City Research Online

City St George's, University of London

Citation: Callegaro, M., Villar, A., Krosnick, J. & Yeager, D. (2014). A Critical Review of Studies Investigating the Quality of Data Obtained With Online Panels. In: Callegaro, M., Baker, R., Bethlehem, J., Goritz, A., Krosnick, J. & Lavrakas, P. (Eds.), Online Panel Research: A Data Quality Perspective. (pp. 23-53). UK: John Wiley & Sons. ISBN 978-1-119-94177-4 doi: 10.1002/9781118763520.ch2

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/14350/

Link to published version: https://doi.org/10.1002/9781118763520.ch2

Copyright and Reuse: Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to City Research Online policy.

City Research Online: http://openaccess.city.ac.uk/ publications@citystgeorges.ac.uk

2

A critical review of studies investigating the quality of data obtained with online panels based on probability and nonprobability samples[1]

Mario Callegaro[a], Ana Villar[b], David Yeager[c], and Jon A. Krosnick[d]

[a]*Google UK*

[b]*City University London, UK*

[c]*University of Texas at Austin, USA*

[d]*Stanford University, USA*

2.1 Introduction

Online panels have been used in survey research as data collection tools since the late 1990s (Postoaca, 2006). The potential great cost and time reduction of using these tools have made research companies enthusiastically pursue this new mode of data collection.

[1] We would like to thank Reg Baker and Anja Göritz for their useful comments on preliminary versions of this chapter.

The vast majority of these online panels were built by sampling and recruiting respondents through nonprobability methods such as snowball sampling, banner ads, direct enrollment, and other strategies to obtain large samples at a lower cost (see Chapter 1). Only a few companies and research teams chose to build online panels based on probability samples of the general population. During the 1990s, two probability-based online panels were documented: the CentER data Panel in the Netherlands and the Knowledge Networks Panel in the United States. Since then, a few probability panels started in the 2000s, including the Face-to-Face-Recruited-Internet-Platform (FFRISP) and the American Life Panel in the United States, the Longitudinal Internet Studies for the Social Sciences (LISS) in the Netherlands (Callegaro & DiSogra, 2008), and a handful of new panels are being built in European countries, including Germany,[2] France[3] (Das, 2012), Norway, and Sweden (Martinsson, Dahlberg, & Lundmark, 2013).

In the minds of many is the question: how do online panels of nonprobability samples compare in terms of quality to online panels of probability samples? The reasons why many online panels were built using nonprobability sampling and recruitment methods stem from methodological as well as financial reasons and are discussed in Chapter 1. In this chapter, we review a set of studies comparing survey estimates obtained from online panels to estimates from other data collection methods in order to assess the quality of the former, capitalizing on more than a decade's worth of studies and experiments.

We aim to provide data-driven answers to four main research questions:

1. How accurate are point estimates computed from online panels of probability and nonprobability samples?
2. How useful are weighting procedures in improving accuracy of these estimates?
3. How do relationships and predictive relations of data collected from online panels of probability and nonprobability samples compare to benchmark surveys?
4. How do experiments on online panels of probability and nonprobability samples replicate over time and across panels?

2.2 Taxonomy of comparison studies

The existing studies comparing statistics from online panels of nonprobability samples to other sources differ with respect to whether the comparison is made against surveys using probability or nonprobability samples, their mode of data collection, and whether benchmark estimates are available. We found six types of designs in the literature depending on these aspects (Table 2.1). These designs are not mutually exclusive; many studies use a combination of two or more designs, for example, an online panel from a nonprobability sample can be compared against an online panel and a telephone survey both using probabilistic sampling.

Next, each type of design will be described, together with their strengths and weaknesses:

Design 1: Comparison of two online panels with nonprobability samples. Design number 1 has the advantage of keeping the mode of data collection constant (online) and possibly

[2] http://reforms.uni-mannheim.de/english/internet_panel/home/index.html; http://www.gesis.org/en/services/data-collection/.

[3] http://www.sciencespo.fr/dime-shs/content/dime-shs-web.

Table 2.1 Possible designs used in studies comparing nonprobability online panels results to other results collected in a different way.

Design	Reference study	Comparison study	Mode	Benchmarks
1	Online panel with a nonprobability sample	Online panel with a nonprobability sample	Self-administered, Online	Yes – No
2	Online panel with a nonprobability sample	Online panel with a probability sample	Self-administered, Online	Yes – No
3	Online panel with a nonprobability sample	Telephone cross-sectional survey with a probability sample	Interviewer, Telephone	Yes – No
4	Online panel with a nonprobability sample	Face-to-face cross-sectional survey with a probability sample	Interviewer, Face-to-Face	Yes – No
5	Online panel with a nonprobability sample	Mail cross-sectional survey with a probability sample	Self-administered	Yes – No
6	Online panel with a nonprobability sample	Same online panel with a nonprobability sample	Self-administered, Online	No

the questionnaire administration constant. Three alternatives for questionnaire administration are possible: (a) each panel redirects their sample to a third party site where the survey is taken; (b) each panel programs and hosts the survey itself; and (c) a survey is centrally located and administered but the look and feel of the questionnaire are specific to each panel provider. In the first case, we have the purest case from an experimental point of view because the visual design of the instrument, the instructions, the prompts and real-time checks are the same for every respondent. However, redirecting panel members to another third party site can introduce nonresponse bias difficult to quantify because some panel members can be reluctant to complete the survey on a site that is not the panel site they belong to. In the second case, the same questionnaire is programmed individually by each panel provider. With this strategy, panel members see the survey on the same site they are familiar with, experiencing the look and feel they are used to. Design 1 allows direct comparison across panels but in order to assess accuracy of each panel, external benchmarks or other forms of data validation need to be available. This is also the case for the other five designs encountered in the literature.

Design 2: Comparison of an online panel with a nonprobability sample to an online panel with a probability sample. Design 2 allows comparison of online panels with different sampling designs, while keeping the mode of data collection constant. This design is similar to design 1, but there are usually a number of restrictions associated with the way probability-based online panels are run: (a) members are typically not allowed to be redirected to other website for survey completion; and (b) in connection with this, surveys are typically programmed in-house. When using design 2, it will be necessary to decide whether or not to include households from the probability-based online panels that did not have Internet at the moment of recruitment and were provided with a device and Internet connection for the study, given that such households would not be, in general, part of the online panels from nonprobability samples.

Design 3 and Design 4: Comparison of an online panel with a nonprobability sample to a face-to-face or a telephone survey with probability sample. These two modes are interviewer-administered and the questions are generally presented to the respondent orally (with the possible addition of show cards to present response options and other materials). As a consequence, any differences could be due to measurement effects as well as coverage, sampling, or differential nonresponse error. Therefore, when comparing results, possible mode effects need to be taken into account.

Design 5: Comparison of an online panel with a nonprobability sample to a mail survey with a probability sample. We found fewer examples of design 5 among the reviewed studies; however, this design has the strength of keeping the mode of administration (self-administered) closer across survey implementations than designs 3 and 4. At the same time, mode effects in mail and web surveys are also possible due to differences in visual design.

Design 6: Replication within panel. Design 6 is very different in nature and has a distinctive goal. Here the same questionnaire is implemented on non-overlapping cross-sectional samples of the same nonprobability-based online panel at different points in time. The questionnaire is generally comprised of questions that are not subject to rapid change and the time across the different administration is usually kept reasonably short (Gittelman & Trimarchi, 2010). The goal of this design is to test if a panel is "deteriorating" in

any way. The hypothesis behind it is that if the quality of the panel is good, the results from one wave to the next one should not be too different. Additional quality metrics are generally computed for each wave such as percentage of speeders, straight-liners, inconsistency in the same questionnaire, and failure to follow an instruction.

All these designs can be further compared to benchmark estimates. Benchmarks are typically demographic and behavioral measures (such as health status, race, or number of rooms in the household), and usually come from official government statistics such as the American Community Survey. Attitudinal benchmarks come from high-quality surveys with probability samples such as the National Election Studies, or the General Social Survey. Until now, benchmarks have generally been collected by an interviewer in surveys that achieve extremely high response rates.

If benchmarks are available and usable for some or all questions, then each panel can be compared against the benchmark, and a measure of error can be computed from that comparison. However, in order to compare the results from surveys to benchmark estimates, two requirements should ideally be met:

1. *Question wording should be identical across the compared surveys*. Question wording is something to keep in mind when comparing studies, regardless of design. Small wording changes have shown to sometimes produce large effects on measurement (e.g., Smith, 1995), therefore to avoid confounding effects, the exact same question wording should be used in all surveys. At the same time, this can be difficult to achieve when mode differs across the surveys being compared and question adaptation becomes necessary. Specifically, benchmarks and other probability-based studies are often collected in interviewer-administered formats where questions are delivered orally, therefore questions selected from these surveys to include in the online panels for later comparison will need to be adapted to the self-administered, visual delivery mode.

2. *The populations represented by each survey need to be comparable*. If the benchmark survey includes population members without Internet access, these will have to be excluded from the estimation if the online panel includes only respondents with Internet access, as is usually the case. Problems may emerge if the definition of the Internet population used by the agency providing the benchmarks does not match the population from which the study respondents were recruited. This is further complicated when no question is asked on the benchmark study that identifies Internet users.

In Section 2.3 we provide a review of accuracy metrics that have been used to evaluate the differences in data quality between online panels and other surveys.

2.3 Accuracy metrics

When comparing results from online panels to benchmarks, different accuracy metrics are used in the literature:

1. *Direct comparisons* (panel by panel) to benchmarks of response distributions are the most commonly reported metric (e.g.,Vonk, van Ossenbruggen, & Willems, 2006; Walker, Pettit, & Rubinson, 2009) and look at the variability of estimates from different

sources. Panel names are usually not disclosed, with the exception of a few studies with a smaller number of panels (e.g., Duffy & Smith, 2005; Malhotra & Krosnick, 2007).

2. The *lowest and highest values* provide the reader with a range of possible estimates computed from data from the surveys used in the study (van Ossenbruggen, Vonk, & Willems, 2006).

3. The *average estimates across panels* are compared to a benchmark in the NOPVO (van Ossenbruggen et al., 2006) and the ARF study (Walker et al., 2009). This metric focuses on one estimate at a time and has the disadvantage of masking differences across panels; even if the overall average of an estimate across panels is equal to the benchmark, individual panels might grossly underestimate or overestimate the phenomenon, which would mean that using a single panel to address a research question would most likely result in biased estimates.

4. To solve the previous measurement issue, Yeager, Krosnick, et al. (2011) propose the *average absolute error* as a metric. The average absolute error is the average of the absolute difference between the modal category of the benchmark and the survey estimate for that category. It has the advantage of avoiding differences to cancel out.

5. The *largest absolute error* is used to summarize more than one estimate and it is measured as the error of the variable estimate in which the survey was least accurate (Yeager, Krosnick, et al., 2011).

6. The *number of significant differences from the benchmark* is the percentage of variables considered in the study that are statistically significantly different from the benchmark. It can be reported panel by panel or as the average percentage across panels (Yeager, Krosnick, et al., 2011).

All the above metrics can be reported either weighted or unweighted and, of course, more than one metric can be reported and compared to each other. We treat the issue of weighting later in the chapter.

2.4 Large-scale experiments on point estimates

Among the numerous studies that compare accuracy of estimates from online panels, many focus on comparing one panel to another survey, and a smaller number compare accuracy of several online panels. For space reasons, we focus on the largest comparisons experiments on point estimates that have been conducted since 2006, starting with the pioneering NOPVO project conducted in the Netherlands.

2.4.1 The NOPVO project

The first published large-scale experiment was initially presented at the 2006 ESOMAR panel research conference. Vonk, van Ossenbruggen, and Willems (2006) illustrated the Dutch online panel comparison (NOPVO) project (http://www.nopvo.nl/english/english.htm). The study compared the results of fielding the same survey on samples of approximately 1000 panel members from 19 different online panels of nonprobability samples in the Netherlands,

which captured 90% of all Dutch online panel respondents at the time (Van Ossenbruggen et al., 2006). An omnibus questionnaire was administered in each panel during the same week of 2006, and was in field during seven days after the initial invitation. No quota sampling was used in selecting each sample from each panel. In total, 18999 panel members were invited to participate and 9514 completed the survey for a completion rate (Callegaro & DiSogra, 2008) of 50.04%.

To investigate data quality, the data were compared, when possible, to known benchmarks from Statistics Netherlands (CBS). Together with the omnibus questionnaire, panel member historical data were attached to the records and used in the analysis. When compared to known benchmarks, respondents across all 19 panels were more likely to be heavy Internet users (81% reported going online daily compared to the CBS benchmark of 68%), less likely to belong to a minority group and more likely to live in big cities. The average estimate of voter turnout, for example, was 90%, but the actual turnout was 79%. Voters for the Christian Democrats were on average underrepresented in the panels (16% vs. 29%) whereas voters of the Socialist Party were overestimated (14% vs. 6%). Some 23% of online panel members claimed to belong to a religious community as compared to a benchmark of 36%. The percentage of respondents who reported doing paid work for more than 15 hours a week varied across all panels from 53% to 82% (28 percentage point difference), whereas the percentage of respondents surfing the web for more than 10 hours had a range of variation of 29 percentage points across the lowest to the highest panel estimate. Although in the original NOPVO study no data were collected online from probability-based samples, a recent study (Scherpenzeel & Bethlehem, 2011) conducted using the Dutch probability-based online panel Longitudinal Internet Studies for the Social Sciences (LISS) compares the same statistics (collected on the LISS panel in 2006) to the benchmark data used by the NOPVO experiment. The bias from the LISS panel, measured as the difference from the benchmark, was smaller than that of the average NOPVO bias in five of the six benchmarks.

2.4.2 The ARF study

Based on concerns raised by early research on panel data quality, the Advertising Research Foundation (ARF) set up the Online Research Quality Council (ORQC) in August 2007 (Walker et al., 2009). One of the council's plans was to arrange a comparison study (NOPVO style) among 17 US online panel providers (all using nonprobability samples) a telephone sample panel, and a mail sample panel. A two-wave study was conducted in October and November 2008. One version of the questionnaire was fielded at a local market level (selected local markets). The online questionnaire was administered by a third independent party and contained: (1) factual and behavioral questions to be compared against known benchmarks; and (2) other common market research attitudinal questions such as intention to purchase items. Factual and behavioral questions were asked with the same question wording as the benchmarks they would be compared against. Of 1038616 invites, 76310 panel members completed the study for a completion rate of 7.34%. Various findings were obtained from this large study, whose estimated "book value" cost exceeded $1 million. When compared to known benchmarks, the study showed a similar pattern to the NOPVO study, with wide variation across panels in the survey estimates of interest. For instance, most panels overestimated smoking behavior; the estimates ranged from 42% (matching the benchmark value

from NHIS) of members admitting having smoked at least 100 cigarettes in their entire life, to 58%, depending on the panel. Cell phone ownership was also overestimated across panels ranging from 85–93%, all above the benchmark value of 79%. Where panels showed the highest variance was in purchase intent and likelihood to recommend questions, typical market research questions. Two products were tested: the intention to purchase a new soup and a new paint. The percentage of panel members who chose the two response options indicating highest likelihood of purchase for the new soup varied from 32%–53% across panels. The authors also found that sample tenure (how long the respondent had belonged to the panel) was negatively related to the intention of purchase. Panel members with self-reported three or more years of membership were less willing (37%) to recommend the new kind of soup than panel members with three months or less of panel tenure (50%). A similar picture emerged for intent to recommend a new type of paint, 48% versus 62%.

The ARF redid the above study in 2012 with a similar design under the umbrella of the Foundation of Quality 2 (FOQ2) taskforce.[4] At the time of writing, there are no publicly available results to report.

2.4.3 The Burke study

The research firm Burke commissioned a study across 20 online panels with nonprobability samples and one online panel with a probability sample (Miller, 2007, 2008). The main purpose of the study was to investigate fraudulent respondents and satisficers. The same questionnaire, which included qualifying (screening) questions, "trap questions," and other standard market research questions was commissioned to the 21 online panels. No quota control in the within-panel sample design was set and the survey length was of about 25 minutes. Completion rates had an extremely large variability, similar to the NOPVO study, going from 3%–91% with an average of 18%. Few of the estimates had the potential to be benchmarked.[5] One of the benchmarked items asked in 11 of the panels was a question about whether the respondent was left-handed or ambidextrous. The absolute average error was of 1.7 percentage points for the proportion of left-handed respondents (ranging from a difference from the benchmark of −2 percentage points to +3 percentage points) and of 4.5 for the proportion of ambidextrous respondents (ranging from a +2 percentage-point to a +6 percentage-point difference from the benchmark). When comparing estimates of usage of blood glucose monitors, the range varies from a minimum of 10% to a maximum of 17% and the incidence of respondents claiming to have pet health insurance from a minimum of 4% to a maximum of 22%.

2.4.4 The MRIA study

A study similar to the ARF study was conducted in 2009 for the Marketing Research and Intelligence Association (MRIA) among 14 Canadian panels, one of which was Probit, an online panel with a probability sample (Chan & Ambrose, 2011). In this study, quotas for age, gender, and income were used to draw the sample. In terms of coverage of the target population, the authors reported that some panels could not deliver enough respondents for Quebec whereas others vastly under represented the French-speaking population. When looking at differences across panels for newspaper, magazine and radio consumption, the variation was small across panels. Further research steps were announced in the article but (to our

[4] http://thearf.org/foq2.php.

[5] No details are provided in the article about the source used for the benchmark estimates.

knowledge) no publication was available at the time of printing. Despite the fact that each panel was anonymized in the article, there was only one panel with a probability sample (Probit), which was therefore self-identified. At the same annual conference in 2010, Probit (EKOS, 2010) reanalyzed the MRIA study using the average of the panels with nonprobability samples and compared it against the Probit estimates collected in the same experiment. Official benchmarks were also added to the study. The authors found that Probit panel members were less likely to be heavy Internet users, to use coupons when shopping, and to have joined the panel for the money or incentives than members of the online panels of nonprobability samples. When compared to the distribution of income for the Internet population according to official benchmarks, online panels of nonprobability samples recruited more members with lower income than the Probit panel, which yielded estimates of income that were however closer to the benchmark.

2.4.5 The Stanford studies

Finally, Yeager, Kosnick, et al. (2011) compared estimates from an RDD telephone survey to estimates from six online panels of nonprobability samples, one online panel with a probability sample, and one cross-sectional sample recruited via river sampling. The exact same online questionnaire was used in all surveys. Data were collected in the fall of 2004 and early 2005 for a total sample size of 1000 respondents per company (study 1). A second round of data collection was done in 2009 with the same probability sample of 2004 and two nonprobability panels of the previous study (study 2). The questionnaire contained items on basic and secondary demographics such as marital status, people living in the households, and home ownership. Other questions asked were frequency of smoking, passport ownership and health status. The uniqueness of the Stanford study is that *every* question was selected so that known gold standards collected by US federal agencies were available for comparison. The authors were then able to compute and compare the absolute average error of each sample source.

Results indicated that the RDD and the probability-based online panel data were on average closer to the benchmarks than any of the online panels with nonprobability samples. The same findings were found for the more recent data collection of 2009: the average absolute error among the same panel providers was close to that in the 2004/2005 study. The probability sample was also more accurate than the two nonprobability samples.

2.4.6 Summary of the largest-scale experiments

To better summarize the findings from these large-scale experiments we have compiled two tables where data from the above studies are compared with known benchmarks coming from official, high-quality surveys with probability samples. In Table 2.2 we have compiled the comparison with smoking benchmarks across different studies. In order to standardize the comparison across studies the average absolute difference metric described above has been used. We could not use other metrics, such as the largest absolute error and the number of significant differences from the benchmark, because detailed panel-by-panel original estimates are not available for the studies considered, with the exception of the Stanford study.

To shed more light on the variability of smoking estimates across panels, in Table 2.3 we reproduce Figure 1 of Walker et al. (2009, p. 474).

Probability sample panels were always closer to the smoking benchmarks than nonprobability sample panels (see Table 2.3). This is true for studies conducted in different years and countries. Online panels of nonprobability samples in the United States and in Canada tend

Table 2.2 Average absolute error of smoking estimates across different studies.

Study	Variable	Benchmark compared to	Average absolute error	Range min–max
Stanford study 1	Non-smoker	1 RDD sample	2.6	–
Stanford study 1	Non-smoker	1 Probability sample panel	4.2	–
Stanford study 1	Non-smoker	Average of 6 nonprobability sample panels	9.6	5.8–17.8
ARF	Ever smoked	Average of 17 nonprobability sample panels	10.0	–
ARF	Currently smoke	Average of 17 nonprobability sample panels	5.6	0–12
MRIA	Currently smoke	Average of 13 nonprobability sample panels + 1 probability sample panel	10.5	–
MRIA	Currently smoke	1 Probability sample panel	2.1	–

Table 2.3 Comparison of weighted percentages regarding smoking behaviors across the 17 nonprobability sample panels in the ARF study.

Source	Currently smoke	Smoked at least 100 cigarettes in your entire life
NHIS/CDC benchmark	*18*	*42*
Panel A	19	42
Panel B	20	47
Panel C	20	47
Panel D	21	48
Panel E	23	49
Panel F	24	50
Panel G	26	50
Panel H	26	50
Panel I	27	50
Panel L	27	51
Panel M	28	51
Panel N	28	51
Panel O	30	52
Panel P	30	55
Panel Q	31	57
Panel R	32	57
Panel S	33	58

Notes: The data come from two different panels which are organized in order of magnitude so the readers should not assume that the results from the same row come from the same panels. Data shown in order of magnitude.

Table 2.4 Average absolute error of average estimates of different variables across different studies.

Study	Variables	Benchmark compared to	Average absolute error	Range min–max
NOPVO	6 variables	Average of 19 nonprobability sample panels	8.5	Cannot be computed[1]
NOVPO	6 variables	1 probability sample panel	4.0	–
Stanford study 1	13 variables	1 RDD sample	2.9	–
Stanford study 1	13 variables	1 probability sample panel	3.4	–
Stanford study 1	13 variables	Average of 6 nonprobability sample panels	5.2	4.5–6.6
ARF	6 variables	Average of 17 nonprobability sample panels	5.2	0–10
Stanford study 2	13 variables	1 RDD sample	3.8	–
Stanford study 2	13 variables	1 nonprobability sample panel	4.7	–
Stanford study 2	13 variables	1 probability sample panel	2.8	–

Note: [1]Data for each single panel included in the NOVPO experiment are not available so we cannot report the minimum and maximum value.

to estimate a higher proportion of smokers than the proportion of smokers in the population according to the benchmark, even after weighting.

The same finding is replicated using other variables (see Table 2.4). Most of the variables analyzed in this study are behavioral or factual in nature such as work status, number of bedrooms in the house, number of vehicles owned, having a passport, drinking and quality of health, having a landline or cell phone, and party voted for in the last election. Here again, probability sample panels and RDD telephone surveys are closer to the benchmarks than online panels based on nonprobability samples.

Sometimes benchmarks are not available, either because more accurate population estimates are impossible to collect for a given variable or because they are not readily available when analyses are conducted. In these cases it is not possible to use an accuracy metric but it is still possible to study the variability of estimates across panels. This kind of data is still relevant and informative for survey commissioners to appreciate how reliable data from online panels might be.

The NOPVO study addressed this question by studying familiarity with brands ("Have you seen a commercial of the following [brand]?"). The original values were not reported in the study; instead a mean value was computed across panels together with the top three estimates plus the bottom three estimates, providing an indication of variability across estimates from different panels. In comparison to the average brand awareness across panels, estimates varied from −5 to +7 percentage points for Citroën, from −9 to +9 for Mazda, from −6 to +6 for T-mobile and from −11 to +5 for Volkswagen (see Table 2.4).

In the ARF estimates about willingness to buy the new soup and paint,[6] the percentage of respondents who selected the top two answers (definitely and probably would buy) varied from a low range of 34% to a high range of 51% for the soup and from 37% to 62% for the new paint (weighted results). In the same ARF study, the mail sample estimate for the intention to buy the new soup was 32%, and for the phone sample 36%.

2.4.7 The Canadian Newspaper Audience Databank (NADbank) experience

In 2006, the Newspaper Audience Databank (NADbank), the Canadian daily newspaper audience measurement agency, initiated a test to assess the feasibility of collecting newspaper readership data using an online panel rather than the until then traditional data collection protocol based on RDD telephone surveys (Crassweller, D. Williams, & Thompson, 2006). In the experiment, the results from their standard telephone data collection (spring and fall) were compared to results from 1000 respondents from an online panel with a nonprobability sample (same time periods) for the Toronto CMA.[7] The online sample estimates for average number of hours per week of TV and Internet usage, as well as for average number of newspapers read per week, were higher than the estimates from the telephone sample (Crassweller et al., 2006). Most importantly, the key readership metrics by newspaper differed with the different sampling approaches and there was no consistent pattern or relationship in the differences.

Based on these initial results NADbank decided to broaden the scope of the test and include more online panels (Crassweller, Rogers, & Williams, 2008). In 2007, another experiment was conducted in the cities of Toronto, Quebec City, and Halifax. Again, the four nonprobability sample panels received identical instructions for project design, implementation, weighting, and projection and were run in parallel with the telephone RDD sample in those markets. The results from the four panels varied substantially in terms of demographic composition (unweighted and after weighting to census data for age, gender, and household size) and in terms of media habits; panels did not mirror the benchmark in any of the metrics of interest.

Compared to the benchmark, all panel estimates of readership for both print (paper versions) and online newspapers were over estimated to varying degrees. This was true for all newspapers in all markets. No one panel performed better than another. The authors concluded that there was no obvious conversion factor to align panel estimates to RDD estimates and that the panel recruitment approach could not provide a sample that reflected the general population. Without such a sample it would be impossible to gain valid insights regarding the population's newspaper readership behavior. The outcome of the test resulted in NADbank maintaining their current RDD telephone sampling methodology. It was clear that at that time "a web-based panel does not provide a representative sample, and secondly that different panels produce different results" (Crassweller et al., 2008, p. 14).

Four years later, NADbank commissioned another study, this time comparing the results from their RDD sample to Probit, an online panel with a probability sample recruited using landline and cell-phone exchanges with an IVR recruitment protocol (Crassweller, J. Rogers, Graves, Gauthier, & Charlebois, 2011). The findings from the online panel were more accurate than the previous comparisons. In terms of unweighted demographics, Probit was better able to match census benchmarks for age and gender than previous panels. The probability-based panel recruitment approach resulted in closer estimates of print readership but over estimated

[6] Assuming this product was available at your local store and sold at an acceptable price, which of the following statements best describes how likely you would be to buy it?

[7] Statistics Canada Census Metropolitan Areas (CMA).

online readership. The authors concluded that this approach was an improvement on previous panel recruitment approaches but still reflected the limitations of the recruitment method (IVR) and the predisposition for mediaphiles to participate in online media surveys. The key strength of the Probit (IVR) approach is that it "has the potential to provide a one-stop shop for online and offline consumers" (p. 6). The authors warned that more work still needed to be done before quantitative research studies can be conducted using online panels of nonprobability samples but concluded that incorporating RDD sampling approaches with the use of online panel measurement provided alternatives for the near future.

2.4.8 Conclusions for the largest comparison studies on point estimates

The main conclusion from this set of studies is that different results will be obtained using different panels or, in other words, that online panels "are not interchangeable". In the NOPVO study Vonk, Ossenbruggen & Willems, (2006, p. 20) advise: "Refrain from changing panel when conducting continuous tracking research". Similar statements are made in the ARF study: "The findings suggest strongly that panels are not interchangeable" (Walker et al., 2009, p. 484), and in the comparison done in Canada by MRIA (Chan & Ambrose, 2011, p. 19) "Are Canadian panels interchangeable? Probably not for repetitive tracking". On a different note, the authors from the Stanford study conclude their paper by saying: "Probability samples, even ones without especially high response rates, yielded quite accurate results. In contrast, nonprobability samples were not as accurate and were sometimes strikingly inaccurate" (Yeager, Krosnick, et al., 2011, p. 737).

2.5 Weighting adjustments

Differences across panels' estimates could potentially disappear after each panel has been weighted. Unfortunately in the reviewed studies that was not the case. The ARF weighting on common demographics made almost no difference in reducing the discrepancy among panels and in comparison to the benchmarks. A second type of weighting was then attempted. In this approach, in addition to post-stratification, duplicates and respondents who belonged to multiple panel were removed. This second approach improved data quality to some extent, but significant differences from the benchmarks still remained (Walker et al., 2009). The ARF study stated: "Sample balancing (weighting) survey data to known census targets, … removed variance but did not completely eliminate it. Likewise, the test of a pseudodemographic weighting variable (panel tenure) did not eliminate variance" (Walker et al., 2009, p. 473).

In the NADbank report the authors conclude that: "There is no firm basis on which to develop a conversion factor or weight that could bridge telephone and online findings" (Crassweller et al., 2008, p. 14). Finally, in the Stanford study, the authors concluded: "Post-stratification of nonprobability samples did not consistently improve accuracy, whereas post-stratification did increase the accuracy of probability sample surveys" (Yeager, Krosnick, et al., 2011, p. 733).

Finally, Tourangeau, Conrad, and Couper (2013) presented a meta-analysis of the effect of weighting on eight online panels of nonprobability samples in order to reduce bias coming from coverage and selection effects. Among different findings, they concluded that the adjustment removed at most up to three-fifths of the bias, and that a large difference across variables still existed. In other words, after weighting, the bias was reduced for some variables but at the same time it was increased for other variables. The estimates of single variables after weighting would shift up to 20 percentage points in comparison to unweighted estimates.

A promising approach that has been developed during the year is the use of propensity score weighting, as discussed in Chapter 12.

2.6 Predictive relationship studies

Findings reported until now suggest that researchers interested in univariate statistics should avoid using panels from nonprobability samples to obtain these estimates. However, more often than not, researchers are interested in investigating relationships between variables, and some argue that multivariate analyses might not be biased when computed using panels of nonprobability samples.

This section summarizes findings from four studies that have compared estimates of association between variables in probability sample panels against nonprobability sample panels.

2.6.1 The Harris-Interactive, Knowledge Networks study

Parallel studies on global climate change and the Kyoto Protocol were administered to an RDD telephone sample, to two independent samples drawn five months apart on the nonprobability sample Harris Interactive panel (HI), and on the probability sample panel Knowledge Networks (KN) (Berrens, Bohara, Jenkins-Smith, Silva, & Wiemer, 2003). The authors compared the relationships between environmental views and ideology across the four samples. When combining the samples and looking at an ordered probit model predicting environmental threat (on an 11-point scale: 0 = No real threat; 10 = brink of collapse), the model showed that ideology was a strong predictor of perceived threat, where the more conservative respondents were, the least of a threat they saw in global warming. There were, however, large significant interactions of the Internet samples (taking the RDD sample as baseline) where the relationship between ideology and perceived threat was less strong in the two nonprobability samples. When controlling for demographics, the effect of the sample source disappeared. In a logistic regression analysis predicting if respondents would vote for or against (0–1) ratification of the Kyoto Protocol given an increased amount of taxes, the authors found that respondents to all the online panels were less supportive of the Kyoto Protocol that respondents to the telephone survey. However, in all samples "the analyst would make the same policy inference (…) – the probability of voting yes on the referendum is significantly and inversely related to the bid price (or cost) of the policy" (p. 20).

2.6.2 The BES study

Parallel to the British Election Studies (BES) of 2005 (a face-to-face survey where addresses were selected from a postal address file in the United Kingdom with a response rate of over 60%), an Internet sample was selected from the YouGov panel (based on a nonprobability sample) with the goal of comparing the accuracy of estimates from both designs (Sanders, Clarke, Stewart, & Whiteley, 2007). The authors found significant differences between the two samples with respect to point estimates of political choice, voter turnout, party identification, and other questions about political issues, where the probability sample was overall, but not always, more accurate than the nonprobability sample. Models predicting three different variables were tested in each sample.

1. The first model used 16 variables to predict voting turnout and found significant differences across samples in five of the 21 estimated parameters. For two of the parameters

(efficacy/collective benefits and education), the relationship was significantly stronger for the face-to-face probability sample. For two other parameters (personal benefits and Midlands Region), the coefficient was significant in one sample but not in the other. Finally, according to the face-to-face probability sample, females were less likely to have voted than males. The opposite was found in the Internet nonprobability sample.

2. The second model was a regression on party choice in the 2005 election, where significant differences were found in 5 of the 27 estimated parameters. Again, for two parameters (Blair effect and Kennedy effect) the coefficient was larger in the face-to-face probability sample than in the Internet nonprobability sample. Two other parameters (party-issue proximity and Southwest region) were significant in one sample and not in the other, and one parameter (age) was negative in the face-to-face sample (suggesting, as one would expect, that older respondents were less likely to vote for the Labour Party) and positive in the Internet nonprobability sample.

3. In the third set of models, rather than comparing coefficients, different competing models were compared to try to find the one that better explained the performance of the rival party. Both samples led to the same conclusions when inspecting the associated explained variance and other goodness-of-fit statistics.

2.6.3 The ANES study

Around the same time, Malhotra and Krosnick (2007) conducted a study comparing the 2000 and 2004 American National Election Study (ANES), traditionally recruited and interviewed face-to-face, to data collected from nonprobability Internet samples. Response rates in the ANES were above 60%; the 2000 ANES sample was compared to a sample obtained from the Harris Interactive panel survey, and the 2004 ANES sample was compared to a sample from the YouGov panel. The questions asked of each sample were not always identical, but only those questions with similar questions and equal number of response options were used to compare the face-to-face probability samples to their Internet nonprobability counterparts. In contrast to the multivariate regression approach followed by Sanders et al., Malhotra and Krosnick analyzed bivariate logistic regressions that predicted "predicted" vote choice, actual vote choice, and actual turnout.

Results showed that the design of the surveys (which used a different mode and sampling strategy) had an impact on survey estimates of voting intention and behavior as well as on estimates of bivariate relationships. For example, in the 2004 study, 10 out of 16 parameters predicting "predicted" vote choice were significantly different in the two sources of data; in the 2000 study, 19 out of 26 parameters were significantly different. When predicting actual vote choice using data from 2000, 12 out of the 26 parameters were significantly different across samples. Weighting the data did not reduce these differences, and they were not entirely explained by different levels of interest in politics of respondents in both types of sample. As in the BES study, even though the true values of the regression parameters are unknown, we do know that point estimates about vote choice and turnout were more accurate in the face-to-face sample than in the nonprobability Internet sample.

2.6.4 The US Census study

The third study investigating differences in relationships between variables compared a series of RDD telephone surveys collecting data from about 200–250 respondents per day for almost

5 months to online surveys fielded on weekly nonprobability samples from the E-Rewards panel. This resulted in about 900 completes per week for approximately 4.5 months (Pasek & Krosnick, 2010). Using questions that were identical or virtually identical, they first compared the demographic composition of the two sets of data and found that the telephone samples were more representative than the Internet samples. When comparing response distributions for the substantive variables, there were also sizeable differences (often differing by 10 to 15 percentage points) between the two samples.

Pasek and Krosnick (2010) first compared bivariate and multivariate models predicting two different variables tested in each sample. When predicting intent to complete the Census Form, 9 of the 10 substantive variables had similar bivariate associations in the expected direction. For example, in both samples, respondents were more likely to report intent to complete the Census form if they thought the Census could help them, or if they agreed that it is important to count everyone. For the tenth variable the relationship was in the expected direction for the telephone sample, but panel respondents who did not think it was important to count everyone were *more* likely to intend to complete the census form. For eight of the substantive variables where the direction of the relationship was the same in both samples, however, the relationships were stronger for the panel sample than for the telephone sample for five variables and weaker for three variables. Demographic predictors were often significantly different in the two samples, supporting different conclusions. When predicting actual Census form completion, differences were less pronounced but still present, suggesting again that which sample is used to investigate the research questions can have an impact on the conclusions that are ultimately reached.

Pasek and Krosnick also compared all possible correlations among the variables measured in both surveys, finding that correlations were significantly stronger in the panel sample than in the telephone sample. It is worth noting that in both the BES and the US Census study the relationship between age and the predicted variable differed significantly between the nonprobability online panel sample and the alternative probability sample. The relationship was significant for both samples but had opposite signs in each. In the nonprobability online survey, the relationship was the opposite of what was expected from theory. In addition, both the ANES and the US Census studies bivariate relationships tended to be significantly stronger for predictors in the online nonprobability sample than in the alternative sample. This suggests that respondents in the former were systematically different from the alternative method respondents.

Although some authors conclude that researchers would make similar conclusions when using probability or nonprobability panels (Berrens et al., 2003; Sanders et al., 2007) when looking at the signs of the coefficients, they are not always in the same direction (Pasek & Krosnick, 2010) and the strength of relationships varies across samples (Malhotra & Krosnick, 2007; Sanders et al., 2007; Pasek & Krosnick, 2010). We hope more studies will follow up this topic.

2.7 Experiment replicability studies

An important question for market researchers and behavioral scientists involves replicability – in terms of both significance and effect sizes – of random-assignment experimental studies that use as participants respondents from online panels. Indeed, market researchers often seek to understand what influences consumers' behaviors and attitudes. Experiments

are an effective method to assess the impact of some change in message or marketing strategy on a person's preference for or likelihood of purchasing a given product. Likewise, public opinion researchers often seek to understand the impact of a candidate's policy on the public's vote. Experiments that present respondents with randomly assigned messages, can allow campaigns to estimate the proportion of the vote that might be won when taking one strategy or another. The estimates of this impact can then be used to calculate the expected gain, in terms of sales or votes that might be found when taking one strategy versus another. This allows for more efficient use of resources. Therefore, it is often of interest to know both *whether* a given change is likely to alter Americans' behaviors or preferences, and also *how much* this change would affect them. Put another way, researchers who conduct experiments using online panels are often interested in both the *significance* of an experimental comparison and the *effect size* of that comparison. What does the research say about replicating experimental results – in terms of both significance and effect sizes – in probability and nonprobability-based samples? The research literature on this topic is sparse. To date, there has been no published exten- sive empirical or theoretical analysis of this question. Much research has focused on whether probability sample panels provide more accurate point estimates of the prevalence of various behaviors or characteristics as just discussed in this chapter, while no published study has comprehensively investigated whether probability versus nonprobability sample panels yield similar conclusions about causal relationships as assessed through experiments. However, there are a number of studies that happened to have used both probability and nonprobability samples when testing causal relationships using experiments (e.g., Bryan, Walton, T. Rogers, & Dweck, 2011; Yeager & Krosnick, 2011, 2012; Yeager, Larson, Krosnick, & Tompson, 2011). Furthermore, disciplines such as social psychology have a long history of discussing the potential impact of sample bias on experimental results (Henrich, Heine, & Norenzayan, 2010; Jones, 1986; Sears, 1986). In this section, then, we review: (1) the key theoretical issues to consider regarding the results of experiments in online panels; (2) the emerging empirical evidence and what future research needs to be conducted in order to sufficiently address this question.

2.7.1 Theoretical issues in the replication of experiments across sample types

One important starting point for theory about the replicability of experiments comes from researchers in social, cognitive, and personality psychology. These researchers have a long history of using nonprobability samples to conduct experiments – specifically, samples of undergraduate students who are required to participate in psychology studies to complete course credit. This large body of work has contributed greatly to our understanding of patterns of thinking and social behavior. However, at various times in the field's history it has responded to criticisms of its database. For instance, Sears (1986) proposed that the narrow age range, high educational levels, and other unique characteristics of college students make them different from adults in ways that may limit the generalizability of findings (see also Henry, 2008). Likewise, Wells (1993), a prominent consumer behavior researcher, said that: "students are not typical consumers" because of their restricted age range and educational levels and that ignoring these uniquenesses "place[s] student-based conclusions at *substantial* risk" (pp. 491–492, emphasis added).

Psychologists have responded to these criticisms by arguing that the objective of much academic research is not to produce point estimates but rather to assess the causal relation

between two conceptual variables in any segment of the population. For instance, Petty and Cacioppo (1996) stated:

> If the purpose of most psychological or marketing laboratory research on college students were to assess the absolute level of some phenomenon in society (e.g., what percentage of people smoke or drink diet coke?) … then Wells's criticism would be cogent. However, this is not the case. [A laboratory study using college students] examines the viability of some more general hypothesis about the relationship between two (or more) variables and ascertains what might be responsible for this relationship. Once the relationship is validated in the laboratory, its applicability to various specific situations and populations can be ascertained.
>
> (pp. 3–4)

Similarly, Ned Jones (1986) has argued that:

> Experiments in social psychology are informative mainly to the extent that they clarify relationships between theoretically relevant concepts. Experiments are not normally helpful in specifying the frequency of particular behaviors in the population at large.
>
> (p. 234)

Indeed, as noted above, research to assess point estimates is distinct from research to understand relations between variables. However, marketing and political researchers are often not interested in whether a given relationship could exist in *any* segment of the population during any time period, but whether it exists *right now in a population they care about*, that is, consumers and voters. Further, as noted above, the size of an effect is often a substantive question. Understanding not only that something *might* matter under some specified set of conditions is sometimes less important when making decisions about how to invest resources than knowing *how much* something matters. And there is no strong statistical rationale for assuming that the size or significance of results from a small biased sample will be true in the population as a whole. To the contrary, statistical sampling theory suggests that any estimate of a parameter will be more accurate when that parameter is estimated using data from a random sample, compared to a biased (nonrandom) sample.

While there is no statistical basis for assuming homogeneity of effect sizes in a biased versus probability-based sample, the logic of random assignment assumes that whatever characteristic that might affect the outcome variable will be distributed equally across the two conditions (see Morgan & Winship, 2007). Given a large enough sample so that participant characteristics are truly randomly distributed across conditions, sample selection bias would only be expected to bias the size of the treatment effect in the event that the sample is biased in terms of some characteristic that is correlated with a person's responsiveness to the experimental manipulation.

For instance, imagine an experiment to test two framings of a campaign issue. If these two framings are judged as equally different by everyone regardless of their cognitive ability, then a nonprobability sample that underrepresents high-education respondents might not result in different treatment effects. However, if only people who think carefully about the issues will notice the difference between the issue framings – that is, if only highly-educated people were expected to show a treatment effect – then a nonprobability sample that includes too-few college educated respondents might show a smaller or even nonexistent treatment effect. Therefore, one theoretical issue that will likely determine the replicability of an experiment

in probability versus nonprobability samples is whether the treatment effect is likely to be different for people with different characteristics, and whether the sampling methods are likely to produce respondents that differ on those characteristics.

A related issue involves research hypotheses that are explicitly designed to test whether a given subgroup of people (for instance, low-education respondents) will show an experimental effect (for instance, whether they will distinguish between the persuasiveness of two advertising campaigns). One assumption might be that *any* sample that includes enough respondents in that sub-group to allow for a test with reasonable power will provide an accurate estimate of the treatment effect for that group. That is, all low-education respondents may be thought to respond identically to the experimental manipulation, whether they were recruited through probability or nonprobability methods. Indeed, this is the perspective of much of psychology, which treats any member of a group (such as "low cognitive ability" vs. "high cognitive ability," (e.g., West, Toplak, & Stanovich, 2008); or "westerners" or "easterners" (Markus & Kitayama, 1991)) as a valid representative of the psychological style of that group. By this logic, it is unimportant whether such a study includes proportions of members of a sub-group that match the population. Instead, the crucial feature is whether the sample has enough people in that group to adequately allow for the estimation of experimental effects.

However, another perspective is that members of subgroups may only be considered to be informative about the thinking styles or behaviors of that subgroup if they were randomly sampled from the population. That is, the "low-education" respondents in a given sample may not resemble low-education respondents more generally in terms of their receptivity to an experimental manipulation. If this is true, then experiments using nonprobability samples to test for effects within a given subgroup may lead researchers astray.

In summary, if researchers are looking for main effects of an experimental manipulation, and if people's responsiveness to that manipulation is uncorrelated with a person's characteristics, then a nonprobability sample would be expected to provide similar estimates of an effect size as a probability-based sample (all other methodological details being equal). However, if responsiveness to the manipulation depends on some characteristic that is over- or under-represented in a nonprobability sample, then experimental effects might vary between that sample and a probability-based sample. Further, if researchers are hoping to assess experimental effects within some subgroup (e.g., low-income respondents, women, Latinos, etc.) and if respondents are not a random sample of people from that subgroup, then it is possible that the subgroup analysis will yield a different result in probability-based and nonprobability-based samples. With these issues in mind, we turn to the limited evidence available, in addition to future studies that are needed to further understand these issues.

2.7.2 Evidence and future research needed on the replication of experiments in probability and nonprobability samples

A large number of studies in psychology and behavioral economics have assessed the different results obtained in experiments with nonprobability samples of college students and nonprobability samples of nonstudent adults. Peterson (2001) meta-analyzed 30 meta-analyses that tested for moderation by sample type and found a great deal of variance in college student versus noncollege student samples. In many cases, findings that were significant and in one direction in one sample were nonsignificant or significant in the opposite direction in the other sample. Similarly, Henrich, Heine, and Norenzayan (2010) compared results from experiments conducted with samples of college students in the United States to results from the same experiments conducted with nonprobability samples of adults in other countries in

Africa or Asia. These authors found many cases of nonreplication or of studies that produced effects in the opposite direction. An obvious limitation in these studies, however, is that both of the samples were recruited using nonprobability methods. It is thus unclear which sample was biased in its estimate of the effect size.

A small number of studies have begun to test for experimental effects using a college-student sample and then have replicated the study using a probability-based sample. One prominent example is a series of experiments conducted by Bryan, Walton, Rogers, and Dweck (2011). These researchers assessed the impact of a brief framing manipulation the day before an election (referring to voting as "being a voter in tomorrow's election" vs. "voting in tomorrow's election") on registered potential voters' actual voting behavior (as assessed by looking for research participants in the validated voter file). In one study conducted with Stanford students, Bryan et al. (2011) found that the framing manipulation increased actual voter turnout by roughly ten percentage points. In a second study conducted with a probability-based sample of voters – members of the GFK Knowledge Panel – the authors replicated the significance of the effect, and the size of the effect was nearly identical. Thus, in at least one case, both significance and effect size were replicated in a probability-based sample.

Two other investigations have conducted randomized experiments to assess the impact of a small change in question wording on the validity of respondents' answers (Yeager & Krosnick, 2011, 2012; Yeager, Larson, et al., 2011). Yeager and Krosnick (2012) examined whether questions types that employ a stem that first tells respondents what "some people" and "other people" think before asking for the respondent's own opinions yields more or less valid data relative to more direct questions. They tested this in both nationwide probability-based samples (the General Social Survey, the FFRISP, and the Knowledge Panel) and in nonprobability-based Internet samples (from Lightspeed Research and Luth Research). These authors found that "some/other" questions yielded less validity, and this was true to an equal extent in both probability and nonprobability-based cases. Furthermore, they reached identical conclusions when they tested the "some/other" format in convenience samples of adolescents (Yeager & Krosnick, 2011). Replicating these overall findings, Yeager, Larson et al. (2011) found that the significance and size of the impact of changes in the "most important problem" question[8] were no different in an RDD telephone survey or in a nonprobability sample of Internet volunteers. Thus, the limited evidence so far does not suggest that there are substantial differences in either replication or size of effects across probability and nonprobability-based samples.

The evidence is not adequate, however, to assess the more general question of whether the two types of samples are always likely to replicate experimental effects. The studies noted above do not have likely *a priori* moderators that could have existed in substantially different proportions across the types of samples. Therefore, it will be important in future research to continue to examine effects that are likely to be different for different people. Furthermore, the studies above were not interested in sub-group analyses. It is an open question whether, for instance, studies assessing the impact of a manipulation for women versus men, or rich versus poor, would yield different conclusions in probability or nonprobability-based samples.

2.8 The special case of pre-election polls

The advantage of pre-election polls is that the main statistics of interest (voter turnout and final election outcome) can be evaluated against a benchmark for all panels. Baker et al.

[8] Respondents were asked: "What do you think is the most important problem facing the country today?"

(2010, p. 743) and Yeager, Larson et al. (2011, p. 734) provide a list of studies showing that nonprobability online panels can provide as good and sometimes better accuracy than probability sample panels. In the United States, for example, this goes as far back as the 2000 election (Taylor, Bremer, Overmeyer, Siegel, & Terhanian, 2001) and in the United Kingdom, this goes back to 2001 (YouGov, 2011). In the 2012 US election, nonprobability panels performed as well and sometimes better than traditional probability polling (Silver, 2012).

At the same time pre-election studies differ in several ways from other survey research. Pre-election polls are focused mainly on estimating one variable (the election outcome), which is most of the time (depending on the country) a binary variable. In addition, pre-election studies are often conducted in an environment where during weeks before the election many other studies, generally pre-election telephone polls, are publicly available. In fact, unlike the majority of surveys, in pre-election polls there are continuous sources of information that help guide additional data collection and refine predictive models based on identification of likely voters, question wording, handling of undecided and nonrespondents, and weighting mechanisms. Thus, differences in accuracy do not just reflect differences in accuracy of nonprobability samples panels but also differences on how all these variables are handled. As the recent AAPOR report from the nonprobability samples states: "Although nonprobability samples often have performed well in electoral polling, the evidence of their accuracy is less clear in other domains and in more complex surveys that measure many different phenomena" (Baker et al., 2013, p. 108).

As Humphrey Taylor recognized early on (Taylor, 2007), the secret to generating accurate estimates in online panels is to recognize their biases and properly correct them. In the specific case of election polls, some companies are better than others in doing so. The case of pre-election polls is encouraging and we hope that many more studies are published trying to extend successful bias correction methodologies to other survey topics.

2.9 Completion rates and accuracy

In online panels of nonprobability samples, response rates cannot be really computed because the number of total people invited to sign up (the "initial base") is unknown. Completion rates can still be computed by dividing the number of unique complete survey responses by the number of email invitations sent for a particular survey (Callegaro & DiSogra, 2008).

In the NOPVO study (Vonk et al., 2006), completion rates ranged from 18%–77%. The authors explained the differences as a function of panel management: some companies "clean up" their database from less active members more than others and they found that fresh respondents were more responsive than members who had been panelists a year or longer. Yeager, Krosnick et al. (2011) studied the effect of completion rates on accuracy of the responses finding that in the nonprobability samples, higher completion rates were strongly associated with higher absolute error ($r = .61$). A similar but slightly weaker relationship was found for the response rates of the seven RDD studies ($r = .47$) and for the response rates of the seven samples drawn from the probability-based Knowledge Networks online panel, ($r = .47$). These results add to an increasing body of literature suggesting that efforts to increase response rates do not result in improvements in accuracy, as previously expected.

2.10 Multiple panel membership

Multiple panel membership is an issue that has attracted the attention of the research community since the beginning of online panels. Also called *panel duplication* (Walker et al., 2009),

Table 2.5 Average number of membership per panel member, and percentage of members belong to five or more panels.

Studies	Year	X panel member	% belonging to 5+	Country
Multiple panels studies				
Chan & Ambrose	2011		45	CA
Walker et al.	2009	3.7	45	US
Gittleman & Trimarchi	2010	4.4	45	US
Gittleman & Trimarchi	2010		25	FR
Gittleman & Trimarchi	2010		19	ES
Gittleman & Trimarchi	2010		23	IT
Gittleman & Trimarchi	2010		28	DE
Gittleman & Trimarchi	2010		37	UK
Gittleman & Trimarchi	2010		38	AU
Gittleman & Trimarchi	2010		39	JP
Vonk et al.	2006	2.7	23	NL
Fulgoni	2005	8.0		US
Single panel studies				
Casdas et al.	2006		11	AU[1]
De Wulf & Bertellot	2006		29[2]	BE
Comley	2005		31	UK[3]

[1]Measured in one panel only, AMR interactive.
[2]Measured in one panel only, XL Online. Some 29% of members declared they belonged to more than one panel.
[3]Measured in one panel only, UK Opinion Advisors.

or *panel overlap* (Vonk et al., 2006), this is a phenomenon found in as many countries as we could find a study for. In Table 2.5 we list the average number of memberships per panel member and the percentage of members belonging to more than five panels, according to different studies. All these studies were undertaken by comparing online panels of nonprobability samples. At the current stage we could not locate studies of probability-based panels or of panels where membership is restricted by invitation only as described in Chapter 1.

It is not uncommon that members belong to multiple panels with as high as 45% of panel members belonging to five or more panels in the most recent estimates in the United States and Canada. The issue of multiple panel membership is important from two points of view: diversity of panel members, and data quality. The first aspect resonates with the concern that Fulgoni (2005) voiced that a minority of respondents might be responsible for the majority of surveys collected.

In the pioneering NOPVO study (Vonk et al., 2006), the number of multiple panel membership varied by recruitment methods: panels who bought addresses or recruited via link or banners had a higher amount of overlap (average of 4.3 and 3.7 panels per member respectively) than panels who recruited by phone or snowballing (2.0 and 2.3 respectively). Panel offering self-registration had an average overlap of 3.3, while panels recruiting via traditional research had an overlap of 2.4. Interestingly but not surprisingly, respondents with high Internet activity had an average multiple panel membership of 3.5 in comparison to low Internet users: 1.8 (i.e., respondents who checked their email once or twice a week). We will return to the issue of frequency of Internet usage with more up-to-date data later on in this

chapter. Casdas, Fine, and Menictas (2006) compared multiple panel member demographics with Australian census data, finding that they were more likely to be younger, less educated, female, working part-time and renting their living space. In the ARF study (Walker et al., 2009) multiple panel membership was again related to the recruitment method: higher multi-panel memberships occurred with unsolicited registrations, affiliate networks and email invitations. Multiple panel membership was also three times higher for African Americans and Latinos.

2.10.1 Effects of multiple panel membership on survey estimates and data quality

Most studies examining the effects of multiple panel membership on data quality have been conducted in the area of traditional market research questions such as shopping attitudes, brand awareness, and intention to purchase. In one of the first multiple panel membership studies, Casdas and colleagues (2006) noted that members belonging to more than two panels were more likely to be more price-driven than brand-driven in comparison to members belonging to one panel only and to a CATI parallel interview. The comparison was done with a multivariate model controlling for demographics characteristics. In terms of brand awareness Vonk et al. (2006) compared multiple panel members' answers to the average awareness results from all the 19 panels in their study. Multiple panel members had above average brand awareness but below average advertisement familiarity. Lastly, in the ARF study (Walker et al., 2009), members belonging to four or five and more panels were more likely to say that they would buy a new soup, or paint (intention to purchase concept test) than panel members belonging to less panels. For example, the percentage of respondents saying that they will definitely buy a new soup was of 12% for members belonging to one panel, 15% for two panels, 16% for three panels, 22% for four panels and 21% for five or more panels.

Vonk, van Ossenbruggen, and Willems (2006) noted a strong correlation ($r = .76$) between being a professional respondent (defined as number of multiple panel memberships + number of surveys completed in the last 12 months) and inattentive respondents (defined as completing the survey in a shorter amount of time and providing shorter answers to open-ended questions).

2.10.2 Effects of number of surveys completed on survey estimates and survey quality

Loyal respondents are desirable from the panel management point of view because they constantly provide answers to the surveys they are invited to. In a context of declining response rates, this can be seen as encouraging. At the same time, we need to explore the possibility that frequent survey-takers provide different answers than less frequent takers and what effect this might have in nonresponse and measurement error. In a Survey Spot panel study, Coen, Lorch, and Piekarski (2005) noted that experienced respondents (who had responded to 4–19 surveys) and very experienced responders (who had responded to 20+ surveys) gave much lower scores than inexperienced respondents (who had completed 1–3 surveys) on questions such as intention to buy, brand awareness, liking, and future purchase frequency. These results were true even after weighting the three groups to make sure they all represented 33% of responses and also after weighting by demographics.

The US bank Washington Mutual (WaMu) switched their market research data collection from telephone surveys to fully nonprobability online panels (the company used more than one). During the gradual switch, researchers at the company noted substantial variations

between online panels, across themselves, and in comparison to RDD telephone studies. The bank then started a program of research, pooling together 29 studies across different online panels for a total of 40000 respondents (Gailey, Teal, & Haechrel, 2008). One of the main findings was that respondents who took more surveys in the past three months (11 or more) gave lower demand ratings for products and services than respondents who took fewer surveys (10 or fewer surveys). When controlling for age, the same patterns held true. The second finding was that not only was the number of surveys a predictor of lower demand (for product and service) but also panel tenure. This prompted the bank to ask every online sample vendor to append survey experience auxiliary variables for their project.

In a very recent study, Cavallaro (2013) compared the responses of tenured Survey Spot members with new members on a variety of questions such as concept testing, propensity to buy, early adoption, and newspaper readership. In the study design, the same questions were asked twice to the same respondents a year apart. The data showed that tenured respondents were less enthusiastic about concepts (e.g., a new cereal brand) and more likely to be "early technology adopters" than new panelists. Differences over time for tenured respondents were small, suggesting that the observed differences between tenured respondents and new respondents are not due to changes in answers but rather to changes in panel composition due to attrition.

From the above studies it seems that the respondents who stay longer in a panel have different psychographic attitudes (at least in the topics discussed above) than new panel members. In this context, it is definitely worth mentioning the pioneering work done on the probability-based CentERdata panel in the Netherlands (Felix & Sikkel, 1991; Sikkel & Hoogendoorn, 2008) where panel members were profiled at the early stage with a set of 22 standardized psychological test on traits such as loneliness, social desirability, need for cognition and innovativeness. When looking at all respondents' scores on the 22 traits and correlating them with the length of stay in the panel, the authors barely found any statistically significant correlations. This study strengthens the Cavallaro (2013) hypothesis that the difference between new and tenured panel members is a matter of attrition, and not of panel conditioning at least on psychological traits. We look forward to new research in this area.

The issue of multiple panel membership is also debated in the context of *professional respondents*. We refer the reader to Chapter 10 of this volume for a thorough discussion on professional respondents and their impact on data quality.

2.11 Online panel studies when the offline population is less of a concern

By definition, the offline population is not part of online panels of nonprobability samples. In other words, individuals from the population of interest without Internet access cannot sign up for nonprobability-based online panels. Although it can be argued that weighting can compensate for the absence of the offline population from a survey error point of view, the percentage of people or households that are not online for a specific country contributes to potential noncoverage error. For this reason, probability-based panels so far have provided Internet access to the non-Internet population units or have surveyed them in a different mode such as mail or telephone (Callegaro & DiSogra, 2008).

In the commercial and marketing sector, the issue of representativeness and noncoverage of the offline population sometimes has a different impact than it has for surveys of the general

population. As discussed by Bourque and Lafrance (2007), for some topics, customers (that is, the target population) might be mostly online (e.g., wireless phone users) while for other topics (e.g., banking) the offline population is "largely irrelevant from a strategic decision-making standpoint" (p. 18).

However, comparison studies focused on the online population only provide increasing evidence that respondents joining online panels of nonprobability samples are different from the general online population in that they are heavier Internet users, and more interested in technology. For example, in a study comparing two online panels of nonprobability samples with a face-to-face survey of a probability sample, Baim and colleagues (2009) found large differences in Internet usage. According to the face-to-face survey, 37.7% of the adult population in the United States used the Internet five or more times a day, compared to a 55.8% in panel A and a 38.1% in panel B. For Internet usage of about 2–4 times a day, the face-to-face survey estimated that 24.8% of the population fell in this category whereas the nonprobability sample panel A estimated 31.9% and B 39.6%. A more recent study conducted in the United Kingdom compared government surveys to the TNS UK online panel (Williams, 2012). When looking at activities done during free time, the demographically calibrated online panel over estimated using the Internet by 29 percentage points and playing computer games by 14 percentage points. The author concludes: "the huge overestimate of Internet and computer games activity levels illustrates a general truth that access panels will not provide accurate prevalence about the use of technology" (p. 43).

Higher time spent online and heavier technology usage in comparison to benchmarks were also found in the Canadian comparison studies of Crassweller, Rogers, and Williams (2008) – higher number of time spent online "yesterday," and by Duffy and Smith (2005) – higher time spent online and higher usage of technology in the United Kingdom. Therefore, studies who are only interested in the online population might also be affected by differences related to Internet usage between those who belong to their target population and the subgroup that signs up for panels that recruit respondents online.

2.12 Life of an online panel member

As mentioned in Chapter 1 of this volume, online panels do not openly share details about their practices and strategies for fear of giving the competition an advantage. For this reason, it is not easy to know what is requested of online panel members. One way to obtain some information is to sign up in online panel portals that allow to do so, and monitor the activity as panel members. The company Grey Matter Research has used this approach. Staff and other volunteers signed up on different US online panels that allowed those who wanted to become members and monitored the traffic for three months (Grey Matter Research, 2012). At sign-up they did not lie on their demographics, nor did they try to qualify for studies they would not qualify for otherwise. In other words, the study was done with participants being on their "best behavior" – each member attempted to complete each survey to the best of their knowledge and in a reasonable time frame of three days maximum from the moment the email invitation was received. In Table 2.6, we report the results of the 2012 study. A similar study had also been conducted three years before (Grey Matter Research, 2009) with similar results.

Each volunteer monitored the number of invitations per panel. As we can see from Table 2.6, the range is quite wide, where the panel with highest invitation level sent on

Table 2.6 Life of an online panel member.

Panel	Average # of invitations in 30 days	% of surveys closed within 72 hours	Average questionnaire length in minutes
1	42.3	9.5	22.1
2	10.3	19.4	20.2
3	20.0	16.0	17.3
4	9.8	33.7	17.7
5	51.3	27.3	17.5
6	11.3	0.0	16.1
7	6.5	0.0	10.7
8	34.0	42.1	21.2
9	8.5	0.0	18.3
10	7.7	22.1	9.6
11	23.0	29.1	19.6
Average	**20.4**	**18.1**	**17.3**

average 51 invitations within 30 days and the panel with the lowest invitation level sent on average 6.5 invitations in 30 days.

A sizeable number of surveys were already closed when the participants attempted to complete them, with an average of 18.1% and a high of 42.1%. Surveys varied in length but they were on average 17.3 minutes long, with the panel with the shortest questionnaires lasting on average 9.6 minutes and the panel with the longest questionnaires having an average of 22.1 minutes. The above picture highlights likely levels of participation requests and burden on online panels. If we take the mean of the panels, for example, we can estimate that an "average" panel member would spend about seven hours filling out questionnaires in a month with high burden panels topping about 16 hours a month (e.g., panel 1) or low burden panels asking less than 2 hours of commitment a month (e.g., panel 7). These results are per single panel; if a respondent is a member of multiple panels, then the commitment quickly increases.

This rare study, which confirmed the results from the company's previous research conducted in 2009, sheds some light on the kind of data obtained by online panels. Active panel members are requested to participate in numerous surveys for a substantial amount of time each month. The importance of the studies lies in realizing the online panels are victim of their own success. It is hard for companies managing online panels to satisfy every client request. That translates into numerous survey requests per month. There is plenty of room for research to investigate the effects of heavy participation in surveys on their data quality.

2.13 Summary and conclusion

In this chapter we have systematized and brought together the disparate and sometimes hard to find literature on the quality of online survey panels, focusing on the critical review of the largest studies on data quality conducted thus far. This review should provide a starting point for additional studies as well as stimulate the publication of existing and new studies. It was apparent from our review that many of these studies appear on conference presentations, blogs, and few are published in peer-reviewed journals. This creates a problem of transparency

because for most studies some of the key survey information, such as the original questionnaire or descriptive statistics, was not available.

The chapter started with the proposal of a taxonomy of different comparison study techniques, together with a review of their strengths and weaknesses. The hope is that the taxonomy can be useful when researchers design future studies on online panel quality. In order to tackle the issue of quality of data obtained from online panels, we looked at three key quality criteria: accuracy of point estimates, relationships across variables, and reproducibility of results. Our recommendation is that researchers and data users/buyers analyzing data coming from online panels should use these criteria to assess the quality of the survey estimates.

The outcome of our review on point estimates, relationships across variables, and reproducibility of results points to quality issues in data obtained from online panels of nonprobability samples. Pre-election online polls are one exception to the general findings, where many web panels of nonprobability samples performed as good as and sometimes better than probability based pre-election polls. Weighting could have the potential to minimize the noncoverage and selection bias observed in online panels, but so far, again with the exception of pre-election polls, this strategy does not seem to be effective.

The final part of the chapter was devoted to common issues debated in the market and survey research arena, specifically the debate on the relationship between completion rates and accuracy, the issue of multiple panel memberships, and studies focusing only on the online population. In the first case, the agreement from the literature is that such relationship does not follow the expected direction. For probability-based panels (though we found only one study: Yeager, Krosnick, et al., 2011), higher completion rates lead to higher bias. For nonprobability panels, what makes a large difference in completion rates seems to be how the company manages the panel in terms of invitation and "panel member deletions" with different, mostly undocumented rules. Multiple panel membership was noticed early on, at the latest since the first study conducted by the NOPVO consortium in the Netherlands in 2006. Given the self-selected nature of panels of nonprobability samples, it is not uncommon for a panel member to sign up for multiple panels. Our review of the limited evidence highlights some issues of data quality for particular questions (e.g., purchase intent or product recommendation) and the fact that members who are more active (in terms of survey completion) than the average tend to have a different psychographic profile from members less active in the panel.

The reader might think that nonprobability sample panels are better suited to study the online population only. However, panel members tend to be heavy Internet users and heavy technology consumers, thus are less representative of the online population overall than sometimes is presumed.

We concluded the chapter by presenting an image of the life of an online panel member. The two studies we reviewed suggest that some panel members spend a high number of hours completing surveys each month. This burden is a new phenomenon, where the population has a higher chance than ever of being selected and receive survey requests, compared to the pre-Internet era when cross-sectional surveys where the norm and even in panel studies frequency of invitation tends to be considerably lower than that of online panels. Protecting the population from overload of survey requests might be important to maintain future cooperation from respondents. Hence, further research is needed investigating the optimal frequency of survey requests (and their length) is for online panel members.

We agree with Farrell and Petersen (2010) that Internet research should not be stigmatized. At the same time, it is worth noting that research conducted using online panels of

nonprobability samples has still numerous quality issues that have not been fully resolved. We hope this review can serve as a baseline for a more transparent research agenda on the quality of data obtained from online panels. However, we lament the fact that the existing commercial studies (NOPVO, ARF, and MRIA) produced insufficient documentation and did not share necessary methodological details such as the full questionnaire and descriptive statistics. We look forward to the findings from the new ARF study conducted by the FOQ2 which are expected to be made available to the entire research community (Therhanian, 2013).

Our taxonomy can help researchers to understand what conclusions can be drawn depending on the research design. The multiple focus on point estimates, relationships across variables, and replicability is the key to scientific advancement in this area. Together with weighting, data modeling, and learning from the successful case of pre-election polls, these aspects of the debate on online panels data quality should be on the agenda of research on online panel samples.

References

Anich, B. (2002). Trends in marketing research and their impact on survey research sampling. *Imprints*, May.

Baim, J., Galin, M., Frankel, M. R., Becker, R., & Agresti, J. (2009). Sample surveys based on Internet panels: 8 years of learning. *Worldwide Readership Symposium*. Valencia, Spain. Retrieved January 1, 2013, from: http://www.gfkmri.com/PDF/WWRS-MRI_SampleSurveysBasedOnInternetPanels.pdf.

Baker, R., Blumberg, S. J., Brick, J. M., Couper, M. P., Courtright, M., Dennis, J. M., Dillman, D. A., et al. (2010). Research synthesis. AAPOR report on online panels. *Public Opinion Quarterly*, *74*, 711–781.

Baker, R., Brick, M. J., Bates, N., Battaglia, M. P., Couper, M. P., Dever, J. A., Gile, K. J., et al. (2013, May). Report of the AAPOR task-force on non-probability sampling.

Berrens, R. P., Bohara, A. K., Jenkins-Smith, H., Silva, C., & Weimer, D. L. (2003). The advent of internet surveys for political research: A comparison of telephone and internet samples. *Political Analysis*, *11*, 1–22.

Bourque, C., & Lafrance, S. (2007). Web survey and representativeness: Close to three in ten Canadians do not have access to the Internet. *Should we care? Canadian Journal of Marketing Research*, *24*, 16–21.

Bradley, N. (1999). Sampling for Internet surveys. An examination of respondent selection for Internet research. *Journal of the Market Research Society*, *41*, 387–395.

Bryan, C. J., Walton, G. M., Rogers, T., & Dweck, C. S. (2011). Motivating voter turnout by invoking the self. *Proceedings of the National Academy of Sciences*, *108*, 12653–12656.

Callegaro, M., & DiSogra, C. (2008). Computing response metrics for online panels. *Public Opinion Quarterly*, *72*, 1008–1032.

Casdas, D., Fine, B., & Menictas, C. (2006). Attitudinal differences. Comparing people who belong to multiple versus single panels. *Panel research 2006*. Amsterdam: ESOMAR. Retrieved from: internal-pdf://Casdas-Fine-Menictas_2006_ESOMAR-0928660738/Casdas-Fine-Menictas_2006_ESOMAR.pdf.

Cavallaro, K. (2013). Theory of adaptation or survival of the fittest? *Quirk's Marketing Research Review*, *27*, 24–27.

Chan, P., & Ambrose, D. (2011). Canadian online panels: Similar or different? *Vue, (January/February)*, 16–20. Retrieved from: http://www.mktginc.com/pdf/VUE%20JanFeb%202011021.pdf.

Coen, T., Lorch, J., & Piekarski, L. (2005). The effects of survey frequency on panelist responses. *ESOMAR World Research Conference 2005*. Budapest: ESOMAR.

Comley, P. (2007). Online market research. In M. van Hamersveld, & C. de Bont (Eds.), *Market research handbook* (5th ed., pp. 401–419). Chichester: John Wiley & Sons, Ltd.

Crassweller, A., Rogers, J., Graves, F., Gauthier, E., & Charlebois, O. (2011). In search of a new approach to measure newspaper audiences in Canada: The journey continues. Paper presented at the Print and Digital Research Forum, San Francisco, CA. Retrieved from: http://nadbank.com/en/system/files/PDRFNADbankEKOS.pdf.

Crassweller, A., Rogers, J., & Williams, D. (2008). Between random samples and online panels: Where is the next .lily pad? Paper presented at the ESOMAR panel research 2008. Retrieved from: http://www.nadbank.com/en/system/files/Final%20paper.pdf.

Crassweller, A., Williams, D., & Thompson, I. (2006). Online data collection: Solution or band-aid? Paper presented at the Worldwide Readership Research Symposia, Vienna. Retrieved March 15, 2012, from: http://www.nadbank.com/en/system/files/Presentation%20WRRS%20Final.pdf.

Das, M. (2012). Innovation in online data collection for scientific research: The Dutch MESS project. *Methodological Innovations Online, 7*, 7–24.

Dillman, D. A., & Messer, B. L. (2010). Mixed-mode surveys. In P. V. Marsden, & J. D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 551–574). Howard House: Emerald Group.

Duffy, B., & Smith, K. (2005). Comparing data from online and face-to-face surveys. *International Journal of Market Research*, *47*, 615–639.

EKOS. (2010, January). Ekos' observation of MRIA study: Canadian online panels: similar or different? Retrieved August 28, 2011, from: http://www.ekos.com/admin/articles/MRIA-Comparison-Panel-Study-2010-01-27.pdf.

Farrell, D., & Petersen, J. C. (2010). The growth of Internet research methods and the reluctant sociologist. *Sociological Inquiry*, *80*, 114–125.

Felix, J., & Sikkel, D. (1991). Attrition bias in telepanel research. *Kwantitatiewe Methoden*, *61*.

Fulgoni, G. (2005). The professional respondent problem in online panel surveys today. Paper presented at the Market Research Association Annual Conference, Chicago, IL. Retrieved from: http://www.docstoc.com/docs/30817010/Partnership-Opportunity-Use-of-comScores-Survey-Panels.

Gailey, R., Teal, D., & Haechrel, E. (2008). Sample factors that influence data quality. Paper presented at the Advertising Research Foundation Online Research Quality Council conference (ORQC). Retrieved January 1, 2013, from: http://s3.amazonaws.com/thearf-org-aux-assets/downloads/cnc/orqc/2008-09-16_ARF_ORQC_WaMu.pdf.

Gittelman, S., & Trimarchi, E. (2010). Online research … and all that jazz!: The practical adaption of old tunes to make new music. *ESOMAR Online Research 2010*. Amsterdam: ESOMAR.

Grey Matter Research. (2009). Dirty little secrets of online panels: And how the one you select can make or break your study. Retrieved from: http://greymatterresearch.com/index_files/Online_Panels.htm.

Grey Matter Research. (2012). More dirty little secrets of online panel research. Retrieved from: http://www.greymatterresearch.com/index_files/Online_Panels_2012.htm.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33(2–3)*, 61–83.

Henry, P. J. (2008). College sophomores in the laboratory redux: Influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychological Inquiry*, *19*, 49–71.

Jones, E. E. (1986). Interpreting interpersonal behavior: The effects of expectancies. *Science, 234(4772)*, 41–46.

Malhotra, N., & 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*, 286–323.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224–253.

Martinsson, J., Dahlberg, S., & Lundmark, O. S. (2013). Is accuracy only for probability samples? Comparing probability and non-probability samples in a country with almost full internet coverage. Paper presented at the 68th Annual Conference of the American Association for Public Opinion Research, Boston, MA. Retrieved from: http://www.lore.gu.se/digitalAssets/1455/1455221_martinsson--dahlberg-and-lundmark--2013--aapor-is-accuracy-only-for-probability-samples.pdf.

Miller, J. (2007). Burke panel quality R&D summary. Retrieved from: http://s3.amazonaws.com/thearf-org-aux-assets/downloads/cnc/orqc/09-10-07_ORQC_Miller.pdf.

Miller, J. (2008). Burke panel quality R&D.

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and causal inference: Methods and principles for social research* (1st ed.). Cambridge: Cambridge University Press.

Pasek, J., & Krosnick, J. A. (2010, December 28). Measuring intent to participate and participation in the 2010 census and their correlates and trends: Comparisons of RDD telephone and non-probability sample internet survey data. U.S. Census Bureau. Retrieved from: http://www.census.gov/srd/papers/pdf/ssm2010-15.pdf.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461.

Petty, R. E., & Cacioppo, J. T. (1996). Addressing disturbing and disturbed consumer behavior: Is it necessary to change the way we conduct behavioral science? *Journal of Marketing Research*, *33*, 1–8.

Postoaca, A. (2006). *The anonymous elect. Market research through online access panels*. Berlin: Springer.

Sanders, D., Clarke, H. D., Stewart, M. C., & Whiteley, P. (2007). Does mode matter for modeling political choice? Evidence from the 2005 British Election Study. *Political Analysis*, *15*, 257–285.

Scherpenzeel, A., & Bethlehem, J. (2011). How representative are online-panels? Problems of coverage and selection and possible solutions. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet: Advances in applied methods and research strategies* (pp. 105–132). New York: Routledge.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515–530.

Sikkel, D., & Hoogendoorn, A. (2008). Panel surveys. In E. De Leeuw, J. Hox, & D. A. Dillman (Eds.), *International handbook of survey methodology* (pp. 479–499). New York: Lawrence Erlbaum Associates.

Silver, N. (2012, November 10). Which polls fared best (and worst) in the 2012 Presidential Race. Retrieved from: http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/.

Smith, T. W. (1995). Little things matter: A sampler of how differences in questionnaire format can affect survey responses. In American Statistical Association (Ed.), *Proceedings of the Joint Statistical Meeting, Survey Research Methods Section* (pp. 1046–1051). Washington, DC: AMSTAT.

Smyth, J. D., & Pearson, J. E. (2011). Internet survey methods: A review of strengths, weaknesses, and innovations. In M. Das, P. Ester, & L. Kaczmirek (Eds.), *Social and behavioral research and the internet. Advances in applied methods and research strategies* (pp. 11–44). New York: Taylor and Francis.

Sudman, S. & Wansink, B. (2002). *Consumer panels*. Chicago: American Marketing Association.

Taylor, H. (2007, January 15). The case for publishing (some) online polls. Retrieved January 1, 2013, from: http://www.pollingreport.com/ht_online.htm.

Taylor, H., Bremer, J., Overmeyer, C., Siegel, J. W., & Terhanian, G. (2001). The record of internet-based opinion polls in predicting the results of 72 races in the November 2000 U.S. Elections. *International Journal of Market Research*, *43*, 127–136.

Terhanian, G. (2013). Comment to the summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology*, *1*, 124–129.

Tourangeau, R., Conrad, F. C., & Couper, M. P. (2013). *The science of web surveys*. Oxford: Oxford University Press.

Van Ossenbruggen, R., Vonk, T., & Willems, P. (2006). Results of NOPVO. Paper presented at the Online panels, close up, Utrecht, Netherlands. Retrieved from: http://www.nopvo.nl/page0 /files/Results_NOPVO_English.pdf.

Vonk, T., van Ossenbruggen, R., & Willems, P. (2006). The effects of panel recruitment and management on research results: A study across 19 online panels. Paper presented at the Panel research 2006, ESOMAR, Barcelona.

Walker, R., Pettit, R., & Rubinson, J. (2009). A special report from the Advertising Research Foundation: The foundations of quality initiative: A five-part immersion into the quality of online research. *Journal of Advertising Research*, *49*, 464–485.

Wells, W. D. (1993). Discovery-oriented consumer research. *Journal of Consumer Research*, *19*, 489–504.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*, 930–941.

Williams, J. (2012). Survey methods in an age of austerity. Driving value in survey design. *International Journal of Market Research*, *54*, 35–47.

Yeager, D. S., & Krosnick, J. A. (2011). Does mentioning "some people" and "other people" in a survey question increase the accuracy of adolescents' self-reports? *Developmental Psychology*, *47*, 1674–1679.

Yeager, D. S., & Krosnick, J. A. (2012). Does mentioning "Some People" and "Other People" in an opinion question improve measurement quality? *Public Opinion Quarterly*, *76*, 131–141.

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, A. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples. *Public Opinion Quarterly*, *75*, 709–747.

Yeager, D. S., Larson, S. B., Krosnick, J. A., & Thompson, T. (2011). Measuring Americans' issue priorities: A new version of the most important problem question reveals more concern about global warming and the environment. *Public Opinion Quarterly*, *75*, 125–138.

YouGov. (2011). YouGov's record. Public polling results compared to other pollsters and actual outcomes. Retrieved January 1, 2013, from: http://cdn.yougov.com/today_uk_import/yg-archives -pol-trackers-record2011.pdf.