baker *et al.*, 2010a; Callegaro and DiSogra, 2008); iii) a minority of studies do not accurately document the quality of the response process (Craig *et al.*, 2013). A number of studies look at the respondents' characteristics (Keusch, 2013; Knapton and Myers, 2005). When comparing the characteristics of panel survey respondents to those of other respondents' samples, contrasting findings stand out. For example, age sometimes does influence participation rate (Cho *et al.*, 2015), whereas at other times it does not (Brown *et al.*, 2012). Moreover, these works identify ways to efficiently maximize the response rate to surveys. Implementing experimental designs, they detect the most effective strategies in improving survey response (Göritz and Luthe, 2013c; Mavletova, 2013). Thus, the main aim of these studies is to offer insights on response process to better design web surveys.

However, all these studies fail to assess the occurrence and magnitude of nonresponse bias, and to address item nonresponse (i.e. a phenomenon that occurs when respondents skip specific survey questions) in a comprehensive approach, at the level of both question and respondent.

Assessing the impact of undercoverage and nonresponse on selection bias is key. However, assessing the characteristics of response and the magnitude of nonresponse at the specific study stage is of paramount importance, because those who participate in a given survey are, again, a self-selected sample of the invited sample. This mechanism might be a potential threat to the accuracy of the survey estimates.

My second research question is to assess the nature, occurrence, and magnitude of nonresponse. In particular, I pursue three aims. The first is to disentangle the types of response and the characteristics of the nonresponse process (e.g. how many people are reached, how many respond, how many refuse to cooperate, etc.), focussing in particular on the discussion of the measures used to describe this process. The second aim is to assess the occurrence and magnitude of unit nonresponse bias, i.e., the systematic error that occurs when the expected nonresponse rate is high and there are differences between respondents' and nonrespondents' estimates. The third aim is to assess the occurrence and magnitude of item nonresponse at the level of the responding sample. This phenomenon occurs when respondents skip specific questions.

The third research stream includes studies on measurement error in non-probability online panel surveys. Scholars investigated a wide variety of causes of measurement error that come from three main sources, i.e. questionnaire design, mode of administration, and the respondent (see Chapter 5 for definitions). In recent years, after the publication of the AAPOR "Report on Online Panels" (Baker *et al.*, 2010a) in 2010, only a few studies focussing on mode effect were carried out (Goldenbeld and de Craen, 2013; Pennay *et al.*, 2018). More specifically, the literature documented few attempts to disentangle a special type of mode effect, that is survey-mode effect within a unique opt-in panel survey, when addressing the measurement error issue (Mavletova, 2013; Mavletova and Couper, 2013). On the other hand, except for the special case of pre-election polls (e.g. Breton *et al.*, 2017; Malhotra and Krosnick, 2007), to the best of my knowledge there is no attempt (in recent years) to disentangle mode effect between non-probability panel surveys and probability-based surveys or official statistics. Moreover, inconclusive findings on data quality from non-probability panel-survey samples stand out.

Measurement error may occur when, filling out a questionnaire at the study specific stage, respondents not only skip questions, but also provide inaccurate answers. This source of error is defined as the difference between an observed response and the underlying true response (Baker *et al.*, 2010a). Therefore, a further research question aims to investigate the occurrence, nature, and magnitude of measurement error in the online-panel survey estimates, and in particular to explore the impact of mode effect on measurement bias.

When the sampling frame is not probability-based, as in online-panel surveys, selection (into the panel) and response (to a specific study) probabilities are difficult to tie to the target and the inference population and, thus, survey estimates based on the analysis of opt-in panels may be biased. For this reason, adjustment techniques are usually applied, either at the sample selection stage or after data collection, to make an online panel more closely mirror the population (Baker *et al.*, 2010b). After data collection, different approaches can be adopted to adjust the survey estimates to external population benchmarks. There is debate on the statistical adjustment procedures to adopt with non-probability samples (Tourangeau *et al.*, 2013; Baker *et al.*, 2010b). In particular, when dealing with non-probability online panels, current practices include the use of calibration weighting, and propensity score adjustments (Elliott and Valliant, 2017; Baker *et al.*, 2013).

My last research question aims to assess the effectiveness of weighting in removing bias from my survey estimates. More specifically, I computed a particular type of propensity score adjustment, i.e. quasirandomization weighting (see the Methods of analysis section for details), and I investigated whether quasirandomization weighting is effective in reducing i) the composition bias of the responding sample, when compared to the Internet and/or general population it claims to represent, and ii) the occurrence and magnitude of measurement bias in the survey estimates.

## 2.3 Data

To pursue my aims, I used a diverse set of data, i.e. the 2015 Multipurpose Survey - Aspects of Everyday Living, considered as the gold standard in my analyses, and two datasets from the non-probability online panel *Opinione.net*, i.e. data on all registered panelists and data from the Italians' Living Conditions (ILC) survey, a study that was conducted on a quota sample of the *Opinione.net* panel members.

Table 1 provides an overview of the data used in this work.

Table 1. Overview of the data.

| Data source | Reference time | Data collection mode | Type of sample | Final sample |
|---|---|---|---|---|
| AEL | 2015 | Face-to-face | Probability | 45,204 |
| *Opinione.net* panelists | 2017 | web | Non-probability | 8,113 |
| ILC (subsample of *Opinione.net*)<br>- incentive: 0.40 euro<br>- reminder: one e-mail reminder<br>- questionnaire: Internet use and life styles<br>- cooperation rate: 52.7% | 2017 | web | Non-probability | 2,007<br>(initial sample: 3,908) |

### 2.3.1 "Gold standard": Multipurpose Survey - Aspects of Everyday Living (AEL)

The Multipurpose Survey - Aspects of Everyday Living (AEL) is a probability face-to-face repeated cross-sectional survey (using paper-and-pencil interview as mode of data collection) of Italian households run every year by ISTAT since 1993 (in 2004, the survey was not carried out). It collects a wide range of information, including data on household composition, education, training, employment, health, media and IT consumption, political participation, leisure time, Internet access and Internet use. It adopts a two-stage clustered sampling method, where primary and secondary sampling units are, respectively, municipalities and households. The sampling frame is address-based, being drawn from regularly updated administrative registers (registration is not on a voluntary basis) that list all the households and individuals who live in a given municipality. The size of the sample from the most recent survey, conducted in 2015, is large and counts 45,204 individuals living in 19,158 households and 836 municipalities. Because the AEL survey is based on a probability sample (with no coverage error in the sampling frame) and is characterised by a high response rate, this survey is usually considered a “gold standard” in validation studies (see, for example, Sala and Lillini, 2015). As a check of robustness (because of the lag in the time frame of the survey data used in my analysis), I compared the distributions by age, sex, and area of residence of the 2015 AEL survey with those of the 2017 Italian administrative data[8]. This comparison showed no statistically significant differences in these three variables.

### 2.3.2 Data on the *Opinione.net* panelists

The panel, used to sample the web survey respondents, was built by the research institute Demetra opinioni.net s.r.l. in 2011. Demetra is one of the six Italian companies (associated with the Italian Association for Market Research - ASSIRM, the national equivalent of the European Society for Opinion and Marketing Research - ESOMAR) that own a non-probability online panel recruited from the general population. *Opinione.net* (https://opinione.net/) is a partner of the well-known international online panel *Consumer Insights Network* (CINT) and meets the industry-standard 28 ESOMAR questions (https://www.cint.com/, for more info). *Opinione.net* is a commercial volunteer panel of 8,782 members (as of 2nd February 2018) and adopts a multi-mode approach to recruit its members, using online modes, such as banners on different portals and websites, as well as offline modes, via landline and mobile phone interviews. People who agree to become panelists are invited to the registration page on the panel website, where they have to enter four pieces of information: e-mail address (and password), gender, year of birth, and postal code of the municipality where they live. According to the “double opt-in enrolment” (Postoaca, 2006) adopted by the *Opinione.net* panel, the recruitment is completed when an e-mail with panel participation confirmation is received by the new member. Following e-mail address validation, each member can update his/her account adding other their own profile data about socio-demographic characteristics, household components, health conditions, hobbies and interests, journeys, food consumption, smoking habits, use of Internet/media/PC/videogames, ownership of electronic devices and cars/motorcycles[9]. Upon completion of each

[8] To perform this analysis I used administrative data from the ISTAT database GeoDemo, available at http://demo.istat.it/. The database was consulted on 21 February 2018.

[9] Such information is updated when they modify their personal profile or complete a survey.

questionnaire, respondents receive a monetary incentive whose value depends on the length of the questionnaire[10].

In my study I used socio-demographic data (i.e. profile data) of all *Opinione.net* registered members as of May 2017 (the same date in which the web survey was conducted) consisting of 8,113 cases.

### 2.3.3 The Italians' Living Conditions (ILC) survey

The Italians' Living Conditions (ILC) web survey was conducted between 16th to 24th May 2017 on a sample of members of the *Opinione.net* panel. The initial sample of 3,908[11] members was selected applying quotas that were defined to be proportional to the gender within age, and geographic area of residence distributions of the Italian population[12].

The questionnaire was constituted of 14 questions taken from the questionnaire of the Multipurpose Survey and, drawing on the "unified-mode design" (Dillman *et al.*, 2014) approach, which recommends adopting the same question wording and formatting features, designed to ensure that the stimulus is kept as similar as possible across the modes (recall that the AEL survey is administered by an interviewer). The AEL questionnaire is constituted of more than 250 questions, too many to be administered in a web survey. For this reason I selected a limited number of questions on the basis of the following three criteria: i) question wording and answer options are most fit for purpose of analysing measurement error; ii) their contents are preferably related to technological aspects, skills and habits, topics often included in other similar questionnaires; iii) the number of missing answers to the question is limited and a degree of variability in the answers provided is observed[13]. To the core questions from the AEL questionnaire, I added one more socio-demographic question, which has the same wording as one included into the panelists' profile dataset. See Table 2 for the list of the topics and the questionnaire structure and Appendix 6 for the questionnaire.

[10] When the respondent accumulates 5 euros, he/she can spend them to buy something on the Amazon website, to buy top-up cards for mobile phones, or to make a donation in favour of people with Hansen's disease.

[11] The value is an estimate (calculated by the survey software) of the number of invitations that would need to obtain 2,000 complete questionnaires.

[12] Administrative data (updated to January 2017) from ISTAT were used as benchmark.

[13] Nonetheless, meeting these criteria leads to a limitation. Indeed, the ILC questionnaire may suffer from question context effect, coming from not asked questions (i.e. questions excluded from the questionnaire) and from the different order in displaying questions (Lee *et al.*, 2017; Nielsen and Kjær, 2011).

When the data collection started, the panel counted 8,113 members. I provided the questionnaire to Demetra and requested a final sample of about 2,000[14] adult (aged eighteen and above) respondents, but I gave no instructions on how to conduct the survey (e.g. sampling method, text message of the invitation e-mail, and number of reminders) and I required no weighting adjustment[15]. The participants received an e-mail[16] invitation to fill in the web questionnaire, clicking on the link included in the text message (see Appendix 7 for the text). Nonrespondents received an automatic e-mail reminder two days after the invitation. The size of the final sample was 2,007 respondents who completed the six-minute web questionnaire, in exchange of a 0.40 euro incentive. The cooperation rate (AAPOR Cooperation Rate 1) was 52.7%.

In addition to the ILC survey dataset, Demetra provided me with the paradata, i.e. data registered by the instrument used to administer the survey, which give information on the process. For example, the ILC survey paradata include the final disposition codes, i.e. the codes used to define the final contact status for each unit of the initial sample.

Table 2. The structure of the questionnaire adopted by the ILC survey.

| Source | Topic | Number of questions |
|---|---|---|
| AEL survey | Internet use | 3 |
| | Food/drink/tobacco consumption | 4 |
| | Socio-political participation | 1 |
| | Watching TV | 2 |
| | Environmental problems | 1 |
| | Occupation | 1 |
| | Household income | 1 |
| | Marital status | 1 |
| Panel - profile data | Education | 1 |

[14] I asked to collect 2,000 questionnaires, but at the end of the data collection I obtained 2,007 questionnaires, because of the method of operating used by the survey software.

[15] The clients sometimes ask Demetra to provide weighted data and to suggest the variables for weighting, that usually are sex, age, education, and behaviours (measured through the official statistics or collected from previous surveys).

[16] The research institute used the software LimeSurvey v2.50 to manage the whole process of data collection, i.e. to send invitations and reminders, to administer the web questionnaire, and to automatically save survey answers into a database.

## 2.4 Methods of analysis

To study the sample selection into the panel study and the resulting data quality of survey estimates, I mainly focus on non-sampling bias, i.e. the systematic distortion of the survey estimates, and the most challenging component of non-sampling error.

To reach my research aims I adopted different techniques, i.e., bivariate[17] and multivariate (i.e. logistic regression models) analyses to address research questions 1.1, 1.2, 2.2, 3.1, and 3.2 (see Table 3), using socio-demographic variables (i.e. sex, age, marital status, education, occupation, and geographic area of residence). I computed specific accuracy metrics to assess sample representativeness and data quality. When appropriate, I also applied a particular type of statistical adjustment strategy, i.e. quasirandomization weighting, to improve the survey estimates. All the analyses were carried out using SPSS (version 25).

I performed my analyses on different populations, including Italian adults (i.e. people aged 18 and over). When analysing the ILC respondents' population, I only used data from complete (i.e. submitted) questionnaires.

An overview of my study design, together with the specific research questions, is shown in Table 3.

Table 3. Overview of the study design.

| Research questions | Populations | Methods |
|---|---|---|
| 1.1 Impact of undercoverage on sample composition | General population, and Internet population | Bivariate analysis |
| 1.2 Impact of nonresponse on sample composition<br>- at the recruitment stage | General population, Internet population, and ILC respondents | Bivariate analysis |

[17] I performed the Chi Square test to check for the statistical significance of each relationship. Although, strictly speaking, the significant testing can be used only with probability samples, its use with non-probability samples is generally accepted practice in the survey industry (Duffy *et al.*, 2005).

Table 3. *Continued*.

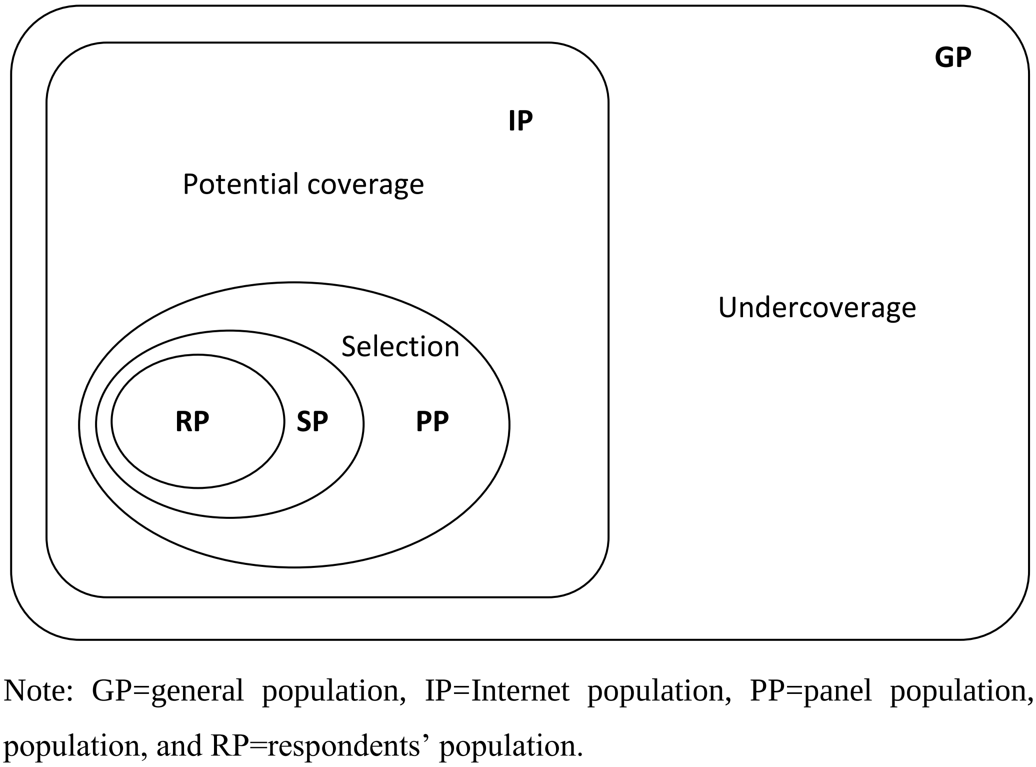
| | | |
|---|---|---|
| - at the joining and profiling stage | General population, Internet population, and panelists | Bivariate analysis |
| - at the specific study stage | Panel members, panel members invited, and ILC respondents | Bivariate analysis |
| 2.1 Types of response and nonresponse process | ILC respondents, and ILC nonrespondents | Response metrics |
| 2.2 Occurrence and magnitude of unit nonresponse bias | ILC respondents, and ILC nonrespondents | Bivariate and multivariate analyses |
| 2.3 Occurrence and magnitude of item nonresponse | ILC respondents | Indicators |
| 3.1 Occurrence and magnitude of measurement bias | General population, and ILC respondents | Bivariate analysis |
| 3.2 Impact of mode effect on measurement bias | General population, and ILC respondents | Multivariate analysis |
| 4.1 Effectiveness of weighting in removing | | |
| - sample composition bias | General population, Internet population, and ILC respondents | Bivariate analysis |
| - measurement bias | General population, and ILC respondents | Bivariate analysis |

Before presenting the methods for each research question, I make a brief digression in order to define the different populations included in my study (see column "Populations" in Table 3).

In Figure 1, partially drawing on Valliant and Dever (Valliant and Dever, 2018, Figure 6.1, p. 111), I propose an illustration of these populations and the sampling process for the ILC survey.

My population of inference is the general population (GP), which is the population I ultimately intend to draw conclusions about. As I used an online panel as a sample source for my study, my target population is the Internet population (IP), that is the general population excluding people who do not have Internet access and, as a consequence, will be excluded from the sampling frame. The Internet population is the potentially covered population on which the recruitment of the panel population (PP) is based. Only a non-probability sample of potential panelists is reached and voluntarily consents to join the panel. Usually, when carrying out a specific study (e.g. the ILC survey), a sample of registered members is drawn from the frame population (i.e. the panel population) and is invited to participate in a web survey. The self-selected portion of this sampled population (SP) that decides to fill out the questionnaire is the respondents' population (RP) and is the final sample on which I collected survey data.

Figure 1. Types of populations and sampling for a specific study on opt-in panelists.



Note: GP=general population, IP=Internet population, PP=panel population, SP=sampled population, and RP=respondents' population.

In my study, I compared all these five populations (i.e. general, Internet, panel, sampled, and respondents') against each other to pursue my aims. In the following sections, I look separately at the methods adopted to answer my research questions.

### 2.4.1 The impact of undercoverage and nonresponse on sample composition

My first research question aims to address the representativeness of *Opinione.net* as a result of the selection process, i) starting from Internet coverage, ii) going through self-selection into the panel, and iii) ending with response to the ILC survey. To assess sample composition at these three stages, I compared different populations, according to the patterns summed up in Table 3 (see above).

After performing a bivariate analysis, I used the results to calculate specific measures to estimate the bias. In particular, I drew on the accuracy metrics proposed by Yeager et al. (2011) and I used the "percentage point error" and the "largest absolute error" as point measures of the systematic error, and the "average absolute error" and the "number of significant differences from the benchmark" as overall measures. The value of all these metrics ranges (potentially) from 0 to 100, except for the number of significant differences from the benchmark that ranges from 0 to 6; the larger their values, the bigger the bias. The "percentage point error" is the percentage difference between the modal category of the benchmark and the survey estimate for that category. The "largest absolute error" is the error (measured as the absolute value of the percentage point error) of the variable on which the survey was least accurate. The "average absolute error" is the average (for the six variables used for the comparison) of the absolute values of the percentage point errors between the modal category of the benchmark and the survey estimate for that category. The "number of significant differences from the benchmark" is the number of variables considered in the survey that are statistically significantly different from the benchmark.

In addition to these data quality metrics, I introduced an innovative measure, i.e. the number of absolute differences greater than three given thresholds, that, for my purpose are set to 5, 10 and 15 percentage points. Moreover, when focussing on the Internet population, I proposed a new conceptualisation of this population that combines Internet access and use. Both my innovations are described in detail in Chapter 3.

### 2.4.2 The nature, occurrence, and magnitude of nonresponse in the ILC survey

In order to answer my second research question, I focused on respondents' and nonrespondents' populations from the ILC survey.

To disentangle the types of response, I used paradata that include the final disposition codes for each panelist who were invited to participate in the ILC survey. To assess the occurrence and magnitude of item nonresponse, I included in the analysis socio-demographic, behavioural and attitudinal variables for the ILC respondent population.

Regarding the techniques, I used the indicators of nonresponse proposed by Toepoel (2015), who classified three types of nonresponse error: unit nonresponse, partial nonresponse, and item nonresponse. "Unit nonresponse" describes the number of individuals within the sample who do not participate in the survey because they are not reached or refuse to answer the questionnaire. AAPOR (2016b), using well-established indicators, developed and made available several formulae for the calculation of response, cooperation, refusal, and contact rates. If we are interested in the final participation in the survey, the mode of data collection is web, and all the units have a known eligibility (as it is in the ILC survey), AAPOR recommends the adoption of cooperation rate. Thus, I calculated this rate. The second type of error is "partial nonresponse", which occurs when the respondent interrupts the completion of the questionnaire and leaves it incomplete (drop-out)[18]. The last type of error is "item nonresponse" that occurs when respondents fail to provide the answer to an individual question within the survey.

To assess the characteristics of the nonresponse process, I computed three response metrics (i.e., cooperation rate, break-off rate, and refusal rate) on my survey data. When focussing on the occurrence and magnitude of nonresponse bias, in addition to bivariate and multivariate analyses, I computed six nonresponse measures (i.e., the percentage point differences, the percentage absolute relative nonresponse bias, the nonresponse bias, the average differences between respondents and nonrespondents, the mean percentage absolute relative bias, and the average nonresponse bias).

Moreover, I looked at the respondents' sample only and assessed the occurrence and magnitude of item nonresponse. Drawing on Callegaro and colleagues' (2015) work I calculated two indicators: i) the "item nonresponse rate", defined as the number of units with item nonresponse divided by all eligible units exposed to the item, and ii) the "unit-level item nonresponse", defined as the number of all item nonresponse for a certain unit divided by the

[18] If the respondent submitted the questionnaire, I classified him/her as respondents, whereas the respondent did not submit the questionnaire I defined him/her as break-off.

number of all items to which a certain unit was exposed. For both indicators I also computed the mean value. In Chapter 4 I give details on these measures.

### 2.4.3 The nature, occurrence, and magnitude of measurement error in the ILC survey

The populations involved in this part of the analysis are the ILC respondents' population and the general population.

The variables I used are socio-demographic characteristics, and behavioural and attitudinal variables. I carried out a bivariate analysis, also performing the Chi Square tests, and I used the results from this analysis to calculate the accuracy metrics, as I did for the analysis on the impact of undercoverage and nonresponse on sample composition.

In addition I considered some indicators of measurement error, suggested by Mueller et al. (2014), Revilla and Ochoa (2015), and Lugtig and Toepoel (2016), selecting those that could be applied to my survey variables. Table 4 links every type of question to the suitable indicators and their methods of calculation.

Table 4. Indicators related to the measurement error.

| Question type | Indicator | Method of calculation |
|---|---|---|
| Set of items | Straightlining | Frequency with which the respondent agrees or disagrees with the items on a scale or chooses an "extreme" or "medium" category |
| Multiple choice question | Socially desirable response | Tendency to answer questions in a manner that will be viewed favourably (regarding current social norms and standards) by others |
| Check-all-that-apply question | Number of provided answers | Counting the total number of answers for each respondent |

Last, I performed a multivariate analysis to investigate the impact of a specific cause of measurement error, i.e. the survey mode (web vs face-to-face), on measurement bias in the estimates from behavioural variables. When interpreting the results from the logistic regression models, I paid specific attention to the estimates obtained from two questions (i.e. alcohol and tobacco consumption) that could be affected by social desirability bias.

### 2.4.4 Quasirandomization weighting to remove bias in the ILC survey estimates

Given the availability of the AEL microdata (i.e. my benchmark survey), I adopted Valliant and Dever’s (2018) approach, and computed a particular type of propensity score adjustment, i.e. quasirandomization weighting. These weights are calculated as the inverse of the propensity scores predicted by the logistic regression model that estimates the probabilities of inclusion in the web survey sample using sex, age, education, and geographic area of residence as independent variables.

I applied quasirandomization weighting before performing:

- the bivariate analysis in the comparisons “ILC survey sample vs Internet population” and “ILC survey sample vs general population”, when addressing the sample representativeness (first research question);
- the bivariate analysis in the comparison between the ILC survey sample and the general population (using both behavioural and attitudinal variables, and indicators of measurement error), when assessing data quality (third research question).

# 3

# Undercoverage and self-selection in non-probability online panels

The overall aim of this chapter is to investigate the impact of two sources of non-sampling error, i.e. undercoverage and nonresponse, on the representativeness of the Italian non-probability online panel *Opinione.net*. I focus, in particular, on nonresponse occurring at the following stages of the life of the panel: i) the recruitment stage, ii) the joining and profiling stage, and iii) the specific study stage. I analyse a unique set of data that includes information on all registered panelists. Being the first study that focuses systematically on selectivity introduced at different stages of the life of the panel, I believe my research makes an important contribution towards the improvement of data quality in non-probability online panels. This work contributes to enhance the current knowledge on this topic in a number of ways, particularly as it is the first study that focuses on a Southern European country and critically discusses the definition of "Internet population", providing a new conceptualisation and implementation of this concept.

## 3.1 Literature review and research aims

Sample representativeness is one of the main problems for non-probability online panels. Despite the relevance of this topic, we know very little about the complicated selection process that is behind it. When surveying the general population using non-probability online panels, sample selection is the outcome of two combined sources of non-sampling error, i.e. coverage of the sampling frame and, in particular, undercoverage, and nonresponse (Bethlehem, 2010; Duffy *et al.*, 2005; Legleye *et al.*, 2015). Similar to other Internet surveys, undercoverage may occur because of the limited Internet coverage and access that may exclude (from the sampling frame) some individuals who do belong to the inference population, i.e. the general population. Nonresponse may occur because some sample members decide not to take part in the survey. It can arise at different stages of the life of a panel: i) the recruitment stage, ii) the joining and profiling stage, iii) the panel maintenance and attrition stage and, iv) the specific study stage (Baker *et al.*, 2010a). Self-selection into the panel, arising at one of the first three stages of the life of the panel, is a source of bias that is specific to non-probability online surveys. Indeed, the sampling frame used in these online surveys is constituted of volunteers who were contacted or may have independently discovered the panel and agreed to become (and remain) part of the panel. Undercoverage and nonresponse may be sources of bias, depending on the scale of the undercoverage or response rate and the differences between individuals included/excluded from the sampling frame or respondents and non-respondents on the variable(s) of interest. Focussing on the Italian case, the overall aim of this work is to investigate the impact of undercoverage and nonresponse on the representativeness of the Italian non-probability online panel *Opinione.net*.

In reviewing the relevant literature, I first focus on research on Internet coverage and access and then on studies on nonresponse. I state the research questions addressed in this work at the end of each section.

### 3.1.1 Internet coverage and access

Research on the quality of the sampling frame (defined here in terms of Internet penetration and the quality of the Internet connection) and its impact on the different survey outcomes is surprisingly scant. It is true that the high European and North American coverage rates are unlikely to introduce a major source of bias - according to the latest Internet World Stats[19], in 2017, the percentages of the European and American populations with Internet access vary between 85.2% and 95%. However, the variability in the quality of the Internet connection between and sometimes within countries may be an issue for survey research, as poor Internet connections may *temporarily* exclude from the sampling frame eligible sample members (and possibly be a source of nonresponse bias and measurement error). This may be the case of Europe and Italy in particular where the so-called “next generation access” (NGA) coverage is more spread in the South and urban areas (EU 2018, p. 125). I am not aware of any papers evaluating the impact of the quality of Internet connection on survey outcomes (e.g. response, measurement error).

[19] Internet World Stat is an international website that features up to date world Internet Usage, Population Statistics, Social Media Stats and Internet Market Research Data, for over 243 individual countries and world regions.

Although in the US there is a wealth of research highlighting the potential coverage bias of web-only surveys and documenting disparities in Internet access and use (for a recent review, see Sterrett *et al*., 2017), the situation is quite different in Europe. The handbooks of online research methods often omit the provision of a comprehensive and in depth overview of Internet access (and possibly Internet bias) in Europe. Exceptions include works by Bethlehem and Biffignandi (2012), Bethlehem (2015) and Valliant and Dever (2018) (who dedicated a section or at least a couple of pages, to Internet access), though these exceptions usual report quite dated data. The only paper providing a comprehensive discussion on this topic is by Mohorko et al. (2013). Comparing data from the 2005-2009 European Eurobarometer survey, the authors found an increase in Internet access at home across Europe (although there was variation within the different countries regarding the rate of the increase) and documented the changes in the socio-demographic characteristics of the individuals with and without Internet access. Although differences in the composition of these two groups were shrinking over time, such differences persisted in 2009, with the Internet population being still likely to be male, young, and highly educated.

Despite the shortage of comparative research in this field, there are some recent works that looked at Internet access and bias in specific European countries e.g. the Netherlands (Eckman, 2016; Toepoel and Hendriks, 2016) and Germany (Blom *et al.*, 2017; Bosnjak *et al.*, 2013). Findings from these studies demonstrate that there is great variation even within countries in continental Europe when assessing the impact of Internet access on bias. In the Netherlands, for example, because of the very high Internet coverage the socio-demographic differences between individuals with and without Internet access have low impact on bias, whereas in Germany such differences seem to lead to samples that are not representative of the whole German population and may bias survey estimates. There are, to date, no studies focussing specifically on the Italian case.

Against this background, the first aim of the work is to investigate the impact of Internet undercoverage on sample representativeness and, in particular, to assess whether the Italian Internet population can be considered representative of the general population. I believe findings from this part of the study contribute to filling in an important gap because there are no studies that focus on South European countries and in Italy in particular. In addition, as I shall discuss in the Methods section, I will draw on Valliant and Dever's (2018) work and adopt an innovative way to measure the Internet population, proposing an indicator that is a combination of Internet accessibility and individuals' technology skills.

### 3.1.2 Nonresponse

Similarly, there is little research on the impact of nonresponse on sample composition in non-probability online panels, despite the wealth of data that research institutes hold on their panelists (Baker *et al.*, 2010a). I anticipate that studies in this field have mainly documented the effects of nonresponse at the recruitment stage on sample composition; there is currently a shortage of research on the impact of nonresponse occurring at the other stages of the life of the panel.

As clearly stated in the AAPOR report, "there is no way of knowing anything precise about the size or nature of the nonresponse that occurs at the recruitment stage" (Baker *et al.*, 2010a, p. 728). To gain some insights into the effects of nonresponse, scholars usually focus on the survey specific responding samples and compare the demographic and socio-economic characteristics of these samples to those of the general (or the Internet) population. Research in this field, mainly carried out in the US, has extensively documented the characteristics of these differences. Dutwin and Buskirk (2017) have recently demonstrated that the unweighted samples from non-probability Internet panels showed substantially higher estimated bias than the low response rate probability samples. They also documented that such bias persisted even after applying different types of statistical adjustment techniques (e.g. propensity weighting). Research from the Pew Research Center (Kennedy *et al.*, 2016) found that responding samples from non-probability online panels under-represent adults with less formal education while simultaneously over-representing non-hispanic whites and adults aged 65 and older. When comparing the characteristics of the responding sample to those of the general population, Cho and colleagues (2015) found differences in age, education, and household income of American panel members. For example, the responding sample over-represents middle age people, those with higher than average educational level, and individuals with higher household incomes. Similarly, Heen *et al.* (2014) found significant differences in age, race, education, household income, political affiliation, marital status, type of accommodation when comparing the responding samples of three online panels and data from the general population. For example, two panels consistently over-represented white people and non-hispanic individuals. When comparing seven non-probability online panels, Craig *et al.* (2013) found that all responding samples under-represented low socio-economic respondents, i.e. adults who did not graduate from high school or had annual incomes less than US $15,000. Despite the consistent evidence on the composition bias of US non-probability online panels, Sell *et al*. (2015) found that the Google responding samples sometimes more closely approximated national averages for ethnicity and race than the reference face to face survey. Research has also documented that non-probability online panels are characterised by strong composition bias both in Japan (Tsuboi *et al.*, 2015) and Europe. In the UK, there are marked differences in the socio-economic and demographic composition of the responding samples as well as high variability within the panels (Erens *et al.*, 2014), whereas in France respondents to online non-probability panels are often middle aged, active individuals and often living in households with only two members (Legleye *et al.*, 2015).

As mentioned previously, there are very few publications on the impact of the joining and profiling stage and the study specific stage nonresponse on sample composition. These studies adopt a similar methodological approach to that used to study the impact of nonresponse at the recruitment stage, and compare the differences in the demographic and socio-economic characteristics of panel members, the Internet population or the general population. Findings from these works are consistent with those that emerged from research on the impact of the joining and profiling stage on sample composition. Alvarez et al. (2003) found very strong socio-economic and demographic differences when comparing US panelists recruited using two different modes (i.e. banners and subscription) with the Internet population. For example, panelists recruited using the “subscription mode” are more likely to be women, white, aged 18-29 than the Internet population. Similarly, when comparing the characteristics of the Danish panelists to those of the general population, Pedersen and Nielsen (2016) found that the panelists are more likely to be women, live in the Capital Region of Denmark, and be younger than 60. To the best of my knowledge, there are no studies on the impact of the specific study stage nonresponse on sample composition.

Drawing on Baker et al. (2010a), the second aim of this work is to explore the impact of nonresponse occurring at the recruitment stage, the joining and profiling stage, and the study stage of the life of a non-probability panel on sample representativeness. In particular, I aim to test whether i) the responding (study) sample is representative of the online and general population, ii) the panelists are representative of the online and/or the general population and iii) the responding (study) sample is representative of the selected sample and the panel. I will focus on the case of the Italian panel *Opinione.net* and will compare the demographic and socio-economic characteristics of "respondents" to those of the reference population.

This work pursues a further aim. To contrast the composition bias of the responding sample, researchers often adopt different adjustment techniques. However, there are contrasting findings on the effectiveness of weighting in reducing or removing this type of bias (Duffy *et al.*, 2005; Dutwin and Buskirk, 2017; Lee *et al.*, 2015). The third aim of this work is to provide further evidence on the impact of weighting on the sample composition bias of the responding sample. Being the first study that provides a systematic assessment of the impact of nonresponse occurring at (nearly) all stages of the life of a non-probability online panel on sample selection, this work is a key contribution to a better understanding of the still well-sealed "black box" of non-probability online panels.

## 3.2 Research design

### 3.2.1 Data

To pursue my aims, I used a diverse set of data: the 2015 Multipurpose Survey - Aspects of Everyday Living (considered the gold standard in my analysis), and two datasets from the non-probability online panel *Opinione.net* (data on all registered panelists) and also data from the Italians' Living Conditions (ILC) survey, a study that was conducted with a quota sample of the *Opinione.net* panel members (see Chapter 2 for details).

Table 1 provides an overview of the data used in this work. I performed the analyses on the adult population (i.e. people aged 18 and over) who completed the whole questionnaire.

Table 1. Overview of the data

| Survey | Reference time | Data collection mode | Type of sample | Response rate | N |
|---|---|---|---|---|---:|
| AEL | 2015 | Face-to-face | Probability | Not available | 37,825 |
| *Opinione.net* panelists | May 2017 | Internet | Non-probability | Not applicable | 8,071 |

Table 1. *Continued*.

| | | | | | |
|---|---|---|---|---|---|
| ILC (subsample of *Opinione.net*) | May 2017 | Internet | Non-probability | 52.7 | 2,007 |

Note: The sample size and the response rate refer to respondents who are 18 and over. The AEL survey documentation does not provide the response rate for the adult population only.

### 3.2.2 Methods

To reach my aims I did the following: i) developed a new way to conceptualise and operationalise the concept of Internet population and ii) computed and compared a number of data quality metrics, based on the results of the bivariate analysis. I focus on the following six socio-demographic variables, available in all three datasets: sex, age, marital status, education, occupation, and geographic area of residence.

*Conceptualisation and measurement of the Internet population*

There are different ways to conceptualise and measure the concept of Internet population (for an overview, see Tourangeau et al. 2013, ch. 2). In many studies, the Internet population is defined as a fraction of the general population that has access to the Internet and is often measured using indicators such as "Do you have Internet at home"? However, this form of operationalisation has limitations for survey research, because it could lead to inaccurate estimates of the coverage rate and, as consequence, of the size of the Internet population, e.g. excluding individuals who access the Internet from other locations than home (e.g. workplace) or including less technologically savvy household members who live in households with Internet access.

To overcome these limitations, I propose a new conceptualisation of the Internet population that partially draws on Valliant and Dever (2018). I propose the definition of the Internet population as 'a fraction of the general population who i) regularly accesses and uses the Internet from any location, regardless of the device used and ii) is able to use the Internet'. To calculate this indicator, I used three variables that measure the frequency of Internet access and the device used to access the Internet and a variable that collects information on the activities performed when using the Internet (see Appendix 8 for the question wording of the variables used to create the indicator). In this work, the Internet population consists of individuals who accessed and used (at least a few times a week) the Internet in the last twelve months and performed - in the last three months - at least one of the following activities: sending/receiving e-mails, making phone calls over the Internet, posting messages, using instant messaging, and participating in social networks. Missing data on the variables used to construct this indicator are very low and vary between 0.3% (Do you or anyone in your household have access to the Internet at home?) and 7% (Have you ever used the Internet?).

*Data quality metrics*

There are different metrics that are commonly used to assess the quality of non-probability online panels (for an overview, see Callegaro, Villar, Yeager, and Krosnick, 2014); in this work, I mainly drew on the accuracy metrics used by Yeager et al. (2011). In particular, as overall measures of systematic error, I used the i) average absolute error and ii) the number of significant differences from the benchmark, whereas as point measures I used iii) the percentage point error and iv) the largest absolute error. I also propose using an innovative metric - v) the number of absolute differences greater than a given threshold, to provide additional indication on the magnitude of the bias. For my purpose the thresholds are set to 5, 10, and 15 percentage points. The value of all these metrics ranges (potentially) from 0 to 100, except for the number of significant differences from the benchmark and the number of absolute differences greater than a given threshold that ranges from 0 to 6; the larger their values, the greater the bias. The average absolute error is the average (for the six variables used for the comparison) of the absolute values of the percentage point errors between the modal category of the benchmark ($y_i$) and the survey estimate ($\hat{y}_i$) for that category, as expressed by the formula:

(1) $AAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$

The percentage point error is the percentage difference between the modal categories of the variable distributions from two surveys. Each statistically significant difference from the benchmark contributes to the “number of significant differences from benchmarks” (I computed Chi Square tests to establish the level of significance of the differences). The largest absolute error is the error (measured as the absolute value of the percentage point error) of the variable on which the survey was least accurate.

*Weighting*

There is debate on the weighting procedure to implement in the case of surveys that are not based on probability samples (Tourangeau *et al.*, 2013; Baker *et al.*, 2010b). In the case of non-probability online panels, current practices include the use of calibration weighting and propensity score adjustments (Baker *et al.*, 2013; Elliott and Valliant, 2017). Given the availability of the AEL microdata (i.e. my benchmark survey), I adopted Valliant and Dever's (2018) approach, computing a particular type of propensity score adjustment, i.e. quasirandomization weighting. These weights are computed as the inverse of the propensity scores predicted by the logistic regression model that estimates the probabilities of inclusion in the ILC survey sample using sex, age, education, and geographic area of residence as independent variables.

As previously mentioned, I am interested in assessing whether the study responding sample is representative of the online and/or the general population. Therefore, to address the third research aim, I ran two separate regressions (on the two different populations) and calculated two sets of quasirandomization weights that I then applied when performing the bivariate analysis and comparing the study responding sample and i) the Internet population and ii) the general population.

## 3.3 Results

Results from my analyses are shown in Table 2. Panel 1 reports percentage differences; positive and negative values indicate that the corresponding categories are over or under-represented, respectively. Panel 1 also shows the level of significance of the Chi Square tests calculated when performing the bivariate analysis. Panel 2 shows the values of the other accuracy metrics. The variable distributions are shown in Appendix 9.

### 3.3.1 The impact of undercoverage on sample representativeness

In 2015, 78.7% of Italians had Internet access whereas the Internet population - individuals who regularly access and use the Internet and have the capacity to do so - is constituted of 48.1% of Italians. Results from my analysis show that the Internet population is not representative of the general population (col. 1). Indeed, the average absolute error and the largest absolute error are 11 and 20.3, respectively. The variable distributions of the two populations are all (statistically significant) different and the magnitude of the bias is large, being six of the differences greater than 10 percentage points (of which four are greater than 15). In particular, the Internet population over-represents single people, educated individuals, and those in employment whereas it under-represents people aged 65 and over, those with no education, and economically inactive individuals.

### 3.3.2 The impact of nonresponse on sample representativeness

In the following sections, I look separately at the impact of nonresponse occurring at the different stages of the life of a panel on sample representativeness and explore the role of weighting in removing the differences in sample composition.

*Nonresponse at the recruitment stage*

Following on from Baker at al. (2010a), I evaluated the impact of nonresponse at the recruitment stage comparing the characteristics of the ILC (responding) sample to those of the online and the general population, before and after weighting (col. 2-3 and 4-5),

Surprisingly, the ILC sample is not a representative sample of the Internet population. Indeed, the average absolute error is 3.2 and the largest absolute error is 7.1. The demographic and socio-economic composition of the two samples differ by age, education, occupation, area of residence, and marital status. However, there are no differences by sex. The magnitude of the bias is quite large, with three of the differences larger than 10 percentage points. The ILC sample over-represents older people, highly educated individuals and those panelists who are economically inactive whereas it under-represents individuals who are in employment.

Quasirandomization weights have a positive impact on the selectivity that nonresponse introduces at the recruitment stage of the life of the panel, overall *reducing* its magnitude; for example, the average absolute error decreases from 3.2 to 2. Interestingly, although weighting is effective in reducing the magnitude of the bias (e.g. weighting is effective in removing the differences that are even greater than 10 percentage points), it introduces an additional source of bias in the gender distributions of the two samples. In addition, the weighted distribution tends to over-represent single people and economically inactive respondents while under-representing divorced/widowed individuals and those in employment.

Similar to the case of the Internet population, the ILC sample is not representative of the general population either (col 4 and 5). The average absolute value is 7.9, the largest absolute error is 15.8 and, with the exception of the area of residence, all differences between the variable distributions of the two samples are statistically significant. The magnitude of the bias is large, being ten differences greater than 10 percentage points. Specifically, the ILC sample over-represents men, single people, educated individuals, and those who are in employment and it under-represents divorced/widowed individuals, people aged 65 and over, economically inactive sample members.

Also in this case, weighting proves to be very effective in *reducing* the bias, e.g. the average absolute error decreased from 7.9 to 1.7 whereas the largest absolute error decreased from 15.8 to 5.1. In addition, weighting is also effective in removing the differences greater than 10 percentage points. However, the weighted distribution still over-represents single people and under-represents divorced/widowed individuals and those aged 65 and over and has introduced some bias in the variable area of residence.

In brief, the ILC (responding) sample is not a representative sample of the Internet nor the general population, the latter being characterised by a larger selection bias. However, quasirandomization proves to be a very effective weighting method to reduce the magnitude of the bias, especially in the case of the general population. Weighting makes the ILC sample more representative of the general population than of the Internet population. Unfortunately, some bias remains even after weighting.

*Nonresponse at the joining and profiling stage*

To assess the effects of nonresponse occurring at the joining and profiling stage, I followed a similar approach and compared the characteristics of the panelists and those of the online and the general population (col. 6 and 7).

As documented in the analysis of the data quality metrics, the panel *Opinione.net* is not a representative sample of the Internet population, although the selection bias is not large. The average absolute error is 3.3, the largest absolute error is 7.3 and the differences between the panelists' demographic and socio-economic variables and those of the Internet population are all statistically significant. The magnitude of the selection bias is relatively small, i.e. six differences are greater than five percentage points and only two are greater than ten percentage points. Specifically, *Opinione.net* over-represents women, single people, high-educated individuals and economically inactive persons. The panel under-represents people with a primary school education.

The panelists are not representative of the general population either. Indeed, the average absolute error is 9.4 and the variable distributions are all statistically significant from those of the AEL survey. The magnitude of the selection bias is large, as documented by the value of the largest absolute error and the number of absolute differences greater than 10. Moreover, the panel over-represents single people, individuals with a higher level of education, respondents aged 25-34, and those in employment. It under-represents divorced/widowed people, those aged 65 and over, less educated individuals, and those who are economically inactive. Although the differences in the distributions by sex and area of residence are statistically significant, self-selection into the panel does not seem to have introduced major sources of bias.

To conclude, *Opinione.net* is not a representative sample of the online and, especially, of the general population, the latter being characterised by a larger bias.

*Nonresponse at the study stage*

In evaluating the impact of nonresponse at the study stage, I compare the characteristics of the ILC (responding) sample to those of the ILC selected sample members and the *Opinione.net* panel members (col. 8 and 9).

The demographic and socio-economic characteristics of the ILC (responding) sample are very similar to those of the sample selected to carry out the ILC survey. Indeed, the average absolute error is 0.7 and the largest absolute error is 1.2. There is only one statistically significant difference and no differences greater than 5 percentage points. There is some evidence that the ILC (responding) sample may underestimate those sample members who live in the North-East part of Italy.

When considering the comparison between the ILC (responding) sample and the panelists, I found a different pattern. The analysis of the overall measures of errors shows that the composition of the responding sample is different from that of the panelists. Indeed, the average absolute error is 4.2 and the largest absolute error is 7.6. With the exception of the variable occupation, the percentage differences between the other variables are all statistically significant. Despite the differences in sample composition, the selection bias is small. There are only three differences greater than 5 percentage points (and no differences greater than 10 and 15 percentage points). The ILC (responding) sample tends to over-represent men and people aged 65 and under-represent single people and those aged 24-35.

In brief, there are no major differences in the sample composition of the ILC (responding) sample and the ILC original sample. In addition, the ILC (responding) sample cannot be considered a representative sample of the *Opinione.net* panel, although the size of the bias is small.

Table 2. Impact of undercoverage and nonresponse on sample representativeness. Percentage differences and Accuracy metrics.

| | Undercoverage | Nonresponse at the recruitment stage | | | | Nonresponse at the joining stage | | Nonresponse at the study stage | |
|---|---|---|---|---|---|---|---|---|---|
| | Internet population compared to general population (col. 1) | ILC sample compared to: | | | | Panelists compared to: | | ILC sample compared to: | |
| | | Internet population | | General population | | Internet population (col. 6) | General populatio n (col. 7) | Panel members invited to survey (col. 8) | Panel members (col. 9) |
| | | No weights (col. 2) | Weights (col. 3) | No weights (col. 4) | Weights (col. 5) | | | | |
| **Panel 1. Percentage differences** | | | | | | | | | |
| Variables | | | | | | | | | |
| *Area of residence* | *** | *** | *** | Not sign. | *** | *** | *** | ** | * |
| North West | 2.2 | -2.8 | -0.3 | -0.5 | -1.3 | -3.7 | -1.5 | -1.2 | 0.9 |
| North East | 1.6 | -1.3 | 0.0 | 0.3 | 3.5 | 1.5 | 3.1 | -2.4 | -2.8 |
| Centre | 0.9 | -1.4 | 0.2 | -0.5 | 0.7 | -3.0 | -2.1 | -0.9 | 1.7 |
| South | -3.4 | 3.5 | -0.4 | 0.2 | -2.7 | 3.9 | 0.5 | 2.2 | -0.4 |
| Islands[a] | -1.3 | 1.9 | 0.6 | 0.5 | -0.2 | 1.4 | 0.0 | 2.3 | 0.5 |
| *Sex* | *** | Not sign. | *** | *** | *** | *** | *** | Not sign. | *** |
| Men | 4.8 | 0.3 | -1.9 | 5.1 | -1.5 | -7.3 | -2.5 | -0.2 | 7.6 |
| Women | -4.8 | -0.3 | 1.9 | -5.1 | 1.5 | 7.3 | 2.5 | 0.2 | -7.6 |

Table 2. *Continued.*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| *Marital status* | *** | *** | *** | *** | *** | *** | *** | Not sign. | *** |
| Single | 12.3 | 1.6 | 5.8 | 13.9 | 7.2 | 8.2 | 20.5 | -0.2 | -6.5 |
| Married | -5.4 | 2.8 | -0.7 | -2.6 | -0.8 | -2.0 | -7.3 | 0.4 | 4.8 |
| Divorced/widowed | -6.9 | -4.4 | -5.1 | -11.4 | -6.4 | -6.2 | -13.1 | -0.2 | 1.8 |
| *Age group* | *** | *** | *** | *** | *** | *** | *** | Not sign. | *** |
| 18-24 | 5.6 | -4.9 | -1.6 | 0.7 | -0.2 | -1.7 | 3.9 | -0.9 | -3.2 |
| 25-34 | 7.2 | -0.5 | -0.5 | 6.7 | 1.4 | 5.3 | 12.5 | -1.0 | -5.8 |
| 35-44 | 6.6 | -3.9 | -0.6 | 2.8 | 0.8 | -1.7 | 4.9 | 1.4 | -2.2 |
| 45-54 | 2.8 | -0.4 | 1.0 | 2.5 | 2.9 | -2.5 | 0.4 | 1.2 | 2.1 |
| 55-64 | -2.0 | 2.9 | 1.2 | 0.9 | 0.2 | -1.0 | -3.0 | 0.2 | 3.9 |
| 65+ | -20.3 | 6.7 | 0.5 | -13.5 | -5.1 | 1.5 | -18.8 | -0.9 | 5.2 |
| *Education* | *** | *** | *** | *** | *** | *** | *** | Not sign. | *** |
| Tertiary | 9.6 | 13.3 | 1.3 | 22.9 | -0.2 | 15.0 | 24.6 | -0.3 | -1.7 |
| Secondary | 13.5 | 2.3 | -0.3 | 15.8 | 0.1 | 0.8 | 14.3 | 0.9 | 1.5 |
| Primary | -6.2 | -14.8 | -1.0 | -21.0 | 0.1 | -14.1 | -20.3 | -1.0 | -0.6 |
| No education | -16.9 | -0.8 | -0.1 | -17.8 | -0.1 | -1.6 | -18.6 | 0.3 | 0.8 |

Table 2. *Continued*.

| *Occupation* | *** | *** | *** | *** | *** | *** | *** | Not sign. | Not sign. |
|---|---|---|---|---|---|---|---|---|---|
| In employment | 17.8 | -7.1 | -8.3 | 10.8 | -0.5 | -4.4 | 13.4 | 0.6 | -2.7 |
| Unemployed | 2.3 | -3.3 | -1.3 | -1.0 | -0.7 | -3.9 | -1.6 | 1.1 | 0.7 |
| Inactive | -20.1 | 10.3 | 9.6 | -9.8 | 1.2 | 8.3 | -11.8 | -1.7 | 2.0 |
| **Panel 2. Other metrics** | | | | | | | | | |
| Average absolute error | 11.0 | 3.2 | 2.0 | 7.9 | 1.7 | 3.3 | 9.4 | 0.7 | 4.2 |
| Number of significant differences from benchmark | 6 | 5 | 6 | 5 | 6 | 6 | 6 | 1 | 5 |
| Largest absolute error | 20.3 | 7.1 | 8.3 | 15.8 | 5.1 | 7.3 | 18.8 | 1.2 | 7.6 |
| Number of absolute differences greater than | | | | | | | | | |
| 5 | 6 | 2 | 4 | 4 | 3 | 6 | 1 | 0 | 3 |
| 10 | 2 | 3 | 0 | 6 | 0 | 2 | 7 | 0 | 0 |
| 15 | 4 | 0 | 0 | 4 | 0 | 0 | 3 | 0 | 0 |

Note: ***p≤0.001; **p≤0.01; *p≤0.05; Not sign.= not statistically significant.

[a] The category “Islands” includes Sardinia, and Sicily.

Modal categories for computing the average absolute error are shown in the Appendix 9.

## 3.4 Conclusions

The overall aim of this chapter is to investigate the impact of undercoverage and nonresponse on the representativeness of the Italian non-probability online panel *Opinione.net*. The first aim of this work is to investigate the impact of Internet undercoverage on sample representativeness and, in particular, to assess whether the Italian Internet population can be considered representative of the general population. Defining the Internet population as those individuals with Internet access and the skills to navigate the Internet, I have documented that the size of the Internet coverage is 48.1% (much smaller than the size of the population with Internet access) and have shown that the Internet population is not representative of the general population as it over-represents single people, educated individuals, and those in employment and under-represents people aged 65 and over, those with no education, and economically inactive individuals. These findings implicitly suggest that undercoverage may be a serious source of bias, because people's attitudes, beliefs, and opinions are often related to their age, level of education and employment conditions. They also suggest that research findings based on the analysis of data collected using web surveys would overestimate the opinions of a specific group of Italians, e.g., the highly educated. I therefore would advise survey practitioners against making (easy) inference to the general population when using data from web surveys.

Using a unique set of data on the panelists who are members of the *Opinione.net* non-probability online panel, this work also aims to explore the impact of nonresponse occurring at three different stages of the life of the panel (i.e. the recruitment stage, the joining and profiling stage, and the study stage) and evaluate the impact of quasirandomization weights on selection bias. A number of interesting findings stand out from my analysis. First, at the recruitment stage, the ILC (responding) sample is not a representative sample of the Internet population and in particular does not represent the general population, being characterised by a larger selection bias. However, quasirandomization is very effective in reducing the magnitude of the bias, especially in the case of the general population. After weighting, the ILC sample is more representative of the general population than the Internet population. This is good news - although some bias remains even after weighting, in some cases non-probability online panels may be used to draw inference on the general population. However, more research is needed to further confirm these preliminary findings. Second, my study has also shown that, at the joining stage, the panel *Opinione.net* is not a representative sample of the Internet population and in particular does not represent the general population, being characterised by a larger bias. Third, it has also documented that, at the specific study stage, there are no major differences in the sample composition of the ILC (responding) sample and the ILC original sample. Again, this is a reassuring finding, confirming that nonresponse at the study specific stage is completely at random. Being the first study of this type, there is urgent need of further research in this area. Finally, my study has shown that the ILC (responding) sample cannot be considered a representative sample of the *Opinione.net* panel, although the size of the bias is small.

My study has one important limitation - my work is based on the analysis of one single non-probability online panel. I cannot exclude the possibility that extending the analysis to other panels may lead to different results. Researching the quality of non-probability online panels poses specific practical issues, because of the lack of (available) data on the panelists and on non-response at the study specific stage. The contribution of the market research industry is therefore key to a better understanding of the quality of the data collected using non-probability online panels.

# 4

# Nonresponse in non-probability online panels

As discussed in Chapters 2 and 3, nonresponse in non-probability online panels may occur at different stages of the life of the panel (i.e. recruitment, joining procedures, profiling, sampling for specific studies, and panel maintenance). In this chapter, I look at the specific study stage. While in the previous chapter I assessed sample representativeness, this chapter assesses the occurrence and magnitude of nonresponse, introducing other indicators that measure the nonresponse phenomenon in the ILC survey. After reviewing recent literature that reports empirical studies on panel survey nonresponse and stating my specific research aims (section 4.1), I describe the research design of this part of my work (section 4.2), I present the results of my analyses (section 4.3), and close the chapter with some concluding remarks (section 4.4).

## 4.1 Literature review and research aims

Research on the nature and magnitude of nonresponse at the specific study stage in non-probability online panels is scant. The few papers on these topics dealt with three main issues: i) the analysis of the types of response (e.g. response, break-off, and refusal) and the discussion of measures of nonresponse process (e.g. response rate, break-off rate, and refusal rate); ii) the occurrence and magnitude of unit nonresponse bias in survey estimates (i.e. differences between respondents and nonrespondents on comparable variables), and iii) the occurrence and magnitude of item nonresponse (i.e. missing answers to specific questions). In reviewing the relevant literature on opt-in panels in survey research, I assessed whether - and how - each issue is addressed at the specific study stage. I state the research questions addressed in this chapter at the end of each section.

### 4.1.1 Types of response and nonresponse process

There is little research that analyses the response process in non-probability online panels in general and, in particular, at the study specific stage. I found no studies that explicitly report the types of response for a given survey, but rather implicitly refer to some of them when calculating the response metrics.

One of the reasons for this lack of knowledge are difficulties connected to the calculation and interpretation of response metrics that are commonly used in surveys. For example, the calculation of response rate poses specific issues. Indeed, as opt-in panels adopt non-probability recruitment strategies to reach potential panel members, the sampling frame is constituted of people with unknown probabilities of selection (AAPOR, 2016a), and therefore, computing response rate is problematic. First, there are broad inferential concerns, then, the denominator of this rate is often unknown (AAPOR, 2016a), and lastly, researchers can calculate different measures at each stage involved in building an online panel.

Scholars have suggested various response metrics and several professional associations have released guidelines for computing metrics for opt-in panels (see Callegaro and DiSogra, 2008 for a review). For example, since 2006 the Standard Definitions document, issued by AAPOR (that published the latest version in 2016 (AAPOR, 2016a)), included “Internet surveys of specific named persons” which details how to compute response rates for web surveys, and specifically for non-probability Internet panels.

Despite the availability of guidelines, there is little agreement on the terminology used to define various metrics (Callegaro and DiSogra, 2008). In particular, i) different wordings may refer to the same method of calculation, and ii) the same wording may refer to different methods of calculation. Here I present the results from the literature that looks at the metrics adopted at the specific study stage.

Regarding different wordings for the same method of calculation, “start rate”, “participation rate”, and “completion rate” are the most widely used but also the most conflicting terms found in the literature. For example, the AAPOR Task Force (Baker *et al.*, 2010a) and ISO 26362 (International Organization for Standardization, 2009) use “participation rate” to refer to the number of respondents who have provided a usable response, divided by the total number of initial personal invitations requesting participation. On the other hand, Callegaro and DiSogra (2008) call the same rate “completion rate”. Göritz and Luthe (2013b, and 2013c) use “participation (or response) rate” to note whether a panelist calls up the study’s first page or not. In contrast, other scholars (Göritz, 2014; Göritz and Luthe, 2013a; Mavletova, 2013; Mavletova and Couper, 2013) define the same rate as “start rate”.

With regard to the same wording for different methods of calculation, “participation rate” is calculated in two different ways by scholars, i.e. including i) the respondents who have provided a usable response (Baker *et al.*, 2010a; International Organization for Standardization, 2009) or ii) the panelists who called up the study’s first page (Göritz and Luthe, 2013b and 2013c).

A number of studies (Arn *et al.*, 2015; Buskirk and Andrus, 2014; Couper *et al.*, 2013; Keusch *et al.*, 2014; Mavletova, 2013; Mavletova and Couper, 2013) focused on nonresponse in the response process, and computed “break-off rate”, that occurs when the survey was opened but not finished (Callegaro and DiSogra, 2008), using the same method of calculation. This provides further details about the quality of the response process.

There are two further issues connected to the use of response metrics in non-probability panel surveys. The first one is that some studies do not report the method of calculation of the metrics they adopt. Within this context of conflicting terminology, it is difficult to disentangle which rate is computed when no or insufficient specifications are provided. For example, Craig and colleagues (2013), and Pedersen and Nielsen (2016), generically refer to “response rate”, without giving an accurate definition of their intended meaning.

The second issue is that, in most recent studies on nonresponse at the specific study stage, I did not find examples of the calculation of two additional response metrics, i.e. “refusal rate” and “contact rate”, for which AAPOR (2016a; 2016b) provided the operational definitions and formulas.

Against this background, the first aim of this work is to describe the nonresponse process in the ILC survey and to document the quality of the response process. In particular, I detected refusals, as well as break-offs, thanks to the availability of paradata from the ILC survey, and then I computed different response metrics, i.e. cooperation rate, break-off rate and refusal rate, following the guidelines provided by AAPOR (2016a; 2016b) and Callegaro and DiSogra (2008). Findings from this part of the study contribute to filling in a notable gap, given that there are only a few studies describing the response/nonresponse process, and therefore contributing to document the quality of the response process.

### 4.1.2 Unit nonresponse bias

Reporting response metrics in a study is not sufficient to identify unit nonresponse, because not only the number but also the characteristics of nonrespondents impact on survey accuracy. When compared to research on response metrics at the specific study stage of the life of an opt-in panel, research on nonresponse bias is even scarcer. Recall that nonresponse bias occurs when the nonresponse rate is high, respondents' characteristics are different from those of nonrespondents, and these differences have an impact on survey estimates. To perform the analysis on nonresponse bias, data on nonrespondents (other than data on respondents) must be available. Unfortunately, panel vendors do not usually provide researchers with panel data for the nonrespondents' sample. Moreover, even when a good deal is known about panel members, the analysis of the differences between respondents and nonrespondents to a given survey is seldom done. (Baker *et al.*, 2010a)

As a consequence, empirical studies were limited to the description of the respondents' socio-demographic profile (Knapton and Myers, 2005) and sometimes of the characteristics of the respondents' panel experience (Keusch *et al.*, 2014; Keusch, 2013), e.g. membership tenure, motives for joining the panel, and number of other online panels enrolled in.

Two studies compared panel survey respondents with other responding samples from surveys adopting different methods of data collection. Cho and colleagues (2015) compared the socio-demographic characteristics of panel survey respondents with those of respondents from a river sample (i.e. a sample of respondents recruited when they are online, often presenting him/her a survey invitation while he/she is engaged in some other online activity) and found that people aged 45 and over, those with higher education, and members of households with higher incomes were more likely to belong to the panel responding sample. Brown and colleagues' (2012) work resulted in contrasting findings. Comparing panel survey respondents with those from other Public Participation Geographic Information Systems (PPGIS) surveys (completed with random household sampling, on-site recruitment, or self-selected public sampling), the authors found no differences in age and level of education, whereas panelists reported lower income levels.

There is only one study in which data for nonrespondents were available. It is an experimental design testing the effect of donation incentives and type (i.e. altruistic vs egoistic) of text appeal in the e-mail invitation on response (Pedersen and Nielsen, 2016). As the authors had background data for all panelists, they could assess nonresponse bias. They did not directly measure the magnitude of bias, but checked for its potential effect on response, including sex, age, and region of residence in the logistic regression that estimates the survey response rate. Results from the model showed that people aged 50 and over are more likely to answer, whereas people aged 18-29, and those living in the Zealand region of Denmark are less likely to answer.

The second aim of this work is to address (unit) nonresponse bias in the ILC survey estimates. As Demetra provided all *Opinione.net* panelists' profile data, I was able to compare the ILC respondents' characteristics with those of nonrespondents. In particular, I investigated i) the magnitude of nonresponse bias and ii) the impact of the socio-demographic characteristics on response propensity. Results from this part of the study contribute to fill in the gap in the existing literature with respect to i) expanding the knowledge on nonresponse bias at the specific study stage of the life of an opt-in panel, as I found only one study focussing on this, and ii) contributing to the debate on nonresponse bias by computing new indicators and performing analyses (see the Methods section for a detailed discussion) that are not currently used to assess nonresponse bias in non-probability online panel surveys.

### 4.1.3 Item nonresponse

In addition (but not necessarily related) to unit nonresponse, that focuses on potential respondents who do not take part in the survey, another different aspect of nonresponse is item nonresponse, that occurs when respondents skip specific survey questions. This is a relevant issue in sample surveys, because missing answers fail to give information on characteristics, behaviours, and opinions of respondents, and therefore might result in inaccurate survey estimates. The level of item nonresponse depends on several factors, including the question, the overall survey design and its strategies to minimise missing data, and the respondents' characteristics (e.g. their motivations and interests). It can vary greatly in web surveys, according to different key design choices, such as offering prompts for missing answers (Tourangeau *et al.*, 2013). Likely, the average magnitude of item nonresponse is generally low, and also for web surveys. In online panels, as well as in market research surveys, very often there is no item nonresponse, because hard prompts[20] are applied to the questions (Callegaro *et al.*, 2015). On the other hand, permitting nonanswers is common in academic surveys (Tourangeau *et al.*, 2013). To measure item nonresponse Callegaro and

[20] The increase in the number of break-offs might represent the disadvantage of forcing an answer. This should be taken into account when designing web surveys and panel maintenance strategies.

colleagues (2015) suggested distinguishing between item nonresponse at the question level and item nonresponse at unit level and proposed one indicator for each level (see the Methods section for details).

Recent studies conducted on samples from non-probability online panels documented that item nonresponse is an uncommon phenomenon. In their work, Göritz and Luthe (2013a, 2013b, and 2013c) reported a minimum item nonresponse of 0.00 and a maximum of 0.07 in different experimental groups (in a range from 0 to 1). Couper and colleagues (2013) computed mean missing data rates across 13 items in two experiments and obtained values that ranged from 0.00 to 0.11, on a scale of 0 to 1. In two experimental designs testing the impact of survey layout (Arn *et al.*, 2015) and survey mode (Buskirk and Andrus, 2014) on response to specific items the percentages of missing answers varied from a minimum of 0.9% to a maximum of 3.6%.

Although the low magnitude of item nonresponse in opt-in panel surveys, preliminary evidence shows that item nonresponse can vary in relation to the type of question. Indeed, two studies (Arn *et al.*, 2015; Couper *et al.*, 2013) exploring the impact of designs with different layouts on response to specific survey questions, found that missing data rates differed significantly across the experimental questionnaire versions. In particular, Arn and colleagues (2015) measured item nonresponse on two questions administered using two designs (i.e., not adapted vs adapted for multi-device use) and found that, albeit low, it was significantly different between the two treatment groups. Similarly, evidence from the experimental designs implemented by Couper and colleagues (2013) showed that, across the three types of grids (i.e., baseline, dynamic, and split), the missing data rate, albeit low, differed significantly.

In the reviewed literature on non-probability online panel surveys item nonresponse was computed to evaluate the response differences between treatment groups, and to test the correlation between missing answers, respondents' behaviour, and respondents' panel experience. For example, scholars evaluated the impact of lotteries (Göritz and Luthe, 2013a; Göritz and Luthe, 2013b; Göritz and Luthe, 2013c), survey layout (Arn *et al.*, 2015), type of grids (Couper *et al.*, 2013), and survey mode (Buskirk and Andrus, 2014) on item nonresponse.

Greszki and colleagues (2014) performed a regression analysis to estimate response time spent on the survey pages that are relevant for 'no answer' as a function of item nonresponse (explanatory variable). Results showed that the correlation of response time with no answer proves statistically significant. In a study about professional respondents in opt-in panel surveys (Hillygus *et al.*, 2014), the authors ran two regression models that estimate the percentage of missing answers (dependent variable). They used, in the first model, the number of panels to which the respondent belongs, and in the second one, the number of completed surveys in the past four weeks, as explanatory variables. They found that those belonging to more online panels or answering more surveys in the past four weeks had a higher percentage of missing responses.

When looking at the overall measures used to calculate item nonresponse in non-probability online panel surveys, studies mainly focused on the level of question (and not of unit) and, as I have already mentioned, compute, percentages of missing answers and (mean) missing data rates.

Drawing on Callegaro and colleagues (2015), the third aim of this work is to address item nonresponse in a comprehensive way, at the level of both question and unit. As the literature review highlighted, to study item nonresponse is not a common task when analysing data from opt-in panel surveys. Thus, this part of my study provides further results on the occurrence and magnitude of item nonresponse in non-probability online panel surveys. Moreover, including the indicator of item nonresponse at unit level, findings from the ILC survey data contribute to enrich the knowledge on the respondents' practice of skipping questions.

## 4.2 Research design

### 4.2.1 Data

The data I used for this part of the study are the two datasets from the non-probability online panel *Opinione.net*, i.e. socio-demographic data on all registered panelists and data from the Italians' Living Conditions (ILC) survey (see Chapter 2 for details).

### 4.2.2 Methods

To pursue my aims I i) identified the features of nonresponse process and calculated the response metrics, ii) estimated nonresponse bias, and iii) measured item nonresponse. The first two groups of analysis deal with unit nonresponse, whereas the third one refers to item nonresponse. I describe below the methods used for each type of analysis.

*Nonresponse process – Response metrics*

To describe the types of response/nonresponse to the ILC survey (i.e. response, break-off, refusal, and other reasons for not responding), and further explore nonresponse, I computed three response metrics: cooperation rate, break-off rate, and refusal rate[21]. As I previously mentioned, I was able to carry out this analysis because Demetra provided me with the paradata saved by the survey software.

To calculate the cooperation rate, I drew on AAPOR standard definitions (2016a, and 2016b). This type of rate is what in literature is usually referred to as "completion rate", and is the percentage of completed interviews, including those that were screened out, divided by all the eligible units ever contacted (i.e. the total number of invitations minus non eligible interviews). In particular, I adopted the Cooperation Rate 1 (AAPOR, 2016b), that is expressed by the following formula:

(1) Cooperation Rate 1 = I / I + P + R + O

Note that *I* are complete interviews, *P* are partial interviews, *R* are refusals and break-offs, and *O* are other non interviews/non-refusals.

Unfortunately, respondents do not always complete the questionnaire and can display two different (non)response behaviours, i.e., they interrupt the completion of the questionnaire or they do not start filling it out.

When respondents get partway through the questionnaire but fail to complete it, they are defined as "abandonments," "break-offs," "drop-outs" or "partials". "Partials" also refers to those respondents who read, or at least view, every question in the questionnaire and submit it after reaching the final question, but decline to answer all of the questions (AAPOR 2016a). In the ILC survey I adopted a restrictive definition of break-offs, i.e., individuals who start filling out the questionnaire, but do not submit it. Partials were included into the complete interviews, as the questionnaires were submitted. To calculate the break-off rate I drew on Callegaro and DiSogra (2008) and applied the following formula:

(2) Break-off Rate = BO / I + P + BO

Note that the number of interviews that are classified as break-offs (*BO*) is divided by the sum of complete (*I*), partial (*P*), and break-off (*BO*) interviews.

[21] I did not calculate the contact rate because the ILC sample does not contain 'non contacts'.

The second (non)response behaviour is refusal, that occurs when respondents receive the invitation to participate in the survey, but decide not to cooperate. To calculate the refusal rate, AAPOR provides a formula that includes both refusals and break-offs, but it also considers an alternative restrictive operational definition that includes refusals only. I used the latter and computed the proportion of all cases in which the respondent refuses to be interviewed, of all potentially eligible cases (AAPOR 2016a). The Refusal Rate 3 (AAPOR, 2016b) is expressed by the formula:

(3) Refusal Rate 3 = R / I + P + R + NC + O

Note that the number of refusals (*R*) is divided by the sum of complete (*I*), partials (*P*), refusals (*R*), non contact (*NC*), and other non interview (*O*).

*Unit nonresponse bias – Using rich sampling frame data to assess nonresponse bias*

There are different methods to assess nonresponse bias, e.g. i) response rate comparisons across subgroups, ii) comparisons to similar estimates from other sources (e.g., census data for the population, or a high-quality government survey), iii) studying variation within the existing survey conducting nonresponse follow-up studies, and iv) using rich sampling frame data or supplemental matched data (Groves, 2006). To assess the impact of nonresponse on the sample composition at the recruitment stage I adopted the second approach (see Chapter 3), whereas in the analysis on nonresponse bias at the specific study stage, I used the fourth approach, i.e. 'rich sampling frame data'. In particular, I used *Opinione.net* panelists' sampling frame to identify the target population that reports many attributes for each population member (Groves, 2006). I compared the respondents' characteristics with those of nonrespondents, using socio-demographic variables (i.e. sex, age, marital status, education, occupation, and geographic area of residence). The reason for this choice is that socio-demographics were the only available and comparable variables for both nonrespondents (as they are stored in the panelists' dataset), and respondents (as they are both stored in the panelists' dataset, and asked again in the ILC questionnaire to update the profile information)[22].

The comparison aims to test whether the socio-demographic profile is correlated with nonresponse bias. Within the approach of the 'rich sampling frame data', I applied various

[22] I checked for other variables from panel data, but any other was comparable in the two samples (i.e., respondents and nonrespondents) or had a number of valid cases that was sufficient to perform consistent analyses.

methods of analysis, proposed by different scholars, following three steps. First, I performed a bivariate analysis, drawing on Bethlehem and colleagues (2011), who mainly focused on the response propensity, looking at the share of respondents and nonrespondents within each category of the socio-demographic variables (thus, computing percentages based on the socio-demographic variable). To check for the independence between the auxiliary variables (i.e. socio-demographics) and the response behaviour, I calculated Cramer's *V* statistic.

Then, I carried out again a bivariate analysis. This time I looked at the distributions of respondents and nonrespondents for each socio-demographic variable (thus, computing percentages based on the response variable). To test for the statistical significance of the differences between respondents and nonrespondents, I performed the Chi Square test. Drawing on the relevant literature about nonresponse bias in surveys (Biemer, 2001; Groves, 2006; Groves and Peytcheva, 2008), I used the results from the bivariate analysis to calculate three point measures (i.e. the percentage point differences, the percentage of absolute relative nonresponse bias, and the nonresponse bias) and three overall nonresponse metrics (i.e. the average differences between respondents and nonrespondents, the mean percentage absolute relative bias, and the average nonresponse bias). The values of all these metrics range (potentially) from 0 to 100, except for the percentage absolute relative nonresponse bias that ranges from 0 to 1; the larger their values, the bigger the bias. The "percentage point differences" and the "average differences between respondents and nonrespondents" are similar, respectively, to the "percentage point error" and the "average absolute error" that I calculated to address sample composition (see Chapter 3 for details), but they are calculated including all the categories of the socio-demographic variables, and not only the modal ones. The "percentage absolute relative nonresponse bias" is the percentage difference between the respondents ($\hat{y}_r$) and the full sample of contacted panelists ($\hat{y}_n$) divided by the percentage of this full sample (Groves and Peytcheva, 2008):

(4) Percentage absolute relative nonresponse bias = $\left|\frac{\hat{y}_r - \hat{y}_n}{\hat{y}_n}\right|$

The "mean percentage absolute relative bias" is the average (for the six variables used for the comparison) of the percentages absolute relative nonresponse bias. The "nonresponse bias" is the product of the nonresponse rate multiplied by the difference between respondent and nonrespondent means (Groves, 2006). As Peytchev (2013) pointed out, different types of estimates can be used to calculate the nonresponse bias. Thus, for purposes of this discussion, nonresponse bias, and also the percentage absolute relative nonresponse bias, are defined for an estimate of the percentage (and not of the mean), as in formulas 4 and 5.

(5) $Bias(\hat{y}_r) = \frac{m}{n}(\hat{y}_r - \hat{y}_m)$

Note that $\hat{y}_r$ is the respondent percentage, $\hat{y}_m$ is the nonrespondent percentage, $m$ is the number of nonrespondents, and $n$ is the total number in the full sample. The covariance term included in expression (5) measures whether the distinctiveness of nonrespondents (relative to respondents) changes as the nonresponse rate changes, with the assumption of no measurement bias (Groves, 2006).

Calculating the average (for the six variables used for the comparison) of the nonresponse biases, I obtained the "average nonresponse bias" (Biemer, 2001).

Lastly, drawing on Bethlehem and colleagues (2011), I ran a logistic regression model to estimate the probability ($\pi_i$) of responding to the ILC survey, as expressed by the formula:

(6) $logit(\pi_i) = \beta_0 + \sum_{j=1}^{J} \beta_j \, x_{ij}$

Note that $\beta_0$ is the reference category of each variable included into the model, $\beta_j$ are the regression coefficients, and $x_{ij}$ are the explanatory variables (i. e., the socio-demographic characteristics).

*Indicators of item nonresponse*

The last part of the analysis focused on the respondents' sample only and assessed item nonresponse. To detect "true" missing answers, the instrument used to administer the questionnaire should permit respondents to proceed without answering all questions. This is the case of the software used to administer the ILC survey. The only mandatory questions of my questionnaire i.e. that do not have the option "no answer", were sex and year of birth, in addition to two other questions (about household income and time spent watching TV) that include the "don't know" answer. Thus, I excluded these four variables from the analysis and I calculated the percentage of asked survey items with missing responses for all the other 30 variables (see the Results section for the complete list).

Drawing on Callegaro and colleagues' (2015) work I calculated two indicators, one at the question level and the other at unit level. The first one is the "item nonresponse rate" (INR) at the level of a specific item, and is defined as the number of units with item nonresponse divided by all eligible units exposed to the item:

$$(7)\ INR = \frac{Number\ of\ units\ with\ item\ nonresponse}{All\ eligible\ units\ exposed\ to\ the\ item}$$

I also calculated the INR for all questions in the questionnaire as the "mean item nonresponse rate" across all items.

The second indicator is the "unit-level item nonresponse" observed at the respondent level, and is defined as the share of items answered among all items to which a certain unit was exposed:

$$(8)\ UIN = \frac{Number\ of\ all\ item\ nonresponse\ for\ a\ certain\ unit}{Number\ of\ all\ items\ to\ which\ a\ certain\ unit\ was\ exposed}$$

As for the INR, I calculated the UIN also for all questions in the questionnaire as the "mean unit-level item nonresponse rate" across all units.

Both indicators range from 0 to 1, where 0 indicates no missing answers and 1 indicates that i) all units provide no answer to a certain item (for the INR indicator) or ii) all items have missing data for a certain unit (for the UIN indicator).

## 4.3 Results

In the following sections, I look separately at the results of my analyses on nonresponse process, nonresponse bias, and item nonresponse.

### 4.3.1 Nonresponse process

Table 1 sums up the nonresponse process, identifying the various types of response/nonresponse from the ILC survey sample. The whole sample of invited panelists consisted of 3,908 members, from whom I obtained 97.5% of eligible questionnaires. The invalid cases are negligible (97 records) and identify attempts to participate in the survey that failed the respondent quality check to avoid getting duplicate or fraudulent respondents. Response to the survey is high (Tourangeau *et al.*, 2013): the percentage of submitted questionnaires is 51.4%. However, the number of refusals is fairly large, i.e. 37.7%, of which only 8 panel members clicked on survey decline link (explicit refusal), and the others unanswered, ignoring the e-mail invitation, or unsubscribed the panel membership (implicit refusal). Nonresponse is the difference between all the contacted panelists (i.e. eligible, and not eligible interviews) and the response, and includes 1,900 units.

Table 1. Nonresponse process to the ILC survey.

| Eligibility | Types of response/nonresponse of the units | N | % |
|---|---|---:|---:|
| Eligible | Response | 2,008 | 51.4 |
| | Break-off | 96 | 2.5 |
| | Other reason for nonresponse | 234 | 6.0 |
| | Explicit refusal | 8 | 0.2 |
| | Implicit refusal:<br>- Unanswered<br>- Unsubscribed | <br>1,396<br>69 | <br>35.7<br>1.8 |
| Not eligible | Bad IP or project token | 16 | 0.4 |
| | Bad ID | 81 | 2.1 |
| (Full sample) | | (3,908) | (100.0) |

The calculation of the response metrics gave the following results: the Cooperation Rate 1 (AAPOR, 2016b) is 52.7%[23], the Refusal Rate 3 (AAPOR, 2016b) is 38.7%, and the Break-off Rate (Callegaro and DiSogra, 2008) is 4.6%. Unfortunately, to compare my results to those of other studies is not an easy task, especially because of i) the lack of complete information about these response metrics in the papers, and ii) the different characteristics of the study designs. Nonetheless, focussing on the cooperation rate only, I may conclude that the response process was ‘virtuous’, as Tourangeau and colleagues (2013) documented a decline of “response rates” for non-probability panels from a high of almost 20 percent in 2002, to the low that was in single digits in 2010.

[23] This high cooperation rate, that is uncommon for a survey conducted with a non-probability panel (Tourangeau *et al.*, 2013), may be due to different aspects of the panel recruitment and maintenance, to the high “incidence rate” (i.e. the percentage of respondents selected for a survey that match the target group of the survey) as the ILC survey was conducted on the general population, and to the incentives.

### 4.3.2 Nonresponse bias

I discuss the results from bivariate analysis, first, looking at the response propensity and the differences between respondents and nonrespondents for each socio-demographic variable, and then focussing on the nonresponse metrics. I conclude with the results from the logistic regression model.

Following on from Bethlehem and colleagues (2011), I calculated Cramer's *V* statistic to test the independence between the auxiliary variables (i.e. socio-demographics) and the response indicator (i.e. response vs nonresponse) for the ILC sample. The results are displayed in Table 2.

Table 2. Tests of independence between the socio-demographic variables and the response behaviour to the ILC survey.

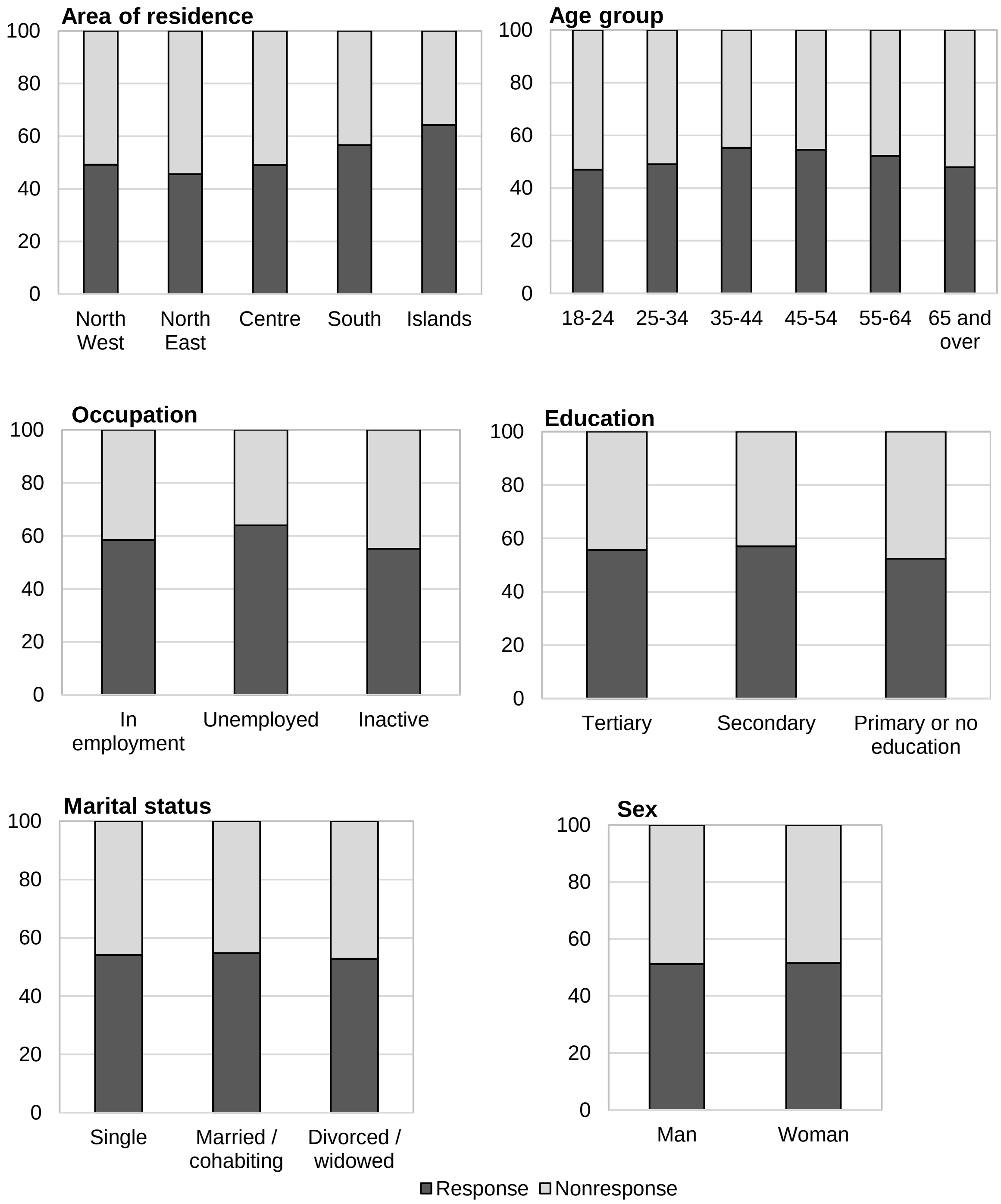
| Socio-demographic variable | Cramer's *V* |
|---|---|
| Area of residence | 0.111 |
| Age group | 0.062 |
| Occupation | 0.053 |
| Education | 0.028 |
| Marital status | 0.011 |
| Sex | 0.004 |

Cramer's *V* ranges from 0 to 1, where 0 corresponds to "no association", and 1 corresponds to "complete association". The values of the Cramer's *V* test, reported in Table 2, indicate that all the socio-demographic variables have a weak relationship with the response behaviour. This is an encouraging finding, because it suggests a minor risk of a nonresponse bias for survey variables that are related to these variables.

Graph 1 shows that the results from the analysis on the share of respondents and nonrespondents for each socio-demographic variable confirm those from the independence tests. There are small differences in the share of respondents and nonrespondents for specific categories of the variables area of residence and age group. In particular, panelists who live in the South of Italy and in the Italian islands[24] are more likely to respond, whereas both younger and older people are less likely to respond than the middle-age groups.

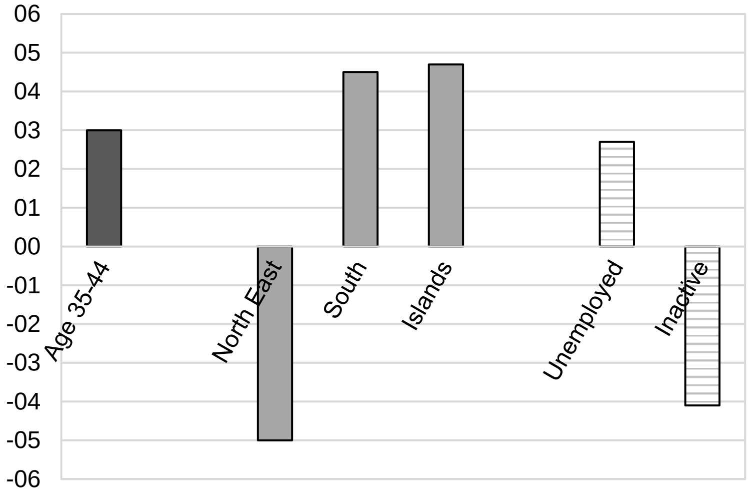
[24] Recall that this category includes Sardinia, and Sicily.

Graph 1. Response propensity of the ILC sample for the socio-demographic variables.



For the second step, I carried out again a bivariate analysis to look at the socio-demographic differences between respondents and nonrespondents. Results from this analysis are consistent with the findings from the analysis on the response composition within each category of the socio-demographic variables (see Graph 1). Graph 2 shows the statistically significant differences between respondents and nonrespondents. The Chi Square tests, calculated when performing the bivariate analysis to check for the statistical significance of the differences between respondents and nonrespondents, highlight two findings, that are displayed in Graph 2. First, in line with the previous analysis, respondents are more likely to live in the South of Italy and in the Italian islands and be in the middle-age groups (specifically, in the class 35-44). Moreover, other additional significant differences in areas of residence and occupation come out and show that nonrespondents are more likely to live in the North East of Italy and to be economically inactive, whereas respondents are more likely to be unemployed.

Graph 2. Percentage point differences between respondents and nonrespondents to the ILC survey.



Note: The graph reports only the statistically significant differences. Area of residence: p≤0.001. Age group and occupation: p≤0.01.

To assess the magnitude of the differences in the estimates, I computed three point measures (i.e. the percentage point differences, the percentage absolute relative nonresponse bias, and the nonresponse bias), and three overall nonresponse metrics (i.e. the average differences between respondents and nonrespondents, the mean percentage absolute relative bias, and the average nonresponse bias). In Table 3 the values of the point measures are reported in Panel 1 and those of the overall nonresponse metrics in Panel 2. Positive and negative values of the point measures indicate that the corresponding categories are over or under-represented, respectively. The variable distributions are shown in Appendix 10.

As documented in the analysis of the nonresponse bias, the socio-demographic characteristics of the responding sample are not different from those of the nonrespondents. The average differences between respondents and nonrespondents is 2.0, the mean percentage absolute relative bias is 0.1, and the average nonresponse bias is 1.0. The magnitude of the nonresponse bias is negligible: i) all the percentage point differences (except one) are lower than 5, ii) all the values of the percentage absolute relative nonresponse bias (except one) are lower than 0.2, and iii) the highest value of nonresponse bias is a negative value at -2.4. Taking the three point measures together, respondents slightly over-represent people aged 35-54, those living in the South of Italy and in the Italian islands, unemployed people, and individuals with secondary education, whereas respondents slightly under-represent those living in the North East of Italy, economically inactive people, and the less- or no-educated individuals. These results seem to reflect the cultural and economic differences between the North and the South (also including the islands) of Italy. As it is common knowledge that the unemployment rate is substantially higher in the South of Italy than in the North, one can speculate that panelists from the South who do not have an income through regular employment might be more attracted to earning money through panel membership than Northern panel members.

Table 3. Occurrence of nonresponse bias. Point measures and nonresponse metrics in the ILC sample.

| Socio-demographic variables | Point measures | | |
|---|---|---|---|
| | Percentage point differences | Percentage absolute relative nonresponse bias | Nonresponse bias |
| **Panel 1. Point measures** | | | |
| *Sex* | | | |
| Man | -0.4 | 0.0 | -0.2 |
| Woman | 0.4 | 0.0 | 0.2 |
| *Age group*** | | | |
| 18-24 | -1.8 | 0.1 | -0.9 |
| 25-34 | -2.1 | 0.0 | -1.0 |
| 35-44 | 3.0 | 0.1 | 1.4 |
| 45-54 | 2.4 | 0.1 | 1.2 |
| 55-64 | 0.4 | 0.0 | 0.2 |
| 65+ | -1.9 | 0.1 | -0.9 |

Table 3. *Continued*.

| | | | |
|---|---|---|---|
| *Area of residence**** | | | |
| North West | -2.4 | 0.0 | -1.2 |
| North East | -5.0 | 0.1 | -2.4 |
| Centre | -1.8 | 0.0 | -0.9 |
| South | 4.5 | 0.1 | 2.2 |
| Islands | 4.7 | 0.2 | 2.3 |
| *Occupation*** | | | |
| In employment | 1.4 | 0.0 | 0.6 |
| Unemployed | 2.7 | 0.1 | 1.1 |
| Inactive | -4.1 | 0.0 | -1.7 |
| *Marital status* | | | |
| Single | -0.4 | 0.0 | -0.2 |
| Married | 0.9 | 0.0 | 0.4 |
| Divorced/widowed | -0.5 | 0.0 | -0.2 |
| *Education* | | | |
| Tertiary | -0.6 | 0.0 | -0.3 |
| Secondary | 2.1 | 0.0 | 0.9 |
| Primary or no education | -1.5 | 0.1 | -0.7 |
| **Panel 2. Nonresponse metrics** | | | |
| Average differences between respondent and nonrespondent | 2.0 | | |
| Mean percentage absolute relative bias | | 0.1 | |
| Average nonresponse bias | | | 1.0 |

Note: ***p≤0.001; **p≤0.01; *p≤0.05. There are no differences that are statistically significant at the level of 0.05.

Results from the Cramer's *V* test and the bivariate analysis showed that area of residence, age group, and occupation were the variables with the strongest and significant relationship with the response behaviour. Thus, I used these variables to perform a multivariate analysis, and check for potential spurious relationships between the socio-demographic variables. I ran a logistic regression model that reveals the net effect of each explanatory variable on the response (i.e. the dependent variable), adjusted for the effect of all the other socio-demographic variables included in the analysis. Table 4 reports the results of the logistic regression.

Table 4. Logistic regression model for the response behaviour.

| Variables | Odds ratios |
|---|---|
| *Area of residence* (North West)[a] | |
| North East | 0.497*** |
| Centre | 0.440*** |
| South | 0.532*** |
| Islands | 0.707* |
| *Age group* (18-24)[a] | |
| 25-34 | 0.821 |
| 35-44 | 0.770 |
| 45-54 | 1.075 |
| 55-64 | 1.140 |
| 65+ | 1.056 |
| *Occupation* (in employment)[a] | |
| Unemployed | 1.100 |
| Inactive | 1.349* |
| *Constant* | 2.321*** |
| *N* | 3,457 |

[a] Category of reference. ***p≤0.001; **p≤0.01; *p≤0.05. There are no estimates that are statistically significant at the level of 0.01.

Three findings stand out. First, compared to living in the North West of Italy, living anywhere else in Italy, reduces the probability of response. In addition, economically inactive people, compared to those in employment, have a higher probability (at the 0.05 level of significance) of participating in the ILC survey. Lastly, contrary to the results from the bivariate analysis, the net effect of age group and of the "unemployed" category of occupation disappear.

To sum up, this part of the work produced encouraging findings. Indeed, evidence from all the analyses on nonresponse bias, potentially due to the socio-demographic differences between respondents and nonrespondents to the ILC survey, prove that the risk of significant distorsions introduced by the response behaviour is limited to the geographic area of residence.

### 4.3.3 Item nonresponse

My analysis on item nonresponse shows that skipping questions is not a common practice when participating in a panel survey.

First, I focus on the question level. As shown in Table 5, the percentages of missing answers, provided by the ILC survey respondents, vary from a minimum of 0.0 to a maximum of 7.7. The values of the item nonresponse rate (INR) range from 0.00 to 0.41 and the mean INR is 0.02. Table 5 displays the specific values of INR for each item. Although the INR is low overall, some items about using the Internet in specific places, and consuming alcoholic drinks report a little bit higher values.

Table 5. Item nonresponse rate (INR) for the 30 items from the ILC survey questionnaire.

| Items | INR |
|---|---|
| Internet use (frequency) | 0.0025 |
| Internet use (place): | |
| at home | 0.0015 |
| at work | 0.0453 |
| at school/university | 0.0767 |
| at other people’s houses | 0.0593 |
| Elsewhere | 0.0538 |
| Internet activities: | |
| sending or receiving e-mails | 0.0045 |
| telephoning over the Internet | 0.0090 |
| posting messages to chat sites/blogs | 0.0105 |
| using instant messaging | 0.0100 |
| participating in social networks | 0.0090 |

Table 5. *Continued*.

| | |
|---|---|
| Consumption of: | |
| Legumes | 0.0025 |
| Potatoes | 0.0105 |
| salty snacks | 0.0030 |
| candy/cakes/ice creams | 0.0015 |
| Alcohol consumption | 0.0015 |
| Number of glasses of alcoholic drinks | 0.4084 |
| Tobacco consumption | 0.0015 |
| Socio-political participation: | |
| I attended a political meeting | 0.0050 |
| I took part in a political demonstration | 0.0040 |
| I listened to a political debate | 0.0040 |
| I gave money to a political party | 0.0055 |
| I gave money to an association | 0.0035 |
| I did voluntary work | 0.0030 |
| Watching TV | 0.0020 |
| Education | 0.0000 |
| Occupation | 0.0010 |
| Marital status | 0.0005 |
| Region of residence | 0.0085 |
| Environmental problems I’m worried about | 0.0000 |

Then, focussing on unit level (i.e. the respondent), the mean number of all item nonresponse for a certain respondent is 0.73. The values of the unit-level item nonresponse (UIN) range from 0.00 to 0.43 and the mean UIN is 0.02.

Table 6 shows the values of the UIN for the respondents that reported lower levels of item nonresponse (1,975 units) to the questions of the ILC survey questionnaire. The respondents who answered all the items to which they were exposed (UIN=0.00) are 1,041, and are mainly young individuals, employed people, and those with secondary education.

Table 6. Unit-level item nonresponse (UIN) for the 1,975* respondents to the ILC survey.

| Units | UIN |
|---:|---:|
| 1,041 | 0.00 |
| 716 | 0.03 |
| 49 | 0.03 |
| 67 | 0.07 |
| 9 | 0.07 |
| 33 | 0.10 |
| 18 | 0.10 |
| 21 | 0.13 |
| 21 | 0.14 |

* The remaining 32 units reported values of the UIN that range from 0.17 to 0.43.

## 4.4 Conclusions

In this chapter I focus on nonresponse at the specific study stage of the life of *Opinione.net*. The overall aim of this part of the study is to assess the characteristics of the response/nonresponse process and the occurrence of unit and item nonresponse.

In particular, the first research aim is to explore the reasons for and the magnitude of nonresponse behaviour in the ILC survey. The cooperation rate is 52.7% and the incidence of break-offs is 4.6%, while the refusal rate is 38.7%. As already stated, it is difficult to compare my results to those of other studies, because of i) the lack of information on the response/nonresponse process and metrics reported by papers, ii) the unclear terminology sometimes used by scholars, and iii) the peculiarity of the various study designs documented by the literature. Nonetheless, if we consider the decline of "response rates" for non-probability panels (from 20% to 1%) pointed out by Tourangeau and colleagues (2013), I can conclude that there was a high response rate to the ILC survey. This is good news, as the incidence of the survey response might contribute to produce distortions in the estimates.

There are different reasons that could explain my findings. In particular, the high cooperation rate may be due to different aspects. First of all, the offline *Opinione.net* panel recruitment strategies motivate people to participate, and the panel maintenance activities (i.e. panelists' assistance through telephone helplines, and online social media, and "cleaning" strategies that unsubscribe the panelists who have been economically inactive for a year) increase the share of dedicated members. Then, the high "incidence rate" (i.e. the percentage of respondents selected for a survey that match the target group of the survey) of the ILC survey, because it was conducted on the general population, and not on specific subgroups. Lastly, the reward (amounting to 0.40 euros) offered upon completion of the six-minute ILC questionnaire may have boosted survey cooperation.

The second aim of this work is to assess the (unit) nonresponse bias. As a general approach I used the ‘rich sampling frame’. The strength of this design is that identical variables are available for both respondents and nonrespondents. Thus, using a unique dataset on all *Opinione.net* panelists, I was able to construct accurate estimates of nonresponse bias for the frame variables (Groves, 2006). Looking at the results, this part of the work presents encouraging findings. Bivariate analysis and measures of nonresponse documented small differences between respondents and nonrespondents. In particular, they showed that respondents slightly over-represent those living in the South of Italy and in the Italian islands, individuals belonging to the middle-age groups, and unemployed people, whereas respondents slightly under-represent those living in the North East of Italy, economically inactive people, and the less- or uneducated individuals. Interestingly, results from the regression analysis show that the net effect of age group and of the “unemployed” category of occupation disappear, revealing spurious relationships between age and response, and between occupation and response. Moreover, if compared to people living in the North West, those from the South of Italy have half the probability of responding to the ILC survey. These results suggest that the area of residence is the main characteristic that influences survey response. Thus, evidence from all the analyses on nonresponse bias prove that the risk of significant distortions introduced by the response propensity is limited to the geographic area of residence. Regarding the other socio-demographic characteristics, nonresponse at the study specific stage is completely at random. This is good news because this mechanism removes the risk of systematic distortions in the estimates from the ILC respondents sample.

Lastly, as pointed out in the literature review, the results from my analysis on item nonresponse confirm that skipping questions is not a common practice when participating in a panel survey. I speculate that the reasons previously discussed for high cooperation rate can explain this result too. In particular, the recruitment strategies, and the panel maintenance activities may select the most dedicated panelists, who are also more prone to provide accurate survey answers. Reassuring findings stand out from my analysis on item nonresponse. Indeed, the values of the item nonresponse rate (INR), and of the unit-level item nonresponse (UIN) are very low. Thus, missing answers are negligible when looking at both the question level and the respondent level.

Although the INR is low overall, I found that some items concerning use of the Internet in specific places, and consumption of alcoholic drinks report slightly higher values. My hypothesis is that some respondents may have chosen to give no answer rather than ticking "never" when answering questions about places where they use the Internet because they do not go there. On the other hand, I speculate that missing data on the "number of glasses of alcoholic drinks" can be an indicator of the unwillingness to report a socially undesirable response, i.e. an answer that will be viewed unfavourably by others (I will discuss this phenomenon in Chapter 5).

In addition, the results from the analysis on UIN showed that the majority of the respondents (i.e. 1,041 units) reported no missing answers. In this respect, the most 'serious' ILC respondents are young individuals, employed people, and those with secondary education.

This part of the study has one main limitation. The 'rich sampling frame' design, used as an approach for assessing nonresponse bias, has a weakness. The variables available for the nonrespondents (belonging to the *Opinione.net* panel) are, by definition, not all those of interest to the survey and are limited to socio-demographic characteristics. As a consequence, I was only able to use these variables to check for the differences between respondents and nonrespondents. On the other hand, it would have been interesting to compare these two groups on a wider range of variables, that include, for example, hobbies, interests, opinions, and behaviours. Results from this further comparison might be useful to go through with an in depth analysis of the nonresponse bias.

In conclusion, when focussing on sample representativeness (see Chapter 2) and nonresponse bias (discussed in this chapter), the findings implicitly suggest that the ILC survey estimates suffer from composition bias, but does not suffer from nonresponse bias. In the next chapter I address another type of distortion that may affect the ILC survey estimates - measurement bias.

# 5

# Measurement error in non-probability online panels

In Chapters 3 and 4 I focused on two sources of non-sampling error - undercoverage and nonresponse. In this chapter, I address another source of non-sampling error - measurement error. More specifically, I look at the study stage and assess the occurrence, nature, and magnitude of measurement error in the ILC survey estimates. Following the outline adopted for the previous two chapters, I review the literature focused on this type of error in non-probability online panels and state my specific research aims (section 5.1). I describe the research design of this part of my work (section 5.2), and I present the results of my analyses (section 5.3). Section 5.4 concludes the chapter.

## 5.1 Literature review and research aims

To frame the discussion on measurement error in non-probability online panel surveys, I draw on the AAPOR "Report on Online Panels" (Baker *et al.*, 2010a). In this report, the authors discuss the potential causes of measurement error that may occur when collecting data from (non-probability) online panel samples (i.e. the questionnaire design, mode of administration, and the respondents). I structure my literature review around the potential causes of measurement error in non-probability online panel surveys.

### 5.1.1 Questionnaire design

The first potential cause of measurement error is the questionnaire design. With the onset of the Internet, the design of web questionnaires has introduced new challenges and potential problems. The literature specific to the implications of the web for survey design has demonstrated a wide range of response effects due to questionnaire and presentation features in web surveys (Couper, 2008). Recent studies carried out on opt-in panel samples focused on i) questionnaire structure, i.e. question order (Jones *et al.*, 2015), and length of the questionnaire (Drewes, 2014), ii) question types, i.e. question-response format (Buskirk and

Andrus, 2014; Drewes, 2014; Revilla *et al.*, 2015), and iii) presentation effects, i.e. graphic layout (Arn *et al.*, 2015; Brown *et al.*, 2012; Couper *et al.*, 2013). Results from these studies on non-probability online panel samples show that i) question order (Jones *et al.*, 2015), and length of the questionnaire (Drewes, 2014) have no impact on various data quality metrics, and ii) the response quality to different question-response formats is influenced neither by the survey mode (i.e. mobile vs PC) (Buskirk and Andrus, 2014; Drewes, 2014) nor by the type of sample (i.e. opt-in panel sample vs probability-based survey sample) (Revilla *et al.*, 2015). Moreover, with regard to the presentation effects, i) the use of an adapted design reduces the straightlining, and increases the willingness to answer open-ended questions (Arn *et al.*, 2015), ii) the mapping effort (i.e. putting markers in the right spot on a geographic map) results in lower data quality (Brown *et al.*, 2012), and iii) in the design of grids (i.e. matrix questions), using dynamic feedback to guide respondents through a multiquestion grid, and splitting the grids into component questions reduce missing data (Couper *et al.*, 2013).

### 5.1.2 Mode of administration

The second potential cause of measurement error is the mode of administration, thus producing mode effects. When dealing with opt-in panels, these effects may be connected to both the shift from interviewer administration to self-administration by computer, and the shift from interviewer administration with probability samples to computer self-completion with non-probability samples (Baker *et al.*, 2010a).

Much of the research on mode effects compared results from the two methods (i.e. non-probability online panels and face-to-face probability surveys) and simply noted differences but without looking specifically at the issue of accuracy, i.e. without comparing survey estimates to benchmark data (Baker *et al.*, 2010a). For example, in the Public participation geographic information systems (PPGIS) surveys, the answers provided by a sample from an opt-in panel were compared to those of a random household sample, and responses from online panelists indicated less mapping effort, resulting in lower data quality, than that obtained from random household respondents (Brown *et al.*, 2012).

Another common technique for evaluating the accuracy of results from non-probability online panels and face-to-face probability surveys has been to compare results with external benchmarks established through official statistics (such as Census data, or election outcomes), or through probability-based and high quality survey estimates (Baker *et al.*, 2010a).

Inconclusive findings stand out from the literature focused on non-probability online panels. In Chang and Krosnick's (2009) study the non-probability sample of panelists resulted to be the most affected by composition bias but yielded the most accurate self-reports, when compared to the probability samples (i.e. RDD telephone and the Internet probability sample). Yeager and colleagues (2011) found that their probability sample surveys (whether by telephone or web) were consistently more accurate than their nonprobability sample surveys. Comparing an opt-in panel sample with a probability-based face-to-face sample of car drivers, Goldenbeld and de Craen (2013) found that online panel respondents were consistently less accurate in reporting ratings (i.e. they had larger percentages midscale ratings), and were more positive and less outspoken in their disagreement on three sets of questions than the probability-based face-to-face survey respondents. Similarly, in a study on sexual behaviours and attitudes set up by Erens and colleagues (2014), respondents from four non-probability web panel surveys showed more neutral answers than the probability-based face-to-face respondents. Revilla and colleagues (2015) found that there is no difference between the quality estimates about satisfaction and trust in institutions obtained from a non-probability online panel sample and a face-to-face sample. Results from an Australian study (Pennay *et al.*, 2018) showed that the five non-probability panel surveys were more biased on the substantive measures and had more variance from the benchmark values than the three probability surveys.

More extensively, I add to the "Mode of administration" group a specific type of mode effect, i.e. the device used to fill out the web questionnaire in non-probability online panel surveys. Studies investigating this potential cause of measurement error mainly focus on experimental designs that explore the impact of different survey modes - in particular, the mobile compared to the PC web completion - on data quality. Findings from these studies are inconclusive. Mavletova (2013) found that mobile web was associated with a shorter length of open answers, similar level of socially undesirable and non-substantive responses, and no stronger primacy effects. In another study, Mavletova and Couper (2013) found no differences between smarthphone and PC respondents, except for the reporting of alcohol consumption. When comparing the mobile group with the PC group, iPhone respondents showed shorter completion times, and had values on the upper range of the characters typed distribution for three open-ended questions (Buskirk and Andrus, 2014).

When looking at the specific indicators of mode effect, satisficing (i.e. shortcutting the response process) and social desirability hypothesis (i.e. the tendency of respondents to provide answers that will give a favorable image of themselves) are the most common behaviours reported by the studies on opt-in panel surveys (Erens *et al.*, 2014; Goldenbeld and de Craen, 2013; Mavletova, 2013).

### 5.1.3 Respondents

The last cause of measurement error is the survey respondent. Indeed, respondents vary from each other in terms of their cognitive capabilities, motivations to participate, panel-specific experiences, and topic interest and experience.

Recent literature on the respondent-level factors that influence measurement error in non-probability online panels focused on i) the impact of incentives, and the respondents' behaviour (i.e. professional respondents, fraudulent respondents, and speeders[25]) as motivations to participate, and ii) the topic interest (i.e. interest in the survey topic, and topic salience).

The main findings of two studies on the impact of lotteries as incentives to take part in surveys showed that the only type of lottery (the others are splitting the lotteries, gifts, and lottery of unknown or known expected value) that significantly increases response quality is the cash lottery (Göritz and Luthe, 2013a and 2013b).

As documented by the results of a comparison between nineteen opt-in panel samples, professional respondents were not a threat to data quality (Matthijsse *et al.*, 2015).

"Fraudulent respondents" were more represented in a sample from Amazon's Mechanical Turk[26] than in a non-probability online panel sample (Smith *et al.*, 2016). In contrast to this result, in two validation studies, respondents who failed verification of their identities were nearly three times as likely to fail at least one quality check (Baker *et al.*, 2014), and some

[25] "Professional respondents" are respondents who join multiple online access panels and regularly completes many surveys (Jones *et al.*, 2015). "Fraudulent respondents" are individuals who assume false identities or simply misrepresent their qualifications either at the time of panel registration or in the qualifying questions of individual surveys (Baker *et al.*, 2010a). "Speeders" are respondents who answer more quickly than would be expected given the nature of the questions and responses (Baker *et al.*, 2010a).

[26] Mechanical Turk is a crowdsourcing Internet marketplace on Amazon's website, where employers or "requesters" post tasks, called HITs (Human Intelligence Tasks), to recruit anonymous "workers" in exchange for a small monetary wage (Smith *et al.*, 2016).

fraudulent respondents seeking incentives tried to maximize chances of entering the survey (Jones *et al.*, 2015).

The latter work by Jones and colleagues (2015) also showed that “speeding” is highly unlikely to occur. Other studies come to similar conclusions on three different samples: respondents from the German Longitudinal Election Study (Greszki *et al.*, 2014), a sample of respondents highly interested in the survey topic from an Austrian opt-in panel (Keusch, 2013), and a sample of US online panelists (Smith *et al.*, 2016).

Lastly, Keusch (2013) focused on the role of topic interest and topic salience in non-probability online panels, and found that respondents highly interested in the survey topic showed less satisficing behaviour than those with low personal interest, whereas the level of topic salience in the e-mail invitation had no influence on data quality in the online panel.

To sum up, studies on measurement error in non-probability online panel surveys investigated a wide variety of effects. After the publication of the AAPOR “Report on Online Panels” in 2010, only a few studies focussing in particular on mode effect were carried out. More specifically, the literature documented few attempts to disentangle a special type of mode effect, that is survey-mode effect within a unique opt-in panel survey, when addressing the measurement error issue. On the other hand, studies compared the quality of the estimates obtained from opt-in panel samples with those from probability samples. Nonetheless, to the best of my knowledge, there are no attempts to specifically disentangle the mode effect between non-probability panel surveys and probability-based high quality surveys (or official statistics). Moreover, inconclusive findings on data quality from non-probability panel-survey samples stand out.

Against this background, my work aims to assess the ILC survey data quality, focussing on the second potential cause of measurement error, i.e. the mode of administration.

First I investigate the occurrence and magnitude of measurement bias in the estimates from behavioural variables.

Moreover, I look at the occurrence of less-than-optimal strategies used by the respondents to get through the ILC survey, as they may yield mode effects. With this part of the analysis I assess whether my findings from the ILC respondents are consistent with those from the literature and, thus, whether this method may be more accurate than face-to-face probability surveys in collecting data on sensitive topics, also in the specific case of non-probability panel surveys.

In addition, I explore the occurrence and the impact of the survey-mode effect on measurement bias in the estimates from behavioural variables. Last, I assess the impact of weighting on measurement bias.

I believe findings from this part of the study contribute to filling a notable gap, especially in the literature on mode effects in non-probability online panel surveys. The main motivations are that i) I am aware of only one study by Revilla and colleagues (2015) focused on a Southern European country (i.e. Spain), and no studies in Italy in particular, ii) scholars obtained neither conclusive nor consistent findings from their studies on measurement error, and iii) to the best of my knowlegde, except for the special case of pre-election polls (e.g. Breton *et al.*, 2017; Malhotra and Krosnick, 2007), there are no recent studies that specifically disentangles survey-mode effect in non-probability online panels[27].

## 5.2 Research design

### 5.2.1 Data

For this part of the study, I used a diverse set of data, i.e. the 2015 Multipurpose Survey - Aspects of Everyday Living (AEL), considered as the gold standard in my analysis, and data from the Italians' Living Conditions (ILC) survey, a web survey that was conducted on a quota sample of the *Opinione.net* panel members (see Chapter 2 for details).

### 5.2.2 Methods

To reach my aims, I performed a bivariate analysis and computed the data quality metrics that I have already used for the analysis on sample composition (see Chapter 3). I also carried out a bivariate analysis to explore the impact of various indicators of measurement error on the measurement bias in the ILC survey estimates. I applied quasirandomization weighting (see Chapter 2) to remove the measurement bias. Moreover, I ran a number of logistic regression models to assess the impact of survey mode on the outcome variables. Table 1 summarizes the variables used for each type of analysis.

[27] Studies on measurement error mainly focused on mode preference (Vandenplas *et al.*, 2016; Vannieuwenhuyze *et al.*, 2010) or measurement equivalence (Hox *et al.*, 2015) in mixed mode surveys or in experimental designs (Bosch *et al.*, 2018) that involve probability samples, and this is not my case.

Table 1. Measurement bias in the ILC estimates: methods of analysis and variables.

| Methods of analysis | Variables |
|---|---|
| Bivariate analysis – accuracy metrics, and logistic regression | Watching TV |
| | Socio-political participation (6 items) |
| | Internet use in the last 12 months |
| | Internet use for various activities (5 items) |
| Bivariate analysis – indicator of measurement error (straightlining) | Internet use in different places (5 items) |
| | Consumption of a variety of foods (4 items) |
| Bivariate analysis – indicator of measurement error (social desirability bias), and logistic regression | Alcohol consumption (outside mealtimes) |
| | Number of glasses of alcoholic drinks (in a week)* |
| | Tobacco consumption |
| Bivariate analysis – indicator of measurement error (number of answers in check-all-that-apply questions) | Environmental problems I'm worried about |

* The logistic regression model was not run for the number of glasses of alcoholic drinks.

*Data quality metrics*

Similarly to the analysis that I performed on sample composition bias (see the Methods section in Chapter 3 for details), I carried out a bivariate analysis, performing the Chi Square tests, to compute the accuracy metrics, i.e. the average absolute error, the number of significant differences from the benchmark, the percentage point error, the largest absolute error, and the number of absolute differences greater than a given threshold, that I set to 5, 15, 25, and 30 percentage points. In addition, I applied quasirandomization weighting to remove the measurement bias (see the Methods of analysis section in Chapter 2).

*Indicators of measurement error*

I performed a bivariate analysis on other variables (see Table 1 above), also computing the Chi Square tests, to explore the occurrence of less-than-optimal strategies used by the respondents to get through the ILC survey. I detected these respondents' behaviours adopting a number of indicators of measurement error, i.e. straightlining, social desirability bias, and the number of answers in check-all-that-apply questions. I also applied quasirandomization weighting to remove the share of bias due to satisficing and social desirability hypothesis (Goldenbeld and de Craen, 2013).

"Satisficing" refers to a failure to put in the necessary effort to optimally answer a survey question (Krosnick, 1991). The general hypothesis of a web survey inducing more satisficing than a face-to-face survey leads to the expectation that web survey respondents, for example, will differentiate (i.e. use a limited number of the available response alternatives) less on rating scales than face-to-face respondents (Holbrook, Green, and Krosnick, 2003). This phenomenon is also called "straightlining". To define straightlining - classified under the label of strong satisficing (Krosnick, 1991) - I used the non-differentiation indicator (i.e. an indicator that measures the variability of answers provided by the respondent to a grid question) proposed by Roßmann and colleagues (2018), which takes the value "1" if all the responses, excluding missing values, have an identical value when answering all the items of the grid question; the indicator equals "0" if at least one response has another value. I considered straightlining on two grid questions. The question about the "Internet use in different places" is constituted of five items, respectively, and adopts a 6-point scale. The question about "food consumption" is constituted of four items, and adopts a 5-point scale (see Appendix 11).
Another example of satisficing behaviour occurs when the respondent ticks a limited number of options in questions that ask to check all the applied alternatives (usually setting a maximum number of responses). In the ILC questionnaire there is one question of this type (also called "check-all-that-apply question") about the environmental problems; the indicator is computed counting the number of problems selected by the respondents.

"Social desirability" refers to the tendency of respondents to provide researchers with responses that will give a favorable image of themselves, or responses they think correspond to the social norm (Frippiat and Marquis, 2010). There is consistent evidence that socially-desirable responding is more likely with interviewer-administered modes of data collection than self-administered modes (Baker *et al.*, 2010a; Heerwegh, 2009; Joinson, 1999). In my analysis I explored the occurrence of this phenomenon looking at the questions about alcohol consumption (in particular, focussing on the "never" answer and the mean number of glasses), and tobacco consumption (in particular, focussing on the "I've never smoked" option).

*Survey-mode effect*

The accuracy metrics, and the indicators of measurement error are methods used to provide evidence on the overall magnitude and nature of the measurement error in the ILC survey estimates. A further method of analysis aims to explore the contribution of a specific source of error (i.e. survey-mode effect) on measurement bias in non-probability online panel estimates. Therefore, I performed a multivariate analysis to assess the impact of survey mode on both the outcome variables previously used to compute the accuracy metrics (i.e. watching TV, Internet use, Internet use for communication activities, and socio-political participation), and to detect the occurrence of social desirability bias (i.e. alcohol and tobacco consumption).

To address the impact of survey mode on the estimates from the ILC survey, and the AEL survey, I adopted a new method of analysis, partially drawing on Mavletova and Couper (2013). These authors implemented an experimental design on a sample of volunteer panelists, where they defined a model including the survey device (i.e. PC vs mobile) as explanatory variable. In my analysis, to disentangle the specific cause of measurement error due to the mode of administration, I estimated the part of measurement bias in the respondents' estimates that can be explained by the net effect of web vs face-to-face mode. In particular, I ran six logistic regression models, one for each outcome variable[28], where the explanatory variable is survey mode (i.e. web vs face-to-face), and the control variables are socio-demographics (i.e. sex, age group, area of residence, education, occupation, and marital status). The models have the following form:

(1) $logit(\pi_i) = \beta_{00} + \beta_1 Mode_{ij} + \sum_{j=2}^{J} \beta_j\, SocioDemo_{ij}$

Note that $\pi_i$ is the probability to respond 'yes' to each measure (e.g., Internet use), $\beta_{00}$ is the reference category of each variable included into the model, $\beta_1$ is the effect of the survey mode (explanatory variable), and $\beta_j$ are the effects of the socio-demographic characteristics (control variables).

## 5.3 Results

In the following sections, I look separately at the results of my analyses on i) the occurrence and magnitude of measurement bias (i.e. the accuracy of the responses), ii) the occurrence of

[28] For the purposes of the logistic regression, I recoded each outcome variable into a dichotomous variable. In Appendix 2 are shown the coding procedures.

less-than-optimal strategies used by the respondents to get through the ILC survey, and iii) the impact of survey mode on measurement bias.

### 5.3.1 The accuracy of the responses

Results from my analyses are shown in Table 2. Panel 1 reports percentage differences; positive and negative values indicate that the corresponding categories are over or under-represented, respectively. For the bivariate analysis in Panel 1 I also computed the Chi Square tests and I found that all the differences between the variables are statistically significant at the level of 0.000. Panel 2 shows the values of the other accuracy metrics. The variable distributions are shown in Appendix 13.

Table 2. Occurrence and magnitude of measurement bias. Percentage differences and Accuracy metrics.

| | ILC sample compared to general population | | |
|---|---|---|---|
| | No weights | Weights | \|No weights - Weights\| |
| **Panel 1. Percentage differences** | | | |
| Variables | | | |
| *Watching TV* | | | |
| I do not watch TV | -0.8 | -0.2 | 0.6 |
| I watch TV every day | -7.0 | -5.0 | 2.0 |
| I watch TV a few days a week | 7.8 | 5.2 | 2.6 |
| *Socio-political participation in the last 12 months ('yes' answers)* | | | |
| I attended a political meeting | 16.0 | 13.7 | 2.3 |
| I took part in a political demonstration | 14.8 | 15.2 | 0.4 |
| I listened to a political debate | 32.9 | 23.4 | 9.5 |
| I gave money to a political party | 11.2 | 11.0 | 0.2 |
| I gave money to an association | 30.1 | 23.1 | 7.0 |
| I did voluntary work | 27.4 | 24.3 | 3.1 |

Table 2. *Continued*.

| | | | |
|---|---:|---:|---:|
| *Internet use in the last 12 months* | | | |
| Every day | 28.4 | 25.2 | 3.2 |
| A few times a week | -22.0 | -19.6 | 2.4 |
| Once a week | -1.7 | -0.9 | 0.8 |
| A few times a month (less than 4) | -3.5 | -3.6 | 0.1 |
| Less than once a month | -1.2 | -1.2 | 0.0 |
| *Internet use for various activities in the last 3 months ('yes' answers)* | | | |
| Send or receive e-mails | 16.2 | 12.0 | 4.2 |
| Telephone over the Internet / video calls | 21.7 | 12.7 | 9.0 |
| Post messages to chat sites/blog/forum | 26.8 | 22.4 | 4.4 |
| Use instant messaging | 23.4 | 15.2 | 8.2 |
| Participate in social network | 23.1 | 20.9 | 2.2 |
| **Panel 2. Other metrics** | | | |
| Average absolute error | 21.5 | 17.2 | 4.3 |
| Number of significant differences from benchmark | 13 | 13 | 0 |
| Largest absolute error | 32.9 | 25.2 | 7.7 |
| Number of absolute differences greater than | | | |
| 5 | 4 | 5 | 1 |
| 15 | 6 | 8 | 2 |
| 25 | 3 | 1 | 2 |
| 30 | 2 | 0 | 2 |

Results from my analysis show that for the behavioural variables considered the ILC respondents report substantially different data from the benchmarks. Indeed, the average absolute error is 21.5, all the differences are statistically significant, and the largest absolute error is 32.9. The variable distributions of the two populations (i.e. the ILC sample, and the general population) are all (statistically significant) different and the magnitude of the bias is large, being eleven of the differences greater than 15 percentage points (of which five are greater than 25). In particular, the ILC responding sample over-represents people who watch TV some days, participate in socio-political activities, use the Internet every day, and use the

Internet for various communication activities. However, there are some exceptions, i.e. people who do not watch TV, and those who use the Internet once a week or less often, where the percentage point differences are smaller than 5.

Quasirandomization weights have a positive impact on data quality, overall reducing the magnitude of the measurement bias; for example, the average absolute error decreases from 21.5 to 17.2, the largest absolute error decreases from 32.9 to 25.2, and there is only one difference higher than 25 or 30 percentage points. Unfortunately, the distortions in the estimates remain high. This finding from the ILC sample and the general population might be due to the combined effect of two sources of error. First, the composition bias, i.e. the socio-demographic differences in the sample composition that I found from the analysis on nonresponse at the recruitment stage. In addition, the measurement bias that may occur when the respondent provides an answer to a given question, that for some reasons do not correspond to the "true" population value.

### 5.3.2 The impact of satisficing and social desirability on measurement bias

In this section, I present the findings from the analysis on a specific threat to data quality, i.e. the occurrence of satisficing and social desirability response when answering particular types of questions (i.e. Internet use in different places, food/alcohol/tobacco consumption, number of glasses of alcoholic drinks, and environmental problems).

Results from my analysis are shown in Table 3. Contrasting findings stand out as regards satisficing behaviours. On the one hand, straightlining seems to be quite a serious threat to data quality. Indeed, the ILC respondents report higher values of non-differentiation than the general population on the items about both Internet use in different places (+9.7 percentage points) and food consumption (+7.3 percentage points). Moreover, the differences between these two populations (i.e. ILC respondents and AEL respondents) are statistically significant. On the other hand, when looking at the number of answers provided to the question on environmental problems, satisficing does not occur. Indeed, the mean difference between the ILC respondents and the general population is negligible, at just 0.3 items.

Social desirability bias is large, especially in the estimates of alcohol consumption, and occurs in the direction of less socially desirable responses in the ILC sample. In particular, people who claimed that they never drink and that they have never smoked are much more represented in the general population than in the ILC sample. The mean difference in the number of glasses of alcoholic drinks is not high (equal to 0.5 glasses), but its value is statistically significant.

Table 3 shows that, similar to the impact on the accuracy of the responses, quasirandomization weighting is effective in reducing, but not in completely removing, the magnitude of the bias for each variable included in the analysis.

Table 3. Impact of satisficing and social desirability bias on data quality. Percentage differences and differences in means.

| Variables | ILC sample compared to general population | | |
|---|---|---|---|
| | No weights | Weights | \|No weights - Weights\| |
| *Satisficing - straightlining* | | | |
| Internet use in different places (%) | 9.7 | 7.2 | 2.5 |
| Consumption of a variety of foods (%) | 7.3 | 4.5 | 2.8 |
| *Satisficing - number of answers to check-all-that-apply questions* | | | |
| Environmental problems I'm worried about (mean) | 0.3 | 0.1 | 0.2 |
| *Social desirability* | | | |
| Alcohol consumption (% 'Never') | -40.5 | -28.6 | 11.9 |
| Number of glasses of alcoholic drinks (mean) | 0.5 | 0.8 | 0.3 |
| Tobacco consumption (% 'Never') | -11.6 | -9.2 | 2.4 |

Note: all differences are statistically significant at the level of 0.000.

### 5.3.3 The impact of survey-mode effect on measurement bias

Table 4 reports findings from the multivariate analysis that estimates the impact of survey-mode effect on measurement bias. The net effect of survey mode on different survey outcomes (adjusted for the effect of the socio-demographic variables included as control variables in the model) is high and statistically significant. In particular, the ILC respondents (web mode) have a higher probability of watching TV, participating in socio-political activities, using the Internet every day (in the last twelve months), using the Internet for at least three communication activities (in the last three months), drinking alcohol, and smoking than the AEL respondents (face-to-face mode). The variables about Internet use (in the last 12 months, and for various communication activities) are those on which the survey mode has the largest net effects.

Web respondents compared to face-to-face respondents are more likely to be "heavy Internet users". This is a common and obvious result, if we consider that the ILC sample is drawn from a panel that was built and is active online.

Findings from the logistic regression models about drinking alcohol (model 5) and smoking (model 6) behaviours confirm those from the bivariate analysis. In particular, for drinking behaviour, the effect of the mode of administration on the probability of consuming alcohol outside mealtimes is strong (odds ratio=5.656) and significant (p≤0.001). The web respondents are 5.6 times more likely than the face-to-face respondents to drink without food. In addition, for smoking habits, the survey mode is a relevant explanatory variable, but its net effect is smaller (odds ratio=1.422).

Table 4. Logistic regression model results (odds ratios) for the mode of administration (ILC-web compared to AEL-F2F).

| Outcome variables | Survey mode (AEL)[a] |
|---|---|
| Watching TV (model 1) | 1.429*** |
| Socio-political participation (model 2) | 4.161*** |
| Internet use in the last 12 months (model 3) | 15.134*** |
| Internet use for various activities (model 4) | 6.075*** |
| Drinking (model 5) | 5.656*** |
| Smoking (model 6) | 1.422*** |

Control variables: sex, age group, area of residence, education, occupation, and marital status.

[a] Category of reference. ***p≤0.001

## 5.4 Conclusions

In this chapter I address measurement error taking three approaches, i.e. overall magnitude of measurement bias, indicators of measurement error, and specific impact of survey mode on measurement bias.

Results from the first approach showed that, in general, the magnitude of measurement bias in the ILC estimates obtained from questions about behaviours is high. The only two exceptions are people who do not watch TV, and those who use the Internet once a week or less often, for whom I found small differences between the ILC respondents and the general population. With respect to these two groups of individuals (i.e. those who do not watch TV, and those who rarely use Internet) I can conclude that my survey sample is representative of the general population. When comparing the bias in the estimates from the analysis on the sample composition (see Chapter 3) with those from the analysis on measurement error (in this chapter), I find that the differences in socio-demographic characteristics are smaller than those in answers to behavioural questions. This means that the ILC respondents are more similar to the Italian population for their socio-demographic characteristics than for their reported behaviours. In particular, the self-selection mechanism (first, into the panel, and then, into the ILC survey) and the measurement process yielded a sample that mainly excluded subgroups of people who behave differently from the general population with respect to Internet use, and socio-political participation.

Unfortunately, applying quasirandomization weighting, albeit effective in reducing the distortions, does not remove bias, and the gap between the estimates on substantive questions from the ILC survey and those from the AEL survey still remains wide. However, the impact of weighting on sample composition bias (see Chapter 3) proves to be more effective. To explain the differential effect of the adjustment strategy, I suggest that the auxiliary variables (i.e. socio-demographics) used to compute the weights do not describe all the differences between the ILC respondents' sample and the general population, and some confounding variables are excluded. As a consequence, quasirandomization weighting is more effective in reducing the bias in the socio-demographic estimates than the bias in the estimates on Internet use, and socio-political participation.

At this level of analysis, the differences in the reported behaviours could be only partially due to measurement error. Indeed, as in Chapter 3 I documented the occurrence of composition bias in the ILC sample, it is not clear which is the contribution of the differences, on one hand, in sample composition and, on the other hand, in the measurement process.

When looking at the results from the analysis on a number of indicators of measurement error, two findings stand out. First, the satisficing hypothesis, documented in the literature, is partially confirmed. Indeed, straightlining is quite a serious threat to ILC data quality, whereas the ILC respondents seem to be as accurate as the AEL respondents in ticking multiple options in check-all-that-apply questions. In addition, the socially desirable response hypothesis is completely validated, as the web survey respondents (i.e. the ILC respondents) are more likely to self-report undesirable responses to sensitive questions than face-to-face respondents (i.e. the AEL respondents). The large differences in alcohol and tobacco consumption might suggest that, when studying sensitive topics, opt-in panels can provide a more accurate data collection method than face-to-face interviews. However, this may be true if other conditions occur (e.g. low incidence of other sources of non-sampling error, such as sample composition and nonresponse bias).

Lastly, for the purposes of the discussion on measurement bias, I focus on the role of mode effect in explaining differences in the estimates from behavioural variables and socially desirable responses. In particular, results from the multivariate analysis on drinking behaviour showed that the mode of administration has a high and significant net effect (adjusted for the effects of the socio-demographic variables, that are smaller) on measurement bias in estimating alcohol consumption. Therefore, I speculate that the distortion in this specific ILC estimate is mainly due to measurement bias. On the other hand, the net effect of survey mode results to be smaller on smoking habits than on drinking behaviour. Thus, regarding tobacco consumption, I speculate that sample composition bias could be the main source of the distortion.

This part of my study has one main limitation. When focussing on measurement bias, I was only partially able to correct for possible sample differences that may have affected results. Therefore, it was not possible to completely separate effects of sampling (face-to-face probability sample vs non-probability online panel sample), and instrument (face-to-face interview vs self-completed online questionnaire) (Goldenbeld and de Craen, 2013).

# Conclusions

My thesis focuses on a relatively new method of data collection in social survey research i.e. non-probability online panels.  As a means of addressing the declining response rates and rising costs, some survey organizations have recruited panels made up of volunteer respondents, who participate in surveys in exchange for some sort of remuneration.

The overall aim of the thesis is to assess the quality of non-probability online-panel survey data in Italy. In particular, I explored the impact of undercoverage and nonresponse on sample selectivity, and the impact of nonresponse and measurement error on the quality of the Italians' Living Conditions (ILC) survey data collected on a sample from the *Opinione.net* opt-in panel[29]. Moreover, I assessed the impact of propensity score weighting in reducing the composition bias, and improving the quality of the ILC survey data.

I used three different datasets: 1) a unique dataset on the panelists of the non-probability online panel *Opinione.net* (profile data are usually not available for researchers), 2) a dataset on the responding sample from the ILC web survey conducted on the *Opinione.net* members, and 3) a dataset on a sample of the Italian population from a probability-based reference survey (i.e. the Multipurpose survey Aspects of Everyday Living - AEL), with no coverage error and high response rates, conducted every year by ISTAT and often used as the gold standard. In addition, unlike the common practice with panel companies, the research institute that manages the *Opinione.net* panel provided me with the ILC survey paradata (i.e. the final disposition codes) that I used for the analysis on nonresponse. To address my research aims, I adopted a variety of methods (i.e. bivariate analysis, logistic regression models, accuracy metrics, and indicators of nonresponse and measurement error), and I also applied a particular type of statistical adjustment (i.e. quasirandomization weighting), to improve the survey estimates.

Several key findings stand out from my analyses. Undercoverage may be a serious source of bias, because the Internet population is not representative of the general population, as it over-represents single people, educated individuals, and those in employment, and under-

[29] I chose to use the *Opinione.net* panel because it is a partner of the well-known international online panel Consumer Insights Network (CINT) and meets the industry-standard 28 ESOMAR questions.

represents people aged 65 and over, those with no education, and economically inactive individuals. Similarly, nonresponse at the recruitment stage[30] may produce biased estimates, because the ILC responding sample, compared to the general population, over-represents men, single people, educated individuals, and those who are in employment, and it under-represents divorced/widowed individuals, people aged 65 and over, and economically inactive sample members. However, weighting is a very effective method to reduce the magnitude of the bias in the ILC estimates. In particular, after weighting, the ILC sample is more representative of the general population than the Internet population. On the contrary, nonresponse at the specific study stage seems not to yield substantial distortions: comparing the socio-demographic characteristics of respondents with those of nonrespondents, these two groups significantly differ only for the geographic area where they live. My research has also shown that measurement error may bias the survey estimates. Indeed, i) the estimates on substantive questions are even more distorted than those on socio-demographic variables (as it resulted from the analysis on sample composition), ii) I found evidence of social desirability bias, and iii) the net effect of the mode of administration on behavioural variables is high and significant. However, similar to the results from the analysis on sample composition, the quasirandomization weighting proves effective in reducing, but not removing bias, which still remains substantial.

All in all, my findings show that using a non-probability online panel as a method to survey the general population in Italy is a challenging task. The accuracy of the estimates obtained from a panel sample, such as the ILC respondents, depends on a number of factors: Internet coverage, self-selection into the panel and into the study, and the quality of the answers to the questionnaire. Because of the self-selection process, some groups of people may mirror the socio-demographic characteristics of the Italian population (e.g. individuals aged 18-24, and unemployed people), whereas some others (e.g. single people, and those with a primary school education) may not. Moreover, for some substantive questions (e.g. watching TV) survey estimates may be accurate, whereas for others (e.g. Internet use) they may not. Similarly, as expected, the weighting adjustment is more effective in reducing bias in the estimates on socio-demographic variables (as these variables were used to calculate the weights), than on behavioural questions.

[30] Recall that, when studying the impact of nonresponse on sample composition, I drew on the AAPOR (Baker *et al.*, 2010a) approach that looks at the different stages of the life of the panel and assesses the occurrence of nonresponse.

My study has four main limitations. First, because of economic constraints, I was not able to carry out the ILC survey on samples drawn from different opt-in panels. Comparative findings could have been useful to support or contradict results from the *Opinione.net* respondents, and draw robust conclusions on the appropriateness of using non-probability online panels as an effective method of data collection in Italy.

Second, data from the ILC survey were collected in 2017 and were compared to the gold standard data collected in 2015. It is quite common, in studies that compare opt-in panel survey data with a reference survey, to use benchmark data that were not collected simultaneously with panel survey data. The lag in the time frame of the panel survey data and the gold standard data might lead to inaccurate conclusions when assessing data quality. Unfortunately, only the 2015 AEL data were available. Nonetheless, as a check of robustness, I compared the distributions by age, sex, and area of residence of the 2015 AEL survey with those of the 2017 Italian administrative data. This comparison showed no statistically significant differences in these three variables.

Third, I explored a limited number of topics, and some of them (the most obvious example is Internet use) may be related to the fact that the panelists belong to the online population. Thus, focussing on other topics, the analysis on data quality in the ILC sample could produce different and (perhaps) less biased estimates.

Last, as largely documented in the literature on data quality in non-probability online panels, my study was unsuccessful in disentangling the different sources of bias, i.e. self-selection bias and measurement bias. However, as an attempt to overcome this limitation, when addressing measurement error I estimated the net effect of the mode of administration on answering substantive questions. Results from this analysis allowed me to explain the part of measurement bias that is specifically caused by mode effect. Nonetheless, as it is quite common in studies that compare opt-in panel survey data with a reference survey, also in my study it was not possible to completely separate effects of sampling (face-to-face probability sample vs non-probability online panel sample), and instrument (face-to-face interview vs self-completed online questionnaire) (Goldenbeld and de Craen, 2013).

My research is a first attempt to fill a gap in studies addressing data quality in non-probability online panels in Italy. However, further research is needed. In particular, future work may involve an extension of my study on *Opinione.net* panel. The research team with which I am involved has submitted a research proposal for the data collection involving web samples

from five Italian non-probability online panels. This further study will contribute to overcoming the first limitation of my thesis.
Moreover, in future work I intend to assess the quality of data collected using the CATI method. Simultaneously with the ILC web survey, I carried out a telephone survey on a sample of Italians drawn from the landline phone register. Comparing the accuracy of the estimates obtained with this additional survey to those from the web survey, I would draw conclusions on the method that performs better when studying the general population in Italy.

My thesis also has some implications. It is clear, from my study as well as from the literature on non-probability online panels, that the use of opt-in panels raises a number of challenging questions for survey methodology research. One is how to choose between probability and non-probability surveys. Rivers (2013) suggests that there is little practical difference between opting out of a probability sample and opting into a non-probability sample. Supporters of non-probability panels point out that modeling, used with these panel data, works well enough for sponsor needs, and as well or better than probability methods in some domains (e.g. election polling). When looking at probability surveys, they also note that while probability theory may underlie the design, the achieved samples in "probability" surveys are increasingly self-selected, as response rates continue to decline (Miller, 2017).

The dilemma between using probability surveys and non-probability surveys is particularly relevant when dealing with the inference concern. Thus, the most urgent question arising from the research on non-probability online panels is: how should the researchers use their results from panel surveys, if they are interested in studying the general population? As Baker et al. (2013) clearly stated, given the absence of a theoretical framework to support inference from non-probability samples, researchers have no basis from which to claim that survey results using samples from non-probability online panels are projectable to the general population (Baker *et al.*, 2013; Baker *et al.*, 2010b). Nonetheless, these statements are not intended to stall further research, but rather to encourage its practice (Langer, 2018). Especially for surveys making broad claims of representation, a careful analysis of the potential bias and its effect on estimates, is important (Schonlau and Couper, 2017).

With this respect, Mercer and colleagues (2017) identify three components that determine whether or not self-selection could lead to biased results: i) exchangeability (i.e. are all confounding variables known and measured for all sampled units?), ii) positivity (i.e. does the sample include all of the necessary kinds of units in the target population, or are certain groups with distinct characteristics missing?), and iii) composition (i.e. does the sample distribution match the target population with respect to the confounding variables, or can it be adjusted to match?). These three requirements may be disregarded at different stages of the life of an opt-in panel.

At the recruitment stage, panels face an immediate threat to the positivity requirement, because individuals who do not use the Internet cannot participate, and for the Internet-savvy population no sampling frame or method (i.e. no complete list of e-mail addresses of the general population) exists that permits direct selection and invitation of sample persons to join the panel (Schonlau and Couper, 2017). As my study confirms (see Chapter 5), the exclusion of non-Internet people may produce large differences in survey outcomes pertaining to technology use, for example. However, I speculate that middle-age and older people are likely to become, in a relatively short time, more technologically savvy so to reduce undercoverage bias. Thus, obtaining a diverse array of potential respondents is crucial to the success of any recruitment method (Mercer *et al.*, 2017). A way to improve the probability of meeting the positivity requirement is to recruit panel members from a diverse set of sources. In my case, to recruit their panelists Demetra adopts not only online but also offline methods, such as CATI interviews conducted drawing respondents from different sampling frames (i.e. the landline phone register, and RDD of mobile phone numbers). In addition, those potential panel members who are recruited offline and do not meet the technological requirements, should be endowed with the skills and tools (e.g., PCs, tablets, Internet connection, etc.) needed to become panelists. However, it may be difficult to know which characteristics distinguish between individuals recruited from different sources, and the recruitment process can yield a pool of people who are similar with respect to a number of characteristics, opinions, behaviours, etc.

At the specific study stage, the sampling method can undermine the exchangeability requirement. In opt-in panels the surveys are usually carried out on non-probability samples. To achieve the desired sample composition while data collection is ongoing, potential respondents are drawn from the panel using quotas, where the researcher pre-specifies a particular distribution across one or more variables. Here the assumption is that individuals who comprise each quota cell are exchangeable with non-sampled individuals who share those characteristics (Mercer *et al.*, 2017). Quotas are often (also in the case of my study) defined across basic demographic variables (i.e. age, sex, race, and education) that are generally insufficient for achieving exchangeability, and thus eliminating bias in the survey estimates. An additional strategy may be adopted, at the specific study stage, to meet the positivity requirement. The researcher could try to draw from the panelists a representative sample of the general population, offering targeted incentives to the subgroups of members who are under-represented in the panel. I cannot exclude the possibility that this response inducement could introduce another source of bias (i.e. satisficing behaviours displayed due to the primary aim of getting the reward), thus the researcher should take account of this risk.

As I highlighted, the exchangeability and positivity requirements may not be met. As a consequence, it may not be feasible to meet the composition requirement. In particular, it is difficult to achieve the desired sample composition through sampling alone, thus post-survey adjustment is still needed. Calibration and propensity score weighting are the two most common approaches to weighting. I used the latter, which involves combining a nonprobability sample (i. e. the ILC sample) with a parallel gold-standard data source (i. e. the AEL survey) as a reference sample. My findings have shown that this adjustment strategy is effective in reducing the overall bias, and works better for some estimates (i.e. some socio-demographic variables) than for others (i.e. behavioural questions). Unfortunately, it failed to completely remove the bias, as the exchangeability and positivity requirements were not completely met. This is what happens commonly in practice because it is difficult to see whether the variables utilized in sampling and adjustment account for any indirect confounding effect resulting from recruitment or sampling that can introduce bias into survey estimates (Mercer *et al.*, 2017).

In conclusion, non-probability online panels in future could improve their effectiveness in surveying the general population in Italy, when the middle-age and older people are likely to become more technologically savvy, for example. In the meantime, against the background of the uncertainties described above, what best practices should researchers adopt in non-probability online-panel surveys? Taking up the suggestions proposed by Mercer et al. (2017), and Schonlau and Couper (2017), scholars should provide academics, practitioners, and policy makers with the methodological details of their studies as this information is key to understanding and interpreting the research findings, and should always be fully and clearly reported. In particular, it is crucial to make explicit the following: the set of assumptions with respect to the three requirements aforementioned, the auxiliary variables used for sampling and adjustment, the incentives offered, the purpose of the survey, the “response rates”[31] obtained, and the intended use of the resulting estimates (Schonlau and Couper, 2017). Indeed, openness in reporting is key to evaluating and critiquing specific research findings, increase our understanding of when and how best to use opt-in panel surveys, and improve methodological practice (Mercer *et al.*, 2017; Schonlau and Couper, 2017).

[31] As discussed in Chapter 4, they are not always reported, there is little agreement on the terminology used to refer to them, and sometimes it is not clear how they are calculated.

# Appendices

**APPENDIX 1.** Summary of the 74 references included in the systematic review.

| ID | Authors and year of publication | Reference | Country of the study | Type of resource | Target population | Number of panels | Number of studies | Focus of the reference |
|---|---|---|---|---|---|---|---|---|
| 01 | Shin, Johnson, and Rao, 2012 | "Survey Mode Effects on Data Quality: Comparison of Web and Mail Modes in a U.S. National Panel Survey" *Social Science Computer Review* 30(2): 212-228. | USA | Journal article | general population | 1 | 1 | both panel itself and as a sample source |
| 02 | Bosnjak et al., 2013 | "Sample composition discrepancies in different stages of a probability-based online panel" *Field Methods* 25(4): 339-360. | Germany | Journal article | general population | 1 | 1 | panel itself |
| 03 | Revilla and Saris, 2013 | "A Comparison of the Quality of Questions in a Face-to-face and a Web Survey" *International Journal of Public Opinion Research* 25(2): 242-253. | The Netherlands | Journal article | general population | 1 | 1 | panel itself |
| 04 | Goeritz and Luthe, 2013a | "How Do Lotteries and Study Results Influence Response Behavior in Online Panels?" *Social Science Computer Review* 31(3): 371-385. | Germany | Journal article | people from all walks of life | 1 | 2 | both panel itself and as a sample source |
| 05 | Scherpenzeel and Toepoel, 2012 | "Recruiting A Probability Sample For An Online Panel: Effects Of Contact Mode, Incentives, And Information" *Public Opinion Quarterly* 76(3): 470-490. | The Netherlands | Journal article | na | 1 | 1 | panel itself |
| 06 | Peugh and Wright, 2012 | "Surveying Rare Populations Using a Probability based Online Panel" *Survey Practice* 5(3): 1-5. | USA | Journal article | American Jewish population (=rare population) | 1 | 1 | both panel itself and as a sample source |
| 07 | Hansen and Pedersen, 2012 | "Efficiency of Different Recruitment Strategies for Web Panels" *International Journal of Public Opinion Research* 24(2): 238-249. | Denmark | Journal article | na | 1 | 1 | panel itself |
| 08 | Revilla, 2013 | "Measurement invariance and quality of composite scores in a face-to-face and a web survey" *Survey Research Methods* 7(1): 17-28. | The Netherlands | Journal article | general population | 1 | 1 | both panel itself and as a sample source |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 09 | Brown et al., 2012 | "Evaluation of an online (opt-in) panel for public participation geographic information systems surveys" *International Journal of Public Opinion Research* 24(4): 534-545. | Australia | Journal article | residents in regional Victoria or Melbourne, who visited at least one of the nine parks in the study region within the last 12 months | 1 | 1 | panel itself |
| 10 | Couper et al., 2013 | "The Design of Grids in Web Surveys" *Social Science Computer Review* 31(3): 322-345. | USA | Journal article | general population | 2 | 2 | panel as a sample source |
| 11 | Keusch, 2013 | "The role of topic interest and topic salience in online panel web surveys." *International Journal of Market Research* 55(1): 59-80. | Austria | Journal article | general population | 1 | 1 | both panel itself and as a sample source |
| 12 | Goldenbeld and de Craen, 2013 | "The comparison of road safety survey answers between web-panel and face-to-face; Dutch results of SARTRE-4 survey" *Journal of Safety Research* 46: 13-20. | The Netherlands | Journal article | car drivers, motorcyclists or other road users | 1 | 1 | panel itself |
| 13 | Mavletova, 2013 | "Data Quality in PC and Mobile Web Surveys" *Social Science Computer Review* 31(6): 725-743. | Russia | Journal article | mobile web population | 1 | 1 | panel as a sample source |
| 14 | de Bruijne and Wijnant, 2013 | "Comparing Survey Results Obtained via Mobile Devices and Computers: An Experiment With a Mobile Web Survey on a Heterogeneous Group of Mobile Devices Versus a Computer-Assisted Web Survey" *Social Science Computer Review* 31(4): 482-504. | The Netherlands | Journal article | people aged 14+ and users of smartphones and tablets | 1 | 1 | both panel itself and as a sample source |
| 15 | Tsuboi et al., 2015 | "Selection bias of internet panel surveys: A comparison with a paper-based survey and national governmental statistics in Japan" *Asia-Pacific Journal of Public Health* 27(2): 2390-2399. | Japan | Journal article | visitors of the agency web page and users of affiliated programs | 1 | 1 | panel itself |
| 16 | Mavletova and Couper, 2013 | "Sensitive topics in PC Web and mobile web surveys: Is there a difference?" *Survey Research Methods* 7(3): 191-205. | Russia | Journal article | mobile web population | 1 | 1 | both panel itself and as a sample source |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 17 | Goeritz and Luthe, 2013b | "Effects of Lotteries on Response Behavior in Online Panels" *Field Methods* 25(3): 219-237. | Germany | Journal article | different walks of life | 1 | 1 | panel as a sample source |
| 18 | Leenheer and Scherpenzeel, 2013 | "Does It Pay Off to Include Non-Internet Households in an Internet Panel?" *International Journal of Internet Science* 8(1): 17-29. | The Netherlands | Journal article | general population | 1 | 1 | panel itself |
| 19 | Binswanger, Schunk, and Toepoel, 2013 | "Panel Conditioning in Difficult Attitudinal Questions" *Public Opinion Quarterly* 77(3): 783-797. | The Netherlands | Journal article | general population | 2 | 1 | panel itself |
| 20 | Wells, Bailey, and Link, 2014 | "Comparison of Smartphone and Online Computer Survey Administration" *Social Science Computer Review* 32(2): 238-255. | USA | Journal article | smartphone owners | 1 | 1 | both panel itself and as a sample source |
| 21 | Goeritz and Luthe, 2013c | "Lotteries and study results in market research online panels" *International Journal of Market Research* 55(5): 611-616. | Germany | Journal article, Journal article | people from all walks of life | 1 | 1 | panel as a sample source |
| 22 | Lugtig, 2014 | "Panel Attrition - Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers" *Sociological Methods & Research* 43(4): 699-723. | The Netherlands | Journal article | general population | 1 | 1 | panel itself |
| 23 | Cella et al., 2013 | "Comparison of US Panel Vendors for Online Surveys" *JMIR Publications* 15(11): e260. | USA | Journal article | general population | 7 | 1 | panel itself |
| 24 | Schonlau, Weidmer, and Kapteyn, 2014 | "Recruiting an Internet Panel Using Respondent-Driven Sampling" *Journal of Official Statistics* 30(2): 291-310. | USA | Journal article | people aged 18 and older | 1 | 1 | panel itself |
| 25 | Toepoel and Lugtig, 2014 | "What Happens if You Offer a Mobile Option to Your Web Panel? Evidence From a Probability-Based Panel of Internet Users" *Social Science Computer Review* 32(4): 544-560. | The Netherlands | Journal article | people aged 18 and older with access to the Internet and in possession of a smartphone | 1 | 1 | panel as a sample source |
| 26 | de Bruijne and Wijnant, 2014a | "Improving Response Rates and Questionnaire Design for Mobile Web Surveys" *Public Opinion Quarterly* 78(4): 951-962. | The Netherlands | Journal article | people aged 16+ who use a smartphone with an Internet connection | 1 | 1 | both panel itself and as a sample source |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 27 | de Bruijne and Wijnant, 2014b | "Mobile Response in Web Panels" *Social Science Computer Review* 32(6): 728-742. | The Netherlands | Journal article | general population | 2 | 1 | panel itself |
| 28 | Erens et al., 2014 | "Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison With a Probability Sample Interview Survey" *Journal of Medical Internet Research* 16(12): e276. | United Kingdom | Journal article | general population | 4 | 1 | panel itself |
| 29 | Blom et al., 2016 | "A Comparison of Four Probability-Based Online and Mixed-Mode Panels in Europe" *Social Science Computer Review* 34(1): 8-25. | The Netherlands, Germany and France | Journal article | general population | 4 | 4 | panel itself |
| 30 | Buskirk and Andrus, 2014 | "Making Mobile Browser Surveys Smarter Results from a Randomized Experiment Comparing Online Surveys Completed via Computer or Smartphone" *Field Methods* 26(4): 322-342. | USA | Journal article | US adults | 1 | 1 | panel as a sample source |
| 31 | Toepoel and Schonlau, 2015 | "Straightlining in Web survey panels over time" *Survey Research Methods* 9(2): 125-137. | The Netherlands | Journal article | general population | 1 | 1 | panel itself |
| 32 | Revilla et al., 2015 | "Can a non-probabilistic online panel achieve question quality similar to that of the European Social Survey?" *International Journal of Market Research* 57(3): 395-412. | Spain | Journal article | users of many websites | 1 | 1 | panel itself |
| 33 | Jones, House, and Zhifeng, 2015 | "Respondent Screening and Revealed Preference Axioms: Testing Quarantining Methods for Enhanced Data Quality in Web Panel Survey" *Public Opinion Quarterly* 79(3): 687-709. | USA | Journal article | grocery shoppers who have purchased fresh blueberries in the last year | 1 | 1 | panel itself |
| 34 | Bonnichsen and Olsen, 2015 | "Correcting for non-response bias in contingent valuation surveys concerning environmental non-market goods: an empirical investigation using an online panel" *Journal of Environmental Planning and Management* 59(2): 245-262. | Denmark | Journal article | target population of the tested area and a panel sample | 1 | 1 | panel itself |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 35 | Struminskaya, de Leeuw, and Kaczmirek, 2015 | "Mode System Effects in an Online Panel Study: Comparing a Probability-based Online Panel with two Face-to-Face Reference Surveys" *Methods, data, analyses* 9(1): 87-110. | Germany | Journal article | Internet users aged 18 and older | 1 | 1 | panel itself |
| 36 | Lee et al., 2015 | "Comparison of telephone RDD and online panel survey modes on CPGI scores and co-morbidities" *International Gambling Studies* 15(3): 435-449. | South Korea | Journal article | general population | 1 | 1 | panel itself |
| 37 | Struminskaya, 2016 | "Respondent Conditioning in Online Panel Surveys: Results of Two Field Experiments" *Social Science Computer Review* 34(1): 99-115. | Germany | Journal article | Internet users aged 18 and older | 1 | 1 | panel itself |
| 38 | Eckman, 2016 | "Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?" *Social Science Computer Review* 34(1): 41-58. | The Netherlands | Journal article | general population | 1 | 1 | panel itself |
| 39 | Golden et al., 2016 | "A multi-group analysis of online survey respondent data quality: Comparing a regular USA consumer panel to MTurk samples" *Journal of Business Research* 69(8): 3139-3148. | USA | Journal article | general population | 2 | 1 | panel itself |
| 40 | Pedersen and Nielsen, 2016 | "Improving Survey Response Rates in Online Panels: Effects of Low-Cost Incentives and Cost-Free Text Appeal Interventions" *Social Science Computer Review* 34(2): 229-243. | Denmark | Journal article | general population | 1 | 1 | both panel itself and as a sample source |
| 41 | Schonlau, 2015 | "What do web survey panel respondents answer when asked ‘Do you have any other comment?’" *Survey Methods: Insights from the field*. Retrieved from http://surveyinsights.org/?p=6899. | The Netherlands | Journal article | immigrants | 2 | 1 | panel as a sample source |
| 42 | Arn, Klug, and Kolodziejski, 2015 | "Evaluation of an Adapted Design in a Multi-device Online Panel: A DemoSCOPE Case Study" *Methods, data, analyses* 9(2): 185-212. | Switzerland | Journal article | market research company's clients | 1 | 2 | both panel itself and as a sample source |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 43 | Bosnjak, Struminskaya, and Weyandt, 2015 | "The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-based General Population Panel" *Methods, data, analyses* 9(2): 261-292. | Germany | Journal article | general population | 1 | 1 | panel as a sample source |
| 44 | Blom, Gathmann, and Krieger, 2015 | "Setting Up an Online Panel Representative of the General Population The German Internet Panel" *Field Methods* 27(4): 391-408. | Germany | Journal article | general population | 1 | 1 | panel itself |
| 45 | Sell, Goldberg, and Conron, 2015 | "The Utility of an Online Convenience Panel for Reaching Rare and Dispersed Populations" *PLoS ONE* 10(12): e0144011. | USA | Journal article | users of the Google Opinion Rewards application who have smartphones operated by Google's Android operating system | 1 | 1 | panel itself |
| 46 | Engel, 2014 | "Response Behavior in an Adaptive Survey Design for the Setting-Up Stage of a Probability-Based Access Panel in Germany" Pp. 207-222 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge. | Germany | Book section | people aged 18+ and entitled to vote for the German Federal Parliament | 1 | 1 | panel itself |
| 47 | Scherpenzeel, 2014 | "Survey Participation in a Probability-Based Internet Panel in the Netherlands" Pp. 223-235 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | panel itself |
| 48 | Rendtel and Amarov, 2014 | "The Access Panel of German Official Statistics as a Selection Frame" Pp. 236-249 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge. | Germany | Book section | general population | 1 | 1 | panel itself |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 49 | Enderle and Münnich, 2014 | "Accuracy of Estimates in Access Panel Based Surveys" Pp. 250-265 in *Improving Survey Methods: Lessons from Recent Research*, edited by U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis. New York: Routledge. | Germany | Book section | general population | 1 | 1 | panel itself |
| 50 | Struminskaya et al., 2014 | "Assessing representativeness of a probability-based online panel in Germany" Pp. 61-85 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | Germany | Book section | Internet users aged 18 and older | 1 | 1 | panel itself |
| 51 | Steinmetz et al., 2014 | "Improving web survey quality: Potentials and constraints of propensity score adjustments" Pp. 273-298 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | The Netherlands | Book section | general population | 1 | 1 | panel itself |
| 52 | Zhang, 2014 | "Estimating the effects of nonresponses in online panels through imputation" Pp. 299-310 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | USA | Book section | people aged 18 and older | 1 | 1 | both panel itself and as a sample source |
| 53 | Malhotra, Miller, and Wedeking, 2014 | "The relationship between nonresponse strategies and measurement error: Comparing online panel surveys to traditional surveys" Pp. 313-336 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | USA | Book section | people aged 18 and older | 2 | 3 | both panel itself and as a sample source |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 54 | Roberts, Allum, and Sturgis, 2014 | "Nonresponse and measurement error in an online panel: Does additional effort to recruit reluctant respondents result in poorer quality data?" Pp. 337-362 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | USA | Book section | people aged 18 and older | 1 | 1 | both panel itself and as a sample source |
| 55 | Drewes, 2014 | "An empirical test of the impact of smartphones on panel-based online data collection" Pp. 367-386 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | Germany | Book section | visitors of a large variety of websites | 1 | 1 | both panel itself and as a sample source |
| 56 | Baker et al., 2014 | "Validating respondents' identity in online samples: The impact of efforts to eliminate fraudulent respondents" Pp. 441-456 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | USA | Book section | US adults | 1 | 1 | panel itself |
| 57 | Grönlund and Strandberg, 2014 | "Online panels and validity: Representativeness and attrition in the Finnish eOpinion panel" Pp. 86-103 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | Finland | Book section | general population | 1 | 1 | panel itself |
| 58 | McCutcheon, Rao, and Kaminska, 2014 | "The untold story of multi-mode (online and mail) consumer panels: From optimal recruitment to retention and attrition" Pp. 104-126 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | USA | Book section | general population | 1 | 1 | panel itself |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 59 | Lugtig, Das, and Scherpenzeel, 2014 | "Nonresponse and attrition in a probability-based online panel for the general population" Pp. 135-153 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | The Netherlands | Book section | general population | 1 | 1 | panel itself |
| 60 | Göritz, 2014 | "Determinants of the starting rate and the completion rate in online panel studies" Pp. 154-170 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | Germany | Book section | people from all walks of life | 1 | 1 | panel itself |
| 61 | Keusch, Batinic, and Mayerhofer, 2014 | "Motives for joining nonprobability online panels and their association with survey participation behavior" Pp. 171-191 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | Austria | Book section | general population | 1 | 1 | panel itself |
| 62 | Scherpenzeel and Toepoel, 2014 | "Informing panel members about study results: Effects of traditional and innovative forms of feedback on participation" Pp. 192-213 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | The Netherlands | Book section | general population | 1 | 1 | panel itself |
| 63 | Hillygus, Jackson, and Young, 2014 | "Professional respondents in nonprobability online panels" Pp. 219-237 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | USA | Book section | visitors of popular websites | 1 | 1 | panel itself |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 64 | Greszki, Meyer, and Schoen, 2014 | "The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels" Pp. 238-262 in *Online Panel Research: A Data Quality Perspective*, edited by M. Callegaro, R. Baker, J. Bethlehem, A. S. Goritz, and J. A. Krosnick. Chichester: Wiley. | Germany and USA | Book section | people aged 18 and older | 2 | 2 | both panel itself and as a sample source |
| 65 | Toepoel and Lugtig, 2016 | "The use of Pcs, smartphones and tablets in a probability based panel survey. Effects on survey measurement error." *Social Science Computer Review* 34(1): 78-94. | The Netherlands | Journal article | general population | 1 | 1 | both panel itself and as a sample source |
| 66 | Heen, Lieberman, and Miethe, 2014 | "A Comparison of Different Online Sampling Approaches for Generating National Samples" *Center for Crime and Justice Policy* 1: 1-8. | USA | Journal article | people aged 18 and older | 3 | 1 | panel itself |
| 67 | Scherpenzeel and Bethlehem, 2011 | "How Representative Are Online Panels? Problems of Coverage and Selection and Possible Solutions" Pp. 105-132 in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | panel itself |
| 68 | Vis and Marchand, 2011 | "Challenges in Reaching Hard-to-Reach Groups in Internet Panel Research" Pp. 271-290 in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | panel itself |

APPENDIX 1. *Continued.*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 69 | Avendano, Scherpenzeel, and Mackenbach, 2011 | "Can Biomarkers Be Collected in an Internet Survey? A Pilot Study in the LISS Panel" Pp. 371-412 in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | panel itself |
| 70 | Kaczmirek, 2011 | "Attention and Usability in Internet Surveys: Effects of Visual Feedback in Grid Questions" Pp. 191-214 in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | both panel itself and as a sample source |
| 71 | Oudejans and Christian, 2011 | "Using Interactive Features to Motivate and Probe Responses to Open-Ended Questions" Pp. 215-244 in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | both panel itself and as a sample source |
| 72 | Ester and Vinken, 2011 | "Measuring Attitudes Toward Controversial Issues in Internet Surveys: Order Effects of Open and Closed Questioning" Pp. 245-268 in *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, edited by M. Das, P. Ester, and L. Kaczmirek. New York: Routledge. | The Netherlands | Book section | general population | 1 | 1 | both panel itself and as a sample source |
| 73 | Matthijsse, de Leeuw, and Hox, 2015 | "Internet panels, professional respondents, and data quality" *Methodology* 11(3): 81-88. | The Netherlands | Journal article | people aged between 18 and 65 | 19 | 1 | panel itself |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 74 | Cho et al., 2015 | "Attitudes Toward Risk and Informed Consent for Research on Medical Practices: A Cross-Sectional Survey" *Annals of Internal Medicine* 162(10): 690-696. | USA | Journal article | US adults | 1 | 1 | panel itself |

**APPENDIX 2**. Definitions of variables used to answer the research questions, coding procedure, and analytical approach.

Definitions of variables:

*RQ1 – Variables applied to describe the types of online panels used in survey methodology*

- “Membership composition” refers to the types of respondents’ populations included in the panel.
- “Recruitment strategy” deals with the methodology adopted to select the panel’s members from the target population.
- “Size of the panel” refers to the number of members included in the panel.
- “Field of establishment” deals with the market or research field where the panel is established.
- “Geographical coverage” refers to the geographical area to which the panel’s members belong.

*RQ2 – Variables related to the quality of online panels as addressed by survey methodologists*

- “Comparison of point estimates with the gold standard” deals with any type of analysis that compares online panel-survey data with other data from a reference national survey or from official national statistics.
- “Comparison of point estimates with another mode of data collection/study design” deals with any type of analysis that compares online panel-survey data with other data collected while conducting random digit-dialing phone surveys, mail surveys, opt-in surveys, face-to-face surveys, and other online panel surveys, which, however, are not considered a gold standard.
- “Weighting techniques” refer to discussions and/or research on weighting adjustments used to correct for different sources of survey errors and, therefore, to improve data quality from online panel surveys.
- “Professional respondents” refer to discussions and/or research on “well-trained or experienced survey-takers who seek out large numbers of surveys, typically for the cash and incentives offered” (Backer *et al.*, 2010a, pp. 756-757).
- “Speeders” refer to discussions and/or research on respondents who do “not thoroughly read the questions and use(s) minimal cognitive effort to provide answers that satisfy the question (to collect their incentive with as little time spent as possible)” (Smith *et al.* 2016).

- “Fraudulent or inattentive respondents” refer to discussion and/or research on panel members who “create multiple profiles to access more surveys or falsely answer screening questions to attempt to qualify for surveys” (Jones *et al.*, 2015) and “provide incorrect responses to questions inserted into the survey flow that require specific responses [...] or respondents who do not have knowledge that would be typically expected for the respondent group” (Smith *et al.*, 2016).
- “Panel-conditioning effect” refers to discussions and/or research on effects that occur when respondents “become more knowledgeable over the course of subsequent waves and (may) change their attitudes or even their behaviour” (Binswanger *et al.*, 2013).
- “Recruitment strategies for setting up the panel” deal with the study of methods, which include tools and questionnaire-design features used to contact people and invite them to become panelists.
- “Retention strategies for maintaining the panel” deal with methods, which include tools and survey-design features used to motivate panelists to stay on the panel and complete questionnaires.
- “Participants’ loyalty to the panel and membership tenure” refer to the study of participants’ cooperativeness and the duration of their membership.
- “Nonresponse issues” refer to the study of nonresponse error as a component of Total Survey Error, occurring when the number of individuals who are not contacted or refuse to participate in the survey is high and when they differ from sample members on some variables of interest (Groves *et al.*, 2009, p. 59). In addition, I included in this category some indicators (e.g. various definitions of response rate) that assess the nonresponse process in panel surveys without necessarily measuring nonresponse error.
- “Measurement error” refers to the study of another component of the Total Survey Error. It occurs when an answer given by a respondent does not accurately represent his or her attitudes or behaviour. This error is measured by the difference between the observed value, given a certain operational definition, and the “true” value (Groves *et al.*, 2009, p. 52). I adopted this category to refer to references dealing with the quality of answers given by the panel members that result from a measurement error (e.g. satisficing behaviour and social-desirability bias).
- “Questionnaire design” refers to studies in which some aspects of questionnaire design usually used by the online panel were discussed/researched.

*RQ3 – Variables used to describe the characteristics of online panel studies*

- "Sampling method" deals with the method used to select a sample of panel members who are invited to fill out a web questionnaire (to participate in an individual panel study).
- "Size of the study sample" refers to the number of panel members invited to participate in an individual panel study.
- "Questionnaire length" is measured as the number of questions included in the questionnaire and/or the estimated time to fill out the web questionnaire of the individual panel study.

*RQ4 – Variables applied to address the purposes for the use of online panels as a sample source for research on survey methodology*

- "Questionnaire design": In this case, the purpose of the study is to examine some questionnaire features, e.g. in experimental designs, to test for specific question options (e.g., question layout, question-order effect, a visual feature) and thereby improve the effectiveness of the web questionnaire.
- "Response process": Here, measures to increase response rate and indicators of nonresponse are coded as purposes when a study deals with strategies that can boost survey participation. Indicators of nonresponse are the variables used by scholars to study nonresponse processes. The focus here is on the issue of nonresponse as a research question, not on the nonresponse error as an indicator of panel-data quality.
- "Measurement error," as mentioned above, is addressed here as a general issue in the field of survey methodology, not as a specific component of panel-data quality.

Table 1. Coding procedure and analytical approach*.

| Aims/Research questions | Variables | Units of analysis | Number of units |
|---|---|---|---|
| Characteristics of online panels (RQ1) | Membership composition | unique online panels | 69 |
| | Recruitment strategy | | |
| | Size of the panel | | 22 |
| | Field of establishment | | 69 |
| | Geographical coverage | | |

Table 1. *Continued*.

| | | | |
|---|---|---|---|
| Dimensions of quality of online panels addressed (RQ2) | Nonresponse issues | references | 25 |
| | Respondents' behaviour (speeders', fraudulents' and professional respondents' behaviour, and panel conditioning) | | 18 |
| | Comparison of point estimates with the gold standard | | 17 |
| | Weighting techniques | | 15 |
| | Participants' loyalty to the panel and membership tenure | | 15 |
| | Measurement error | | 15 |
| | Comparison of point estimates with other modes of data collection/study designs | | 11 |
| | Recruitment strategies for setting up the panel | | 8 |
| | Retention strategies for maintaining the panel | | 4 |
| | Questionnaire design | | 1 |
| Characteristics of online panel studies (RQ3) | Sampling method | unique panel studies | 81 |
| | Size of the study sample | | 66 |
| | Questionnaire length (number of questions and/or estimated time to fill in the questionnaire) | | 36 |
| Purposes of the usage of online panels as a sample source for research on survey methodology (RQ4) | Measurement error | unique panel studies | 27 |
| | Response process | | 23 |
| | Questionnaire design | | 13 |

* For definitions of variables see above in this Appendix.

**APPENDIX 3**. Dimensions of quality of online panels addressed by survey methodologists (RQ2).

Table 2. Indicators of nonresponse error used to study data quality of online panels (N = 74 references, N = 25 references mentioning at least one indicator).

| Indicators of nonresponse error | References | | |
|---|---|---|---|
| | N | % (out of 25) | % (out of 74) |
| *At recruitment stage* | 9 | 36.0 | 12.2 |
| recruitment rate/number of recruits | 8 | 32.0 | 10.8 |
| profile/panel registration rate | 5 | 20.0 | 6.8 |
| *At specific study stage* | 16 | 64.0 | 21.6 |
| starting/participation/completion rate | 13 | 52.0 | 17.6 |
| screening/eligibility rate | 4 | 16.0 | 5.4 |
| response propensity | 3 | 12.0 | 4.1 |
| break-off rate | 2 | 8.0 | 2.7 |
| cumulative response rate | 2 | 8.0 | 2.7 |
| absorption rate | 1 | 4.0 | 1.4 |

Table 3. Recruitment strategies as indicators of online panel data quality (N = 74 references, N = 8 references mentioning at least one type of recruitment strategy).

| Recruitment strategies | References | | |
|---|---|---|---|
| | N | % (out of 8) | % (out of 74) |
| monetary incentive | 5 | 62.5 | 6.8 |
| reminder (letter, SMS, and e-mail) | 3 | 37.5 | 4.1 |
| multi-mode contact (face-to-face, phone, and mail) | 3 | 37.5 | 4.1 |
| gift (tablet PCs, 3G Internet, and a small piece of quality chocolate) | 2 | 25 | 2.7 |
| (special content of the) advance letter | 2 | 25 | 2.7 |
| respondent-driven sampling method | 1 | 12.5 | 1.4 |
| sponsor | 1 | 12.5 | 1.4 |
| reasons for joining the online panel | 1 | 12.5 | 1.4 |

Table 4. Loyalty to the panel as indicator of online panel data quality (N = 74 references, N = 15 references mentioning at least one element of loyalty).

| Loyalty to the panel | References | | |
|---|---|---|---|
| | N | % (out of 15) | % (out of 74) |
| factors influencing attrition (background information and panel-specific factors) | 15 | 100.0. | 20.3 |
| attrition or retention rate | 6 | 40.0 | 8.1 |
| response rates of long stay or more recent panelists | 3 | 20.0 | 4.1 |
| frailty effect | 1 | 6.7 | 1.4 |
| relations between response quality and attrition propensity | 1 | 6.7 | 1.4 |

Table 5. Indicators of measurement error used to study data quality of online panels (N = 74 references, N = 15 references mentioning an indicator).

| Indicators of measurement error | References | | |
|---|---|---|---|
| | N | % (out of 15) | % (out of 74) |
| satisficing behaviour | 10 | 66.7 | 13.5 |
| quality estimates obtained adopting a Multitrait-Multimethod (MTMM) matrix | 2 | 13.3 | 2.7 |
| bias between true values and estimates | 1 | 6.7 | 1.4 |
| social desirability bias | 1 | 6.7 | 1.4 |
| strong axiom of revealed preference (SARP) violations | 1 | 6.7 | 1.4 |

Table 6. Comparison with a gold standard as indicator of online panel data quality (N = 74 references, N = 17 references mentioning at least one group of variables).

| Variables used for the comparison with a gold standard | References | | |
|---|---|---|---|
| | N | % (out of 17) | % (out of 74) |
| socio-demographics | 15 | 88.2 | 20.3 |
| attitudinal variables | 6 | 35.3 | 8.1 |
| behavioural variables | 6 | 35.3 | 8.1 |
| urbanicity | 3 | 17.6 | 4.1 |
| Internet access and use | 3 | 17.6 | 4.1 |
| other (gender identity, religious confession, and health status) | 3 | 17.6 | 4.1 |

Table 7a. Variables used for comparison with other modes of data collection/study designs (N = 74 references, N = 11 references mentioning at least one group of variables).

| Variables | References | | |
|---|---|---|---|
| | N | % (out of 11) | % (out of 74) |
| socio-demographics | 7 | 63.6 | 9.5 |
| attitudinal variables | 4 | 36.4 | 5.4 |
| behavioural variables | 2 | 18.2 | 2.7 |
| panel-specific factors (length of stay in panel, numbers of panels belonged to, and average number of surveys completed per week) | 2 | 18.2 | 2.7 |
| Internet access | 1 | 9.1 | 1.4 |
| urbanicity | 1 | 9.1 | 1.4 |
| other (co-morbidities and religious affiliation) | 2 | 18.2 | 2.7 |

Table 7b. Survey modes/study designs used for comparison with the reported online panels (N = 74 references, N = 11 references mentioning at least one survey mode/study design).

| Survey modes/study designs | References | | |
|---|---|---|---|
| | N | % (out of 11) | % (out of 74) |
| other online panels | 4 | 36.4 | 5.4 |
| other sampling frames for a web survey (self-selected, random household or on-site recruitment sample) | 3 | 27.3 | 4.1 |
| F2F survey | 3 | 27.3 | 4.1 |
| RDD survey | 2 | 18.2 | 2.7 |
| mail survey | 1 | 9.1 | 1.4 |
| CATI survey | 1 | 9.1 | 1.4 |

Table 8. Weighting techniques as indicators of online panel data quality (N = 74 references, N = 15 references mentioning a weighting technique).

| Weighting techniques | References | | |
|---|---|---|---|
| | N | % (out of 15) | % (out of 74) |
| post-stratification weights | 5 | 33.3 | 6.8 |
| design weights | 3 | 20.0 | 4.1 |
| propensity scores | 3 | 20.0 | 4.1 |
| a combination of different types of weights | 3 | 20.0 | 4.1 |
| imputation of missing responses | 1 | 6.7 | 1.4 |

**APPENDIX 4**. Characteristics of individual online panel studies (RQ3).

Table 9. Sampling designs for an individual survey (N = 83 unique online panel studies, N = 81 unique online panel studies reporting the type of sampling method).

| Types of sampling method | Unique studies | | |
|---|---|---|---|
| | N | % (out of 81) | % (out of 83) |
| probability sampling | 30 | 37.0 | 36.1 |
| non-probability sampling | 21 | 25.9 | 25.3 |
| both probability and no sampling | 2 | 2.5 | 2.4 |
| no sampling, all panelists | 28 | 34.6 | 33.7 |

Table 10. Size of the samples for an individual survey (N = 83 unique online panel studies, N = 66 unique online panel studies reporting the number of people selected for the survey).

| Number of people selected for the survey | Unique studies | | |
|---|---|---|---|
| | N | % (out of 66) | % (out of 83) |
| 300-1,500 | 12 | 18.2 | 14.5 |
| 1,501-3,200 | 23 | 34.8 | 27.7 |
| 3,201-5,999 | 7 | 10.6 | 8.4 |
| 6,000-10,000 | 14 | 21.2 | 16.9 |
| 10,001-20,000 | 2 | 3.0 | 2.4 |
| 20,001-154,000 | 8 | 12.1 | 9.6 |

Table 11. Questionnaire length in an individual survey: number of questions (N = 83 unique online panel studies, N = 15 unique online panel studies reporting the number of questions in the questionnaire).

| Number of questions in the questionnaire | Unique studies | | |
|---|---|---|---|
| | N | % (out of 15) | % (out of 83) |
| 2 | 1 | 6.7 | 1.2 |
| 3 | 1 | 6.7 | 1.2 |
| 6-11 | 1 | 6.7 | 1.2 |
| 12 | 1 | 6.7 | 1.2 |
| 13 | 1 | 6.7 | 1.2 |
| 19 | 1 | 6.7 | 1.2 |
| 24 | 2 | 13.3 | 2.4 |
| 26 | 2 | 13.3 | 2.4 |
| 28 | 1 | 6.7 | 1.2 |
| 67 | 1 | 6.7 | 1.2 |
| 83 | 1 | 6.7 | 1.2 |
| 120 | 1 | 6.7 | 1.2 |
| 130 | 1 | 6.7 | 1.2 |

Table 12. Questionnaire length in an individual survey: completion time (N = 83 unique online panel studies, N = 30 unique online panel studies reporting the completion time).

| Completion time (in minutes) | Unique studies | | |
|---|---|---|---|
| | N | % (out of 30) | % (out of 83) |
| 1.3-14.2 | 1 | 3.3 | 1.2 |
| 5 | 2 | 6.7 | 2.4 |
| 7.5 | 1 | 3.3 | 1.2 |
| 8 | 1 | 3.3 | 1.2 |
| 10 | 5 | 16.7 | 6.0 |
| 12 | 1 | 3.3 | 1.2 |
| 12.5 | 3 | 10.0 | 3.6 |
| 14.1 | 1 | 3.3 | 1.2 |
| 15 | 4 | 13.3 | 4.8 |
| 17.5 | 1 | 3.3 | 1.2 |
| 20 | 1 | 3.3 | 1.2 |
| 22.5 | 2 | 6.7 | 2.4 |
| 27.5 | 1 | 3.3 | 1.2 |
| 30 | 6 | 20.0 | 7.2 |

**APPENDIX 5**. Issues addressed using online panels as a sample source for research in survey methodology (RQ4).

Table 13. Indicators of measurement error addressed using online panels (N = 83 unique online panel studies).

| Indicators of measurement error | Unique studies | |
|---|---|---|
| | N | % (out of 83) |
| satisficing in closed questions | 21 | 25.3 |
| satisficing in open-ended questions | 12 | 14.5 |
| time | 12 | 14.5 |
| mode effect | 2 | 2.4 |
| conspicuous response behaviour in grid statements | 1 | 1.2 |
| socially undesirable responses | 1 | 1.2 |

Table 14. Issues on response/nonresponse process addressed using online panels (N = 83 unique online panel studies).

| | Unique studies | |
|---|---|---|
| | N | % (out of 83) |
| *Indicators of response/nonresponse process* | 21 | 25.3 |
| survey outcome rates (start rate. break-off rate, completion rate, nonresponse rate, and number of completed questionnaires) | 20 | 24.1 |
| number of call attempts/e-mail reminders to obtain panel/survey participation | 4 | 4.8 |
| response rate for specific questions | 3 | 3.6 |
| initial panel recruitment refusals/rate/screening | 2 | 2.4 |
| respondent reluctance in panel/survey recruitment | 2 | 2.4 |
| number of days after the field date it took to for the respondent to complete the survey | 2 | 2.4 |
| imputation as a way to estimate nonresponse bias | 1 | 1.2 |
| response speed (in the first 24 hours) | 1 | 1.2 |
| *Measures to increase response rates* | 6 | 7.2 |
| incentives (lotteries, monetary donations, and text appeals) | 4 | 4.8 |
| invitation text/mode | 2 | 2.4 |
| offering study results | 1 | 1.2 |

Table 15. Questionnaire design features addressed using online panels (N = 83 unique online panel studies).

| Questionnaire design features | Unique studies | |
|---|---|---|
| | N | % (out of 83) |
| question layout choices | 4 | 4.8 |
| interactive and visual features | 4 | 4.8 |
| questionnaire versions/layouts | 2 | 2.4 |
| items per screen | 2 | 2.4 |
| question order effect | 1 | 1.2 |
| question wordings | 1 | 1.2 |
| number and order of answer options | 1 | 1.2 |
| open-ended versus closed-ended | 1 | 1.2 |
| “Other” versus “Other, specify” | 1 | 1.2 |
| opinions about different features of the questionnaire (orientation, color, design, and usability) | 1 | 1.2 |

**APPENDIX 6**. ILC survey questionnaire.

| Section | Question |
|---|---|
| Internet use | 1. How often on average did you use the Internet in the last 12 months?<br>• Every day<br>• A few times a week<br>• Once a week<br>• A few times a month (less than 4)<br>• Less than once a month |
| | 2. How often did you use the Internet in the last 3 months at home, at work, at school/univeristy, or elsewhere? *(Tick an answer for each item)*<br>• At home<br>• At work (if different from home)<br>• At school/university<br>• At other people's houses<br>• Elsewhere<br>[scale: 1 every day; 2 a few times a week; 3 once a week; 4 a few times a month; 5 less than once a month; 6 never] |
| | 3. For which of the following communication activities did you use the Internet in the last 3 months? *(Tick an answer for each item)*<br>• Sending or receiving e-mails<br>• Telephoning over the Internet / video calls (via webcam) over the Internet (using applications, e.g. Skype, Facetime)<br>• Posting messages to chat sites, blogs, newsgroups or online discussion forum<br>• Using instant messaging<br>• Participating in social networks (creating user profile, posting messages or other contributions to Facebook, Twitter, etc.)<br>[No/Yes] |
| Food consumption | 4. How often do you usually eat the following foods?<br>*(Tick an answer for each item)*<br>• Legumes (dried or canned)<br>• Potatoes<br>• Salty snacks (chips, pop corn, pretzels, olives etc.)<br>• Sweets (cakes, sweet snacks, ice-cream etc.)<br>[scale: 1 more than once a day; 2 once a day; 3 a few times a week; 4 less than once a week; 5 never] |
| Drinking | 5. Do you drink wine or alcohol outside mealtimes?<br>• Every day<br>• A few times a week<br>• Less often<br>• Never |
| | *(If 5=1 or 2)*<br>6. Overall, in a week, how many glasses of wine or alcohol do you usually drink outside mealtimes?<br><br>Number of glasses per week □□ |

APPENDIX 6. *Continued*.

| | |
|---|---|
| Smoking | 7. Do you smoke?<br>• Yes<br>• No, but I used to smoke<br>• No, I have never smoked |
| Socio-political participation | 8. In the last 12 months: *(Tick an answer for each item)*<br>• I attended a political meeting<br>• I took part in a political demonstration<br>• I listened to a political debate<br>• I gave money to a political party<br>• I gave money to an association<br>• I did voluntary work<br>[No/Yes] |
| Watching TV | 9. Do you watch TV?<br>• No<br>• Yes, every day<br>• Yes, a few days a week |
| | *(If 9=2 or 3)*<br>10. In the days you watch TV, how long do you spend watching TV a day?<br>Hours □□ and minutes □□<br>I don't know |
| Socio-demographics | 11. What is the highest level of education you successfully completed?<br>• Less than primary education<br>• Primary education<br>• Lower secondary education<br>• Upper secondary education<br>• First/ Second stage of tertiary education |
| | 12. What is your employment status?<br>• Full/part-time employed<br>• Looking for work<br>• Unpaid work, e.g. domestic<br>• Student<br>• Permanently disabled<br>• Retired<br>• Other status |
| | 13. What is your marital status?<br>• Single<br>• Civil partnership<br>• Married<br>• Separated<br>• Divorced<br>• Widowed |
| | 14. Which Region are you living in?<br>*(Select one from the drop-down box)* |

APPENDIX 6. *Continued*.

| | |
|---|---|
| Environmental problems | 15. Which of the following environmental problems are you worried about? *(Tick max 5)*<br>• Global warming, the hole in the ozone layer<br>• Extinction of species<br>• Climate change<br>• Poor waste management<br>• Noise pollution<br>• Air pollution<br>• Soil pollution<br>• Water pollution<br>• Hydrogeological instability<br>• Man-made disasters<br>• Deforestation<br>• Electromagnetic pollution<br>• Destruction of the landscape caused by excessive construction<br>• Depletion of the earth's resources<br>• Other (specify) |

**APPENDIX 7**. Text of the invitation e-mail to the ILC survey.

Here’s our new survey!

[FIRSTNAME],
You are receiving this e-mail as a member of the Opinione.net Panel. We are once again offering you the chance to give your opinion.

ATTENTION: to fill out this questionnaire you need to have a certain profile, thus it is possible that, after you start the completion, the instrument will not allow you to complete the questionnaire.

To take part in the survey, please click on the link below:
[SURVEY LINK]

The questionnaire is about [LENGTH OF QUESTIONNAIRE] minutes, according to the matching between your answers and the respondent’s profile required by the study. The incentives you will receive depend on the estimated time to complete the questionnaire. Upon questionnaire completion [INCENTIVE] euros will be transferred to your panel account.
Thank you again for your collaboration!
Opinione.net
-----------------------------
The questionnaire completion is completely anonymous and is specifically addressed to you. Nobody else should fill out the questionnaire: that’s why your opinion is important for us!
-----------------------------
Your participation to this survey is completely free. If you do NOT want to take part in THIS survey, please click on the link below:
[SURVEY DECLINE LINK]

**APPENDIX 8**. Question wording of the variables used to create the indicator "Internet population".

Questions related to Internet access:

Q12.1 Do you or anyone in your household have access to the Internet at home? (by any device)

1. No
2. *Yes*

Q12.3 What are the reasons for not having access to the Internet at home? (tick all that apply)

1. *Have access to Internet elsewhere (e.g. at work, at the place of study, in other people's homes)*
2. Don't need Internet (because not useful, not interesting, etc.)
3. Equipment costs too high
4. Access costs too high (telephone, DSL subscription etc.)
5. Lack of skills
6. Privacy or security concerns
7. Broadband internet is not available in our area
8. Other

Questions related to Internet use:

Q1 How often on average did you use the Internet in the last 12 months?

1. *Every day*
2. *A few times a week*
3. Once a week
4. A few times a month (less than 4)
5. Less than once a month

Q3 For which of the following communication activities did you use the Internet in the last 3 months? [*yes*/no answer for each item]

1. *Sending or receiving e-mails*
2. *Telephoning over the Internet / video calls (via webcam) over the Internet (using applications, e.g. Skype, Facetime)*
3. *Posting messages to chat sites, blogs, newsgroups or online discussion forum*
4. *Using instant messaging*
5. *Participating in social networks (creating user profile, posting messages or other contributions to Facebook, Twitter, etc.)*
6. Posting opinions on civic or political issues via websites (e.g. blogs, social networks, etc.)
7. Taking part in on-line consultations or voting to define social (civic) or political issues (e.g. urban planning, signing a petition)
8. Participating in professional networks (creating user profile, posting messages or other contributions to LinkedIn, Xing, etc.)
9. Uploading self-created content (text, photos, music, videos, software etc.) to any website to be shared

Note: the answer options used to create the indicator "Internet population" are in italics.

**APPENDIX 9**. Sample composition from each data source.

|  | General population | Internet population | Panel members | Panel members invited to ILC survey | ILC sample |  |  |
|---|---|---|---|---|---|---|---|
|  |  |  |  |  | unweighted | weighted to mirror the general population | weighted to mirror the Internet population |
| *Area of residence* |  |  |  |  |  |  |  |
| North West | 26.7 | 28.9 | 25.2 | 27.3 | 26.2 | 25.4 | 28.6 |
| North East | 19.1 | 20.7 | 22.2 | 21.9 | 19.4 | 22.6 | 20.7 |
| Centre | 20.0 | 21.0 | 17.9 | 20.5 | 19.6 | 20.7 | 21.1 |
| South | 23.1 | 19.7 | 23.6 | 21.1 | 23.3 | 20.4 | 19.3 |
| Islands | 11.0 | 9.7 | 11.0 | 9.2 | 11.6 | 10.8 | 10.3 |
| (N) | (50,221,662) | (25,698,929) | (8,065) | (3,907) | (2,007) | (50,338,059) | (25,332,076) |
| *Sex* |  |  |  |  |  |  |  |
| Men | 48.0 | 52.8 | 45.5 | 53.3 | 53.1 | 46.5 | 50.9 |
| Women | 52.0 | 47.2 | 54.5 | 46.7 | 46.9 | 53.5 | 49.1 |
| (N) | (50,293,523) | (25,728,117) | (8,071) | (3,907) | (2,007) | (50,338,059) | (25,332,076) |
| *Marital status* |  |  |  |  |  |  |  |
| Single | 28.9 | 41.2 | 49.4 | 43.0 | 42.8 | 36.1 | 47.0 |
| Married | 53.0 | 47.6 | 45.6 | 50.0 | 50.4 | 52.2 | 46.9 |
| Divorced/widowe | 18.1 | 11.2 | 5.0 | 7.0 | 6.8 | 11.7 | 6.1 |
| N | (50,293,523) | (25,728,117) | (8,071) | (3,695) | (2,007) | (50,319,828) | (25,321,155) |

APPENDIX 9. *Continued*.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Age group* | | | | | | | |
| 18-24 | 8.2 | 13.8 | 12.1 | 9.8 | 8.9 | 8.0 | 12.3 |
| 25-34 | 13.7 | 20.9 | 26.1 | 21.4 | 20.4 | 15.1 | 20.3 |
| 35-44 | 18.0 | 24.6 | 22.9 | 19.4 | 20.8 | 18.8 | 24.0 |
| 45-54 | 19.1 | 21.9 | 19.5 | 20.4 | 21.5 | 21.9 | 22.9 |
| 55-64 | 15.2 | 13.1 | 12.2 | 15.8 | 16.0 | 15.4 | 14.4 |
| 65+ | 25.9 | 5.6 | 7.1 | 13.3 | 12.4 | 20.8 | 6.1 |
| (N) | (50,293,523) | (25,728,117) | (8,071) | (3,896) | (2,007) | (50,338,060) | (25,332,076) |
| *Education* | | | | | | | |
| Tertiary | 14.5 | 24.2 | 39.1 | 37.7 | 37.5 | 14.3 | 25.5 |
| Secondary | 37.4 | 50.9 | 51.7 | 52.3 | 53.2 | 37.6 | 50.6 |
| Primary | 29.3 | 23.2 | 9.0 | 9.4 | 8.4 | 29.5 | 22.2 |
| No education | 18.7 | 1.8 | 0.1 | 0.6 | 0.9 | 18.6 | 1.7 |
| (N) | (50,293,523) | (25,728,117) | (6,916) | (3,569) | (2,007) | (50,338,059) | (25,332,077) |
| *Occupation* | | | | | | | |
| In employment | 43.4 | 61.2 | 56.8 | 53.3 | 54.1 | 42.8 | 52.9 |
| Unemployed | 12.4 | 14.7 | 10.8 | 10.6 | 11.4 | 11.7 | 13.4 |
| Inactive | 44.2 | 24.1 | 32.4 | 36.1 | 34.4 | 45.5 | 33.7 |
| (N) | (50,293,523) | (25,728,117) | (6,445) | (3,461) | (1,931) | (48,887,718) | (24,495,904) |

Note: AEL data on the general and the Internet populations are weighted using the weights provided with the dataset by ISTAT.

**APPENDIX 10**. Socio-demographic characteristics of the ILC survey respondents and nonrespondents.

| | Respondents | Nonrespondents |
|---|---|---|
| *Sex* | | |
| Man | 53.1 | 53.5 |
| Woman | 46.9 | 46.5 |
| (N) | (2,007) | (1,900) |
| *Age group*** | | |
| 18-24 | 8.9 | 10.7 |
| 25-34 | 20.4 | 22.4 |
| 35-44 | 20.8 | 17.8 |
| 45-54 | 21.5 | 19.1 |
| 55-64 | 16.0 | 15.6 |
| 65+ | 12.4 | 14.3 |
| (N) | (2,007) | (1,889) |
| *Area of residence**** | | |
| North West | 26.2 | 28.6 |
| North East | 19.4 | 24.4 |
| Centre | 19.6 | 21.4 |
| South | 23.3 | 18.8 |
| Islands | 11.5 | 6.8 |
| (N) | (2,007) | (1,900) |
| *Occupation*** | | |
| In employment | 53.9 | 52.5 |
| Unemployed | 11.7 | 9.0 |
| Inactive | 34.4 | 38.5 |
| (N) | (2,001) | (1,460) |
| *Marital status* | | |
| Single | 42.8 | 43.2 |
| Married | 50.4 | 49.5 |
| Divorced/widowed | 6.8 | 7.2 |
| (N) | (2,007) | (1,688) |
| *Education* | | |
| Tertiary | 37.5 | 38.1 |
| Secondary | 53.2 | 51.1 |
| Primary or no education | 9.3 | 10.8 |
| (N) | (2,007) | (1,573) |

Note: ***p≤0.001; **p≤0.01; *p≤0.05. There are no differences that are statistically significant at the level of 0.05.

**APPENDIX 11**. Question wording of the variables on which I considered straightlining.

2. How often did you use the Internet in the last 3 months at home, at work, at school/univeristy, elsewhere? (tick an answer for each item)

- At home
- At work (if different from home)
- At school/university
- At other people's houses
- Elsewhere

[scale: 1 every day; 2 a few times a week; 3 once a week; 4 a few times a month; 5 less than once a month; 6 never]

4. How often do you usually eat the following foods? (tick an answer for each item)

- Legumes (dried or canned)
- Potatoes
- Salty snacks (chips, pop corn, pretzels, olives)
- Sweets (cakes, sweet snacks, ice-cream etc.)

[scale: 1 more than once a day; 2 once a day; 3 a few times a week; 4 less than once a week; 5 never]

**APPENDIX 12**. Dichotomous dependent variables for the logistic regression models.

- Watching TV (0=No; 1=Yes)
- Socio-political participation (0=less than 2 'Yes' to question about socio-political activities; 1= at least 2 'Yes' to question about socio-political activities)
- Internet use, frequency in the last 12 months (0=less often; 1=every day)
- Internet use for various activities (0= less than 3 'Yes' to question about Internet activities; 1=at least 3 'Yes' to question about Internet activities)
- Drinking (0=less often or never; 1= at least a few times a week)
- Smoking (0=No; 1=Yes)

**APPENDIX 13**. Behavioural variable distributions from the general population, and the ILC respondents' sample.

| Variables | General population | ILC respondents | |
|---|---|---|---|
| | | No weights | Weights |
| *Watching TV* | | | |
| I do not watch TV | 7.4 | 6.6 | 7.2 |
| I watch TV everyday | 81.5 | 74.6 | 76.6 |
| I watch TV a few days a week | 11.1 | 18.8 | 16.2 |
| (N) | (49,891,896) | (2,003) | (50,118,388) |
| *Socio-political participation in the last 12 months ('yes' answers)* | | | |
| I attended a political meeting | 4.6 | 20.6 | 18.3 |
| (N) | (49,501,523) | (1,997) | (50,035,723) |
| I took part in a political demonstration | 4.1 | 18.9 | 19.3 |
| (N) | (49,441,297) | (1,999) | (50,222,212) |
| I listened to a political debate | 20.5 | 53.4 | 43.9 |
| (N) | (49,412,126) | (1,999) | (50,168,468) |
| I gave money to a political party | 1.9 | 13.2 | 12.9 |
| (N) | (49,449,724) | (1,996) | (50,184,247) |
| I gave money to an association | 15.7 | 45.8 | 38.8 |
| (N) | (49,381,973) | (2,000) | (50,136,852) |
| I did voluntary work | 10.8 | 38.2 | 35.1 |
| (N) | (49,439,831) | (2,001) | (50,242,586) |
| *Internet use in the last 12 months* | | | |
| Every day | 68.6 | 97.0 | 93.8 |
| A few times a week | 24.5 | 2.5 | 4.9 |
| Once a week | 2.1 | 0.4 | 1.2 |
| A few times a month (less than 4) | 3.6 | 0.1 | 0.1 |
| Less than once a month | 1.2 | 0.0 | 0.0 |
| (N) | (29,736,343) | (2,002) | (50,195,511) |
| *Internet use for various activities in the last 3 months ('yes' answers)* | | | |
| Sending or receiving e-mails | 82.5 | 98.7 | 94.4 |
| (N) | (28,464,866) | (1,998) | (50,074,530) |
| Telephoning over the Internet / video calls | 34.5 | 56.2 | 47.2 |
| (N) | (28,275,991) | (1,989) | (50,078,335) |
| Posting messages to chat sites, blogs, etc. | 50.3 | 77.0 | 72.7 |
| (N) | (28,251,063) | (1,986) | (49,951,346) |
| Using instant messaging | 61.4 | 84.8 | 76.6 |
| (N) | (28,208,949) | (1,987) | (49,973,928) |
| Participating in social networks | 57.2 | 80.3 | 78.1 |
| (N) | (28,355,717) | (1,989) | (49,927,965) |

Note: AEL data on the general population are weighted using the weights provided with the dataset by ISTAT.

# References

AAPOR (2016a), «Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys - 9th edition» (https://www.aapor.org/AAPOR_Main/media/publications/Standard-Definitions20169theditionfinal.pdf), link retrieved on 02/10/2018.

AAPOR (2016b), «Standard Definitions. Response Rate Calculator V4.0» (https://www.aapor.org/Standards-Ethics/Standard-Definitions-(1).aspx), link retrieved on 24/09/2018.

Alvarez, R. M., Sherman, R. P., and VanBeselaere, C. (2003), «Subject Acquisition for Web-Based Surveys», *Political Analysis*, 11, 1, pp. 10-48.

Arn, B., Klug, S., and, Kołodziejski, J. (2015), «Evaluation of an Adapted Design in a Multi-Device Online Panel: A DemoSCOPE Case Study», *Methods, Data, Analyses*, 9, 2, pp. 185-212.

Ayrton, R. (2017), «Time for a revival? A historical review of the social survey in Great Britain and the United States», National Centre for Research Methods (NCRM). Methodological Review paper (http://eprints.ncrm.ac.uk/3999/1/Time%20for%20a%20revival%20-%20A%20historical%20review%20of%20the%20social%20survey%20in%20Great%20Britain%20and%20the%20United%20States.pdf), link retrieved on 03/09/2018.

Baker, R., Blumberg, S., Brick, J. M., Couper, M. P., Courtright, M. et al. (2010a), «Research Synthesis: AAPOR Report on Online Panels», *Public Opinion Quarterly*, 74, 4, pp. 711-781.

Baker, R., Blumberg, S., Brick, J. M., Couper, M. P., Courtright, M.et al. (2010b), «AAPOR Report on Online Panels» (https://www.aapor.org/Education-Resources/Reports/Report-on-Online-Panels), link retrieved on 19/10/2018.

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P. (2013), «Report of the AAPOR Task Force on Non-Probability Sampling» (http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf), link retrieved on 27/10/2018.

Baker, R., Miller, C., Kachhi, D., Lange, K., Wilding-Brown, L. (2014), «Validating respondents' identity in online samples: The impact of efforts to eliminate fraudulent respondents», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 441-456.

Batinic, B., and Moser, K. (2005), «Determinanten der Rücklaufquote in Online-Panels [Determinants of response rates in online panels]», *Zeitschrift für Medienpsychologie*, 17, pp. 64-74.

Beierlein, C., Kuntz, A., and Davidov, E. (2016), «Universalism, conservation and attitudes toward minority groups.», *Social Science Research*, 58, pp. 68-79.

Bethlehem, J. (2010), «Selection Bias in Web Surveys», *International Statistical Review*, 78, 2, pp.  161-188.

Bethlehem, J. (2015), «Web Surveys in Official Statistics», in U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis (eds.), *Improving Survey Methods. Lessons from Recent Research*, New York, Routledge, pp. 156-169.

Bethlehem, J., and Biffignandi, S. (2012), *Handbook of Web Surveys*, Chichester, Wiley.

Bethlehem, J., Cobben, F., and Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*, Hoboken, Wiley.

Biemer, P. P. (2001), «Nonresponse Bias and Measurement Bias in a Comparison of Face to Face and Telephone Interviewing», *Journal of Official Statistics*, 17, 2, pp. 295-320.

Biemer, P. P. (2010), «Total Survey Error: Design, Implementation, and Evaluation», *Public Opinion Quarterly*, 74, 5, pp. 817-848.

Binswanger, J., Schunk, D., and Toepoel, V. (2013), «Panel Conditioning in Difficult Attitudinal Questions», *Public Opinion Quarterly*, 77, 3, pp. 783-797.

Blom, A. G., Gathmann, C., and Krieger, U. (2015), «Setting Up an Online Panel Representative of the General Population The German Internet Panel», *Field Methods*, 27, 4, pp. 391-408.

Blom, A. G., Herzing, J. M. E., Cornesse, C., Sakshaug, J., Krieger, U., and Bossert, D. (2017). «Does the Recruitment of Offline Households Increase the Sample Representativeness of Probability-Based Online Panels? Evidence from the German Internet Panel», *Social Science Computer Review*, 35, 4, pp. 498-520.

Blumberg, S. J., and Luke, J. V. (2017), «Wireless Substitution: Early Release of Estimates From the National Health Interview Survey, July–December 2017», Early Release Report, National Center for Health Statistics. (https://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless201806.pdf), link retrieved on 26/10/2018.

Bosch, O. J., Revilla, M., DeCastellarnau, A., and Weber, W. (2018), «Measurement Reliability, Validity, and Quality of Slider Versus Radio Button Scales in an Online Probability-Based Panel in Norway», *Social Science Computer Review*, pp. 1-14.

Bosnjak, M., Haas, I., Galesic, M., Kaczmirek, L., Bandilla, W., and Couper, M. P. (2013), «Sample Composition Discrepancies in Different Stages of a Probability-Based Online Panel», *Field Methods*, 25, 4, pp. 339-360.

Breton, C., Cutler, F., Lachance, S., and Mierke-Zatwarnicki, A. (2017), «Telephone versus Online Survey Modes for Election Studies: Comparing Canadian Public Opinion and Vote Choice in the 2015 Federal Election», *Canadian Journal of Political Science*, 50, 4, pp. 1005-36.

Brick, J. M., and Williams, D. (2013), «Explaining Rising Nonresponse Rates in Cross-Sectional Surveys», *The Annals of the American Academy of Political and Social Science*, 645, 1, pp. 36-59.

Brown, G., Weber, D., Zanon, D., and de Bie, K. (2012), «Evaluation of an Online (Opt-in) Panel for Public Participation Geographic Information Systems Surveys», *International Journal of Public Opinion Research*, 24, 4, pp. 534-545.

Buskirk, T. D., and Andrus, C. H. (2014), «Making Mobile Browser Surveys Smarter: Results from a Randomized Experiment Comparing Online Surveys Completed via Computer or Smartphone», *Field Methods* 26, 4, pp. 322-342.

Buskirk T. D., and Dutwin, D. (2016), «Telephone sample surveys: dearly beloved or nearly departed? Trends in errors from dual frame and cell phone RDD surveys in the age of declining response rates», presented at the *Annual Conference AAPOR*, May 12-15, Austin, TX.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A. S., Krosnick, J. A., and P. J. Lavrakas (eds.) (2014), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley.

Callegaro, M., Baker, R., Bethlehem, J., Göritz, A.S., and Lavrakas, P.J. (2014), «Online panel research. History, concepts, applications and a look at the future», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 1-22.

Callegaro, M., and DiSogra, C. (2008), «Computing Response Metrics for Online Panels», *Public Opinion Quarterly*, 72, 5, pp. 1008-1032.

Callegaro, M., Lozar Manfreda, K., and Vehovar, V. (2015), *Web Survey Methodology*, London, SAGE.

Callegaro, M., Villar, A., Yeager, D., and Krosnick, J. A. (2014), «A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 23-53.

Chang, L., and Krosnick, J.A. (2009), «National Surveys via RDD Telephone Interviewing Versus the Internet Comparing Sample Representativeness and Response Quality», *Public Opinion Quarterly*, 73, 4, pp. 641-678.

Cho, M. K., Magnus, D., Constantine, M., Soo-Jin Lee, S., Kelley, M., Alessi, S., Korngiebel, D., et al. (2015), «Attitudes Toward Risk and Informed Consent for Research on Medical Practices: A Cross-Sectional Survey», *Annals of Internal Medicine*, 162, 10, pp. 690-696.

Comley, P. (2007), «Online Market Research», in M. van Hamersveld, and C. de Bont (eds.), *Market Research Handbook* (5th edition), Chichester, Wiley, pp. 401-420.

Couper, M. P. (2000), «Web Surveys. A Review of Issues and Approaches», *Public Opinion Quarterly*, 64, 4, pp. 464-494.

Couper, M. P. (2005), «Technology Trends in Survey Data Collection», *Social Science Computer Review*, 23, 4, pp. 486-501.

Couper, M. P. (2008), *Designing Effective Web Surveys*, New York, Cambridge University Press.

Couper, M. P. (2017), «New Developments in Survey Data Collection», *Annual Review of Sociology*, 43, 1, pp. 121-145.

Couper, M. P., and Bosnjak, M. (2010), «Internet Surveys», in P. V. Marsden, and J. D. Wright (eds.), *Handbook of Survey Research* (2nd edition), Bingley, Emerald, pp. 527-550.

Couper, M. P., Kapteyn, A., Schonlau, M., and Winter, J. (2007), «Noncoverage and Nonresponse in an Internet Survey», *Social Science Research*, 36, pp. 131-148.

Couper, M. P., Tourangeau, R., Conrad, F. G., and Zhang, C. (2013), «The Design of Grids in Web Surveys», *Social Science Computer Review* 31, 3, pp. 322-345.

Craig, B. M., Hays, R. D., Pickard, A. S., Cella, D., Revicki, D. A., and Reeve, B. B. (2013), «Comparison of US Panel Vendors for Online Surveys», *Journal of Medical Internet Research* 15, pp. 11 e260.

Das, M., Ester, P., and Kaczmirek, L. (eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, New York, Routledge.

De Leeuw, E. D. (2018), «Mixed-Mode: Past, Present, and Future», *Survey Research Methods*, 12, 2, pp. 75-89.

De Leeuw, E. D., and Berzelak, N. (2016), «Survey Mode or Survey Modes?», in C. Wolf, D. Joye, T. W. Smith, and Y.-C. Fu (eds.), *The Sage Handbook of Survey Methodology*, Los Angeles, SAGE, pp. 142-156.

De Leeuw, E. D., and Hox, J. J. (2015), «Survey Mode and Mode Effects», in U. Engel, B. Jann, P. Lynn, A. Scherpenzeel, and P. Sturgis (eds.), *Improving Survey Methods. Lessons from Recent Research*, New York, Routledge, pp. 22-34.

De Leeuw, E., and Toepoel, V. (2018), «Mixed-Mode and Mixed-Device Surveys», in D.L. Vannette, and J.A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, London, Palgrave Macmillan, pp. 51-61.

Dillman, Don A. (1978), *Mail and Telephone Surveys: The Total Design Method*, New York, Wiley.

Dillman, Don A. (2000), *Mail and Internet Surveys: The Tailored Design Method*, New York, Wiley.

Dillman, Don A., Smyth, J. D., and Christian, L. M. (2014), *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*, New York, Wiley.

Drewes, F. (2014), «An empirical test of the impact of smartphones on panel-based online data collection», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 367-386.

Duffy, B., Smith, K., Terhanian, G., and Bremer, J. (2005), «Comparing Data from Online and Face-to-Face Surveys», *International Journal of Market Research*, 47,6, pp. 615-639.

Dutwin, D., and Buskirk, T. D. (2017), «Apples to Oranges or Gala versus Golden Delicious?», *Public Opinion Quarterly*, 81 (S1), pp. 213-239.

Dutwin, D., and Lavrakas, P. J. (2016), «Trends in telephone outcomes 2008-2015», *Survey Practice*, 9, 3, https://www.surveypractice.org/article/2808-trends-in-telephone-outcomes-2008-2015.

Eckman, S. (2016), «Does the Inclusion of Non-Internet Households in a Web Panel Reduce Coverage Bias?», *Social Science Computer Review*, 34, 1, pp. 41-58.

Elliott, M. R., and Valliant, R. (2017), «Inference for Nonprobability Samples», *Statistical Science*, 32, 2, pp. 249-264.

Engel, U., Jann, B., Lynn, P., Scherpenzeel, A., and Sturgis, P. (eds.) (2015), *Improving Survey Methods: Lessons From Recent Research*, European Association of Methodology Series, New York, Routledge.

Erens, B., Burkill, S., Couper, M. C., Conrad, F., Clifton, S., Tanton, C., Phelps, A., et al. (2014), «Nonprobability Web Surveys to Measure Sexual Behaviors and Attitudes in the General Population: A Comparison With a Probability Sample Interview Survey», *Journal of Medical Internet Research*, 16, 12, e276.

EU (2018), «Broadband Coverage in Europe 2017» (https://ec.europa.eu/digital-single-market/en/news/study-broadband-coverage-europe-2017), link retrieved on 19/12/2018.

Faasse, J. (2005), «Panel proliferation and quality concerns», in Proceedings of *ESOMAR Conference on Worldwide Panel Research: Developments and Progress,* Budapest, Hungary, pp. 159-169.

Frippiat, D., and Marquis, N. (2010), «Web surveys in the social sciences: An overview», *Population*, 65, 2, pp. 285-312.

Gobo, G., and Mauceri, S. (2014), *Constructing Survey Data: An Interactional Approach*, SAGE.

Goldenbeld, C., and de Craen, S. (2013), «The Comparison of Road Safety Survey Answers between Web-Panel and Face-to-Face; Dutch Results of SARTRE-4 Survey», *Journal of Safety Research*, 46 (September), pp. 13-20.

Göritz, A. S. (2009), «Using Online Panels in Psychological Research», *Oxford Handbook of Internet Psychology*, February, https://doi.org/10.1093/oxfordhb/9780199561803.013.0030.

Göritz, A. S. (2014), «Determinants of the starting rate and the completion rate in online panel studies», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 154-170.

Göritz, A. S., and Luthe, S. C. (2013a), «Effects of Lotteries on Response Behavior in Online Panels», *Field Methods*, 25, 3, pp. 219-237.

Göritz, A. S., and Luthe S. C. (2013b), «How Do Lotteries and Study Results Influence Response Behavior in Online Panels?», *Social Science Computer Review*, 31, 3, pp. 371-385.

Göritz, A. S., and Luthe, S. C. (2013c), «Lotteries and Study Results in Market Research Online Panels», *International Journal of Market Research*, 55, 5, pp. 611-626.

Göritz, A. S., and Moser, K. (2000), «Repräsentativität im Online-Panel [Representativeness in online panels]», *Der Markt*, 155, pp. 156-162.

Greszki, R., Meyer, M., and Schoen, H. (2014), «The impact of speeding on data quality in nonprobability and freshly recruited probability-based online panels», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 238-262.

Groves, R.  M. (2006), «Nonresponse Rates and Nonresponse Bias in Household Surveys», *Public Opinion Quarterly*, 70, 5, pp. 646-675.

Groves, R. M., Fowler, F. J., Couper, M. P., Lepkowski, J. M., Singer, E., and Tourangeau, R. (2009), *Survey Methodology* (2nd edition), New York, Wiley.

Groves, R. M., and Kahn, R. L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*, New York, Academic.

Groves, R. M., and Peytcheva, E. (2008), «The Impact of Nonresponse Rates on Nonresponse Bias», *Public Opinion Quarterly*, 72, 2 pp. 167-189.

Häder, S., Häder, M., and Kühne, M. (eds.), (2012), *Telephone Survey in Europe. Research and Practice*, Berlin, Springer.

Heen, M. S. J., Lieberman, J. D., and Miethe, T. D. (2014), «A Comparison of Different Online Sampling Approaches for Generating National Samples», Center for Crime and Justice Policy (CCJP). (https://www.unlv.edu/sites/default/files/page_files/27/ComparisonDifferentOnlineSampling.pdf), link retrieved on 27/10/208.

Heerwegh, D. (2009), «Mode differences between face-to-face and web surveys: an experimental investigation of data quality and social desirability effects», *International Journal of Public Opinion Research*, 21, 1, pp. 111-121.

Hewson, C., Vogel, C., and Laurent D. (2016), *Internet Research Methods*, London, SAGE.

Hillygus, D. S., Jackson, N., and Young, M. (2014), «Professional respondents in nonprobability online panels», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective*, Chichester, Wiley, pp. 219-237.

Holbrook, A. L., Green, M. C., and Krosnick, J. A. (2003), «Telephone versus face-to-face interviewing of national probability samples with long questionnaires: Comparisons of respondent satisficing and social desirability response bias», *The Public Opinion Quarterly*, 67, 1, pp. 79-125.

Holbrook, A. L., Krosnick, J. A., and Pfent, A. (2007), «The Causes and Consequences of Response Rates in Surveys by the News Media and Government Contractor Survey Research Firms», in J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, and R. L. Sangster (eds.), *Advances in Telephone Survey Methodology*, Hoboken, Wiley, pp. 499-528.

Hox, J. J., De Leeuw, E. D., and Klausch, T. (2017), «Mixed mode Research: Issues in Design and Analysis», in P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N.C. Tucker, and B. T. West (eds.), *Total Survey Error in Practice*, New York, Wiley, pp. 511-530.

Hox, J. J., De Leeuw, E. D, and Zijlmans, E. A. O. (2015), «Measurement Equivalence in Mixed Mode Surveys», *Frontiers in Psychology*, 6, 87, https://doi.org/10.3389/fpsyg.2015.00087.

International Organization for Standardization (2009), ISO 26362:2009, *Access Panels in Market, Opinion, and Social Research – Vocabulary and Service Requirements*, Geneva, ISO.

Internet World Stats (2018), «Internet Usage Statistics. The Internet Big Picture» (https://www.internetworldstats.com/stats.htm), link retrieved on 27/10/2018.

Joinson, A. (1999), «Social desirability, anonymity, and Internet-based questionnaires», *Behavior Research Methods, Instruments, & Computers*, 31, 3, pp. 433-438.

Jones, M. S., House, L. A., and Gao, Z. (2015), «Respondent Screening and Revealed Preference Axioms», *Public Opinion Quarterly*, 79, 3, pp. 687-709.

Kennedy, C., Mercer, A., Keeter, S., Hatley, N., McGeeney, K., and Gimenez, A. (2016), «Evaluating Online Nonprobability Surveys», Pew Research Center. (http://www.pewresearch.org/2016/05/02/evaluating-online-nonprobability-surveys/), link retrieved on 28/10/2018.

Keusch, F. (2013), «The Role of Topic Interest and Topic Salience in Online Panel Web Surveys», *International Journal of Market Research*, 55, 1, pp. 59-80.

Keusch, F., Batinic, B., Mayerhofer, W. (2014), «Motives for joining nonprobability online panels and their association with survey participation behavior», in M. Callegaro, R. Baker, J. Bethlehem, A. S. Göritz, J. A. Krosnick, and P. J. Lavrakas (eds.), *Online Panel Research: A Data Quality Perspective,* Chichester, Wiley, pp. 171-191.

Knapton, K., and Myers, S. (2005), «A Study of Non-Response Patterns», Articles, Quirks.Com (https://www.quirks.com/articles/a-study-of-non-response-patterns), link retrieved on 26/10/2018.

Krosnick, J. (1991), «Response strategies for coping with the cognitive demands of attitude measures in surveys», *Applied Cognitive Psychology*, 5, 3, pp.213-236.

Langer, G. (2018), «Probability Versus Non-Probability Methods», in D. L. Vannette, and J. A. Krosnick (eds.), *The Palgrave Handbook of Survey Research*, London, Palgrave Macmillan, pp. 351-362.

Lavrakas, P. J. (1997), «Methods for Sampling and Interviewing in Telephone Surveys», in. L. Bickman, and D. J. Rog (eds.), *Handbook of Applied Social Research Methods*, Thousand Oaks, CA, SAGE, pp. 429-472.

Lavrakas, P. J., Benson, G., Blumberg, S., Buskirk, T., Cervantes, I. F., Christian, L. et al. (2017), «Report from the AAPOR Task Force on the Future of U.S. General Population Telephone Survey Research», AAPOR. (https://www.aapor.org/Education-Resources/Reports/The-Future-Of-U-S-General-Population-Telephone-Sur.aspx), link retrieved on 04/09/2018.

Lee, C., Back, K., Williams, R. J., and Ahn, S. (2015), «Comparison of Telephone RDD and Online Panel Survey Modes on CPGI Scores and Co-Morbidities», *International Gambling Studies*, 15, 3, pp. 435-449.

Lee, R. M., Fielding, N. G., and Blank, G. (2017), «Online Research Methods in the Social Sciences: An Editorial Introduction», in N. G. Fielding, R. M. Lee, and G. Blank (eds.), *The SAGE Handbook of Online Research Methods* (2nd edition), SAGE, pp. 3-16.

Legleye, S., Charrance, G., Razafindratsima, N., Bajos, N., Bohet, A., Moreau, C., and the Fecond research Team (2015), «The use of a non-probability Internet panel to monitor sexual and reproductive health in the general population», *Sociological Methods & Research*, 47, 2, pp. 314-348.

Lipps, O., Pekari, N., and Roberts, C. (2015), «Undercoverage and Nonresponse in a List-sampled Telephone Election Survey», *Survey Research Methods*, 9, 2, pp. 71-82.

Lozar Manfreda, K., and Vehovar, V. (2007), «Web survey methodology (WebSM) portal», in V. Reynolds, and A. Rodney (eds.), *Handbook of research on electronic surveys and measurements*, Hershey, Idea Group Reference, pp. 248-252.

Lugtig, P. (2014), «Panel Attrition Separating Stayers, Fast Attriters, Gradual Attriters, and Lurkers», *Sociological Methods & Research*, 43, 4, pp. 699-723.

Lugtig, P., and Toepoel, V. (2016), «The Use of PCs, Smartphones, and Tablets in a Probability-Based Panel Survey Effects on Survey Measurement Error», *Social Science Computer Review*, 34, 1, pp. 78-94.

Malhotra, N., and Krosnick, J. A. (2007), «The Effect of Survey Mode and Sampling on Inferences about Political Attitudes and Behavior: Comparing the 2000 and 2004 ANES to Internet Surveys with Nonprobability Samples», *Political Analysis*, 15, 3, pp. 286-323.

Marsden, P. V., and Wright, J. D. (2010), «Survey Research and Social Science: History, Current Practice, and Future Prospects», in P. V. Marsden and J. D. Wright (eds.), *Handbook of Survey Research* (2nd edition), Bingley, Emerald, pp. 3-26.

Matthijsse, S. M., de Leeuw, E. D., and Hox, J. J. (2015), «Internet Panels, Professional Respondents, and Data Quality», *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 11, 3, pp. 81-88.

Mavletova, A. (2013), «Data Quality in PC and Mobile Web Surveys», *Social Science Computer Review*, 31, 6, pp. 725-743.

Mavletova, A., and Couper, M. P. (2013), «Sensitive Topics in PC Web and Mobile Web Surveys: Is There a Difference?», *Survey Research Methods*, 7, 3, pp. 191-205.

Mercer, A. W., Kreuter, F., Keeter, S., and Stuart, E. A. (2017), «Theory and Practice in Nonprobability Surveys», *Public Opinion Quarterly*, 81 (SI).

Miller, P. V. (2017), «Is There a Future for Surveys?», *Public Opinion Quarterly*, 81 (SI).

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., The PRISMA Group (2009), «Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement», *PLoS Med*, 6, 7, e1000097.

Mohorko, A., de Leeuw, E., and Hox, J. (2013), «Coverage Bias in European Telephone Surveys: Developments of Landline and Mobile Phone Coverage across Countries and over Time», *Survey Methods: Insights from the Field*, January 2013, pp. 828-841.

Mueller, K., Straatmann, T., Hattrup, K., and Jochum, M. (2014), «Effects of Personalized Versus Generic Implementation of an Intra-Organizational Online Survey on Psychological Anonymity and Response Behavior: a Field Experiment», *Journal of Business and Psychology*, 29, pp. 169-181.

Natale, P. (2004), *Il sondaggio*, Bari, Laterza.

Nedelec, J. L. (2018), «Individual differences and co-occurring victimization online and offline: The role of impulsivity», *Personality and Individual Differences*, 133, pp. 77-84.

Neyman, J. (1934), «On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection», *Journal of the Royal Statistical Society*, 97, 4, pp. 558-625.

Nielsen, J. S., and Kjær. T. (2011), «Does Question Order Influence Sensitivity to Scope? Empirical Findings from a Web-Based Contingent Valuation Study», *Journal of Environmental Planning and Management*, 54, 3, pp. 369-381.

Pedersen, M. J., and Nielsen, C. V. (2016), «Improving Survey Response Rates in Online Panels Effects of Low-Cost Incentives and Cost-Free Text Appeal Interventions», *Social Science Computer Review*, 34, 2, pp. 229-243.

Pennay, D. W., Neiger, D., Lavrakas, P. J., and Borg, K. (2018), «The Online Panels Benchmarking Study: a Total Survey Error comparison of findings from probability-based surveys and nonprobability online panel surveys in Australia», CSRM & SRC Methods Paper, 2.

Peytchev, A. (2013), «Consequences of Survey Nonresponse», *The ANNALS of the American Academy of Political and Social Science*, 645, 1, pp. 88-111.

Postoaca, A. (2006), *The Anonymous Elect. Market Research through Online Access Panels*, Berlin, Springer.

Revilla, M., and Ochoa, C. (2015), «What are the Links in a Web Survey Among Response Time, Quality, and Auto-Evaluation of the Efforts Done?», *Social Science Computer Review*, 33, 1, pp. 97-114.

Revilla, M., Saris, W., Loewe, G., and Ochoa, C. (2015), «Can a Non-Probabilistic Online Panel Achieve Question Quality Similar to That of the European Social Survey?», *International Journal of Market Research*, 57, 3, pp. 395-412.

Rivers, D. (2013), «Comment», *Journal of Survey Statistics and Methodology*, 1, pp. 111-117.

Roßmann, J., Gummer, T., and Silber, H. (2018), «Mitigating Satisficing in Cognitively Demanding Grid Questions: Evidence from Two Web-Based Experiments», *Journal of Survey Statistics and Methodology*, 6, 3, pp. 376-400.

Rookey, B.D., Hanway, S., and Dillman, D.A. (2008), « Does a Probability-based Household Panel Benefit from Assignment to Postal Response as an Alternative to Internet-only?», *Public Opinion Quarterly*, 72, pp. 962-984.

Sala, E., and Lillini, R. (2015), «Undercoverage Bias in Telephone Surveys in Europe: The Italian Case», *International Journal of Public Opinion Research*, 29, 1, pp. 133-156.

Scherpenzeel, A. C., and Bethlehem, J. G. (2011), «How Representative Are Online Panels? Problems of Coverage and Selection and Possible Solutions», in M. Das, P. Ester, and L. Kaczmirek (eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, New York, Routledge, pp. 105-132.

Schonlau, M. (2015), «What Do Web Survey Panel Respondents Answer When Asked 'Do You Have Any Other Comment?'», *Survey Methods: Insights from the Field*, November 2015, pp. 6899-6906.

Schonlau, M., and Couper, M. P. (2017), «Options for Conducting Web Surveys», *Statistical Science*, 32, 2, pp. 279-292.

Sell, R., Goldberg, S., and Conron, K. (2015), «The Utility of an Online Convenience Panel for Reaching Rare and Dispersed Populations: E0144011», *PLoS One*, 10, 12, e0144011. http:// doi:10.1371/journal.pone.0144011.

Simons, A. M. W., Koster, A., Groffen, D. A. I., and Bosma, H. (2017), «Perceived Classism and Its Relation with Socioeconomic Status, Health, Health Behaviours and Perceived Inferiority: The Dutch Longitudinal Internet Studies for the Social Sciences (LISS) Panel», *International Journal of Public Health*, 62, 4, pp. 433-440.

Smith, S. M., Roster, C. A., Golden, L. L., and Albaum, G. S. (2016), «A Multi-Group Analysis of Online Survey Respondent Data Quality: Comparing a Regular USA Consumer Panel to MTurk Samples», *Journal of Business Research*, 69, 8, pp. 3139-3148.

Smyth, J. D., and Pearson, J. E. (2011), «Internet Survey Methods: A Review of Strengths, Weaknesses, and Innovations», in M. Das, P. Ester, and L. Kaczmirek (eds.), *Social and Behavioral Research and the Internet: Advances in Applied Methods and Research Strategies*, New York, Routledge, pp. 11-44.

Sterrett, D., Malato, D., Benz, J., Tompson, T., and English, N. (2017), «Assessing Changes in Coverage Bias of Web Surveys in the United States», *Public Opinion Quarterly*, 81, (SI), pp. 338-356.

Struminskaya, B., Weyandt, K., and Bosnjak, M. (2015), «The Effects of Questionnaire Completion Using Mobile Devices on Data Quality. Evidence from a Probability-Based General Population Panel», *Methods, Data, Analyses*, 9, 2, pp. 261-292.

Toepoel, V. (2015), *Doing Surveys Online*, London, SAGE.

Toepoel, V., and Hendriks, Y. (2016), «The Impact of Non-Coverage in Web Surveys in a Country with High Internet Penetration: Is It (Still) Useful to Provide Equipment to Non-Internet Households in the Netherlands?», *International Journal of Internet Science*, 11, 1, pp. 33-50.

Tourangeau, R. (2017), «Mixing Modes: Tradeoffs among Coverage, Nonresponse, and Measurement Error», in P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N.C. Tucker, and B. T. West (eds.), *Total Survey Error in Practice*, New York, Wiley.

Tourangeau, R., Conrad, F. G., and Couper, M. P. (2013), *The Science of Web Surveys*, Oxford, Oxford University Press.

Tourangeau, R., and Plewes, T. J. (2013), *Nonresponse in Social Science Surveys: A Research Agenda*, Washington DC, The National Academies Press.

Tsuboi, S., Yoshida, H., Ae, R., Kojo, T., Nakamura, Y., and Kitamura, K. (2015), «Selection Bias of Internet Panel Surveys A Comparison with a Paper-Based Survey and National Governmental Statistics in Japan», *Asia-Pacific Journal of Public Health*, 27, 2, pp. 2390-2399.

Tucker, C., and Lepkowski, J. M. (2008) «Telephone Survey Methods: Adapting to Change», in J. M. Lepkowski, C. Tucker, J. M. Brick, E. D. De Leeuw, L. Japec, P. J. Lavrakas, M. W. Link, R. L. Sangster (eds.), *Advances in Telephone Survey Methodology*, Hoboken, Wiley, pp. 3-26.

Valliant, R., and Dever. J. A. (2018), *Survey Weights: A Step-by-Step Guide to Calculation*, College Station, Texas, Stata Press.

Vandenplas, C., Loosveldt, G., and Vannieuwenhuyze, J. T. A. (2016), «Assessing the Use of Mode Preference as a Covariate for the Estimation of Measurement Effects between Modes. A Sequential Mixed Mode Experiment», *Methods, Data, Analyses*, 10, 2, pp. 119-142.

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2010), «A Method for Evaluating Mode Effects in Mixed-Mode Surveys», *Public Opinion Quarterly*, 74, 5, pp. 1027-1045.

Vonk, T., van Ossenbruggen, R., and Willems, P. (2008), «A comparison study across 19 online panels (NOPVO 2006)», in I. A. L. Stoop, and M. Wittenberg (eds.), *Access Panels and Online Research, Panacea or Pitfall?: Proceedings of the DANS Symposium, Amsterdam, October 12th 2006*, Amsterdam: Uitgeverij Aksant, pp. 53-77.

Yeager, D. S., Krosnick, J. A., Chang, L. C., Javitz, H. S., Levendusky, M. S., Simpser, A., and Wang, R. (2011), «Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples», *Public Opinion Quarterly*, 75, 4, pp. 709-747.

# Acknowledgements

I would like to thank my tutor, Emanuela Sala, who introduced me to the survey-methodology field and to the international research community, guided me during my PhD programme, and suggested the initial idea of my thesis.

I would also like to thank my co-tutor, Katja Lozar Manfreda (University of Ljubljana, Ljubljana - Slovenia) for her teachings about new technologies in survey research, for the lively discussions with her research team (in particular, with Gregor Čehovin), and for her much appreciated support in preparing the systematic review on online panels.

Special thanks go to my colleague Marco Gui (Department of Sociology and Social Research, University of Milano-Bicocca), Angelo Tomaselli and his research team, and in particular Beatrice Bartoli (Demetra opinioni.net s.r.l.). I would like to thank them all for their enthusiastic support, valuable advice and collaboration that made my study possible.

I am also grateful for the important feedbacks received from my external evaluators, Malcolm Williams (Cardiff University, Cardiff - Galles) and Wander van der Vaart (University of Humanistic Studies, Utrecht - The Netherlands), who provided very useful suggestions to improve my thesis and expressed their appreciation for my work.

And finally a warm thank you to my beloved family and friends, who have always been glad to share every little accomplishment with me.

# Funding

My research was funded by the Doctoral School and the Department of Sociology and Social Research, both based at the University of Milano-Bicocca, and by the “Innovation Project Grant” supported by Fastweb S.p.A. An additional financial support was provided by the International Mobility Programme - Erasmus+ that cofinanced my visiting semester at the Faculty of Social Sciences, University of Ljubljana.