The Application of Evidence-Based Methods in Survey Methodology

Inauguraldissertation
zur Erlangung des akademischen Grades
eines Doktors der Sozialwissenschaften
der Universität Mannheim

Vorgelegt von

Jessica Christine Daikeler

Hauptamtlicher Dekan der Fakultät für Sozialwissenschaften:

Prof. Dr. Michael Diehl

Erstbetreuerin:

Prof. Dr. Michael Bosnjak

Zweitbetreuer:

Prof. Dr. Florian Keusch

Erstgutachter:

Prof. Dr. Florian Keusch

Zweitgutachter:

Prof. Dr. Michael Braun

Tag der Disputation:

21. Mai 2019

Acknowledgements

Many people have supported me during the work on this dissertation. First of all, I would like to thank my mentor and supervisor Michael Bosnjak. His intensive support and encouragement especially in the initial phase of this dissertation and his continuous feedback have made an essential contribution to the success of this dissertation. Furthermore, he has enhanced my understanding of research through his strong focus on experimental designs and meta-analytical research, which has opened up many opportunities for me.

I also thank Florian Keusch and Michael Braun for reviewing my thesis and providing me valuable feedback. Furthermore, I would like to thank my wonderful colleagues inside and outside of Gesis. First of all, Henning Silber with whom I had the luck to sit in an office for many years and who encouraged and supported me with so many things during my dissertation, and also co-authored two papers of this dissertation. Britta Gauly for her view of things as a non-survey methods person and her friendship. Thanks also to Tobias Gummer, working with him is both very instructive and fun. Ruben Bach for his valuable support in one of these papers, his wise general feedback and to lend me an ear. My thanks also go to Stephanie Eckman, who co-authored one of these papers and contributed significantly to its rapid success with excellent feedback and her structured way of working. I would also like to thank Katja Lozar-Manfreda for the co-authoring and her excellent feedback as well as Brady West for inviting me to spend a very productive time at the University of Michigan - Ann Arbor, and members of the MPSM Meth Lab for having me in their group!

Special thanks to Alex, who always supports me. You are the best!

Contents

Acknowledgements **i**

1 Introduction and Summary **1**

1.1 Classification of evidence-based research 1

1.2 Evidence-based survey methodology - A review of the literature 5

1.3 Why this dissertation? 7

1.4 Summary of chapters 9

2 Motivated Underreporting in Smartphone Surveys **15**

2.1 Abstract 15

2.2 Introduction 16

2.2.1 Misreporting in filter and follow-up questions 17

2.2.2 Response behavior in PC and smartphone surveys 19

2.2.3 Response behavior in different questions formats and devices 20

2.3 Data and methods 21

2.3.1 Check of randomizations 21

2.3.2 Data quality indicators 23

2.3.3 Analysis plan 26

2.4 Results 26

2.4.1 Triggered filter questions and follow-up data quality by question format (H1) and (H2) 27

2.4.2 Triggered filter questions and follow-up data quality by device (H3) and (H4) 28

2.4.3 Triggered filter questions and follow-up data quality in an interaction of device and format (H5) and (H6) 29

2.5 Discussion 31

2.6 Appendix 34

2.6.1 Survey invitation PC and smartphone 34

2.6.2 Text of filter and follow-up questions 34

2.6.3 Data quality indicators 36

2.6.4 Format, filter and interaction effects 37

3 Web Versus Other Survey Modes **41**

3.1 Abstract 41

3.2 Introduction 42

3.3 Background 43

3.4 Method 47

3.4.1 Eligibility criteria and search strategy 48

3.4.2 Coding procedures 49

3.4.3 Statistical method 51

3.5 Results 52

3.5.1 Study characteristics 52

3.5.2 Mean response rate difference: Web surveys versus other survey modes . 53

3.5.3 Moderator analysis: Replication 55

3.5.4 Moderator analysis: Extension 56

3.6 Discussion 58

3.6.1 Limitations and further research 62

3.7 Appendix 65

3.7.1 Search strategy 65

3.7.2 Variable overview 65

3.7.3 Robustness checks 68

3.7.4 Measurement overview 68

3.7.5 Publication bias and sensitivity analysis 69

3.7.6 Summary statistics of moderators 71

4 Which Country-Level Factors Are Associated With Web Survey Response Rates? 85

4.1 Abstract 85

4.2 Introduction 86

4.3 Country-specific predictors of web survey response rates 87

4.3.1 Social factors 88

4.3.2 Economic factors 89

4.3.3 Technological development 90

4.3.4 Survey participation propensity 90

4.4 The present study 91

4.5 Method 92

4.5.1 Overview of meta-analytic procedure 92

4.5.2 Eligibility criteria and search strategy 93

4.5.3 Statistical method and effect sizes 95

4.6 Results 97

4.6.1 Study characteristics and sensitivity 97

4.6.2 Cultural differences in web surveys 98

4.6.3 Country-specific predictors for the success of web surveys 100

4.7 Discussion 103

4.7.1 Practical implications 105

4.7.2 Limitations and further research 105

4.8 Appendix 107

4.8.1 Data sources 107

4.8.2 Heterogeneity of effect sizes 109

4.8.3 Robustness checks 109

5 How to Conduct Effective Interviewer Training: A Meta-Analysis 126

5.1 Abstract . . . 126

5.2 Introduction . . . 127

5.3 Conceptual development of research questions . . . 129

5.3.1 Effect of refusal avoidance training on survey response rates . . . 132

5.3.2 Effect of interviewer training on data quality . . . 132

5.3.3 Effect size heterogeneity . . . 133

5.3.4 Training features that may improve data quality . . . 133

5.4 Data and Methods . . . 135

5.4.1 Eligibility criteria and search strategy . . . 135

5.4.2 Coding procedure . . . 137

5.4.3 Effect size metric and statistical method . . . 137

5.4.4 Publication bias and sensitivity analyses . . . 139

5.5 Results . . . 140

5.5.1 What is the effect of interviewer training on data quality? (Q1– Q3) . . . 141

5.5.2 Moderator analysis: Which features render interviewer training successful? (Q4 and Q5) . . . 144

5.6 Conclusion and discussion . . . 147

5.7 Appendix . . . 151

5.7.1 Publication bias . . . 151

5.7.2 Coding . . . 153

5.7.3 Random effects model and meta regression summary statistics . . . 153

6 Conclusion and Discussion **161**

List of Tables

2.1 Example of filter questions in interleafed vs. grouped format 19

2.2 Randomization to format and device 23

2.3 Definition of data-quality indicators 24

2.4 Mean number of triggered filter questions by question format and device 28

2.5 Indicators of data quality in follow-ups, by device and format 30

2.6 Summary statistics for data-quality indicators 36

2.7 Regression outcomes 37

3.1 Meta-analytic summary statistics - random effects model without moderators . . 55

3.2 Meta-analytic summary statistics - random effects model - replication 57

3.3 Meta-analytic summary statistics - random effects Model - extension 59

3.4 Overview of study design characteristics 60

3.5 Comparison of search terms and search engines 65

3.6 Conference overview 65

3.7 Variable and moderator overview 67

3.8 Sampling error weighted mean response rate difference overview 68

3.9 Quality statistics for moderator analysis 71

4.1 Country-specific indicator: Sources 94

4.2 Social, economic, technological and survey participation propensity determinants for the success of web surveys 101

4.3 Data Sources 109

4.4 Heterogeneity differences in web and comparison mode 109

4.5 Robustness check by mode and US studies 111

4.6 Cultural dimension 112

5.1 Overview of the literature on interviewer tasks addressed in interviewer training experiments 131

5.2 Eligibility criteria 136

5.3 Description of effect sizes 140

5.4 Moderator overview 147

5.5 Publication bias check: Egger's regression test 152

5.6 Coding scheme 153

5.7 Sampling error weighted mean effect sizes and heterogeneity 154

List of Figures

1.1 The Evidence pyramid (Figure adapted from Haynes et al. (1997)). 3

1.2 From information to wisdom (Figure from Marshall (2013)). 4

1.3 Data source and research design overview of publications in four journals (2016-2019) - *based on own calculations* 6

2.1 Display of filter and follow-up questions on devices 22

2.2 Mean percentage and standard deviations in percent of data quality indicators in follow-up questions 25

2.3 Triggered filter questions by format and device (in %) 29

2.4 Data quality in follow-up questions 31

2.5 Survey invitation PC 34

2.6 Survey invitation smartphone 34

3.1 PRISMA literature search flow diagram 50

3.2 Cumulative forest plot 54

3.3 Forest plot of significant categorical moderators 60

3.4 Funnel plot 69

3.5 Normal quantile plot 70

4.1 Macro-level factors for web response 88

4.2 Comparison mode and survey country overview 98

4.3 Response rate overview across countries 99

5.1 The literature search process 138

5.2 Forest plots for data quality indicators: Trained vs. untrained interviewers . . . 142

5.3 Forest plots for data quality indicators: Trained vs. untrained interviewers . . . 145

5.4 Publication bias: Funnel plots for data quality indicators 152

Chapter 1

Introduction and Summary

This dissertation is dedicated to the application of evidence-based methods in survey research. Although survey research is a relatively young discipline, knowledge and contradictory findings abound in this field, as in other disciplines. In the first section I will provide a general introduction to evidence-based research, followed by an overview of evidence-based research in the field of survey methodology. Then, I will set up the motivation for my dissertation and provide a summary of each chapter.

1.1 Classification of evidence-based research

"Non-reproducible single occurrences are of no significance to science." - Popper (1956, p.66)
With this quote Karl Popper already named in 1956 a highly relevant issue in science – the replicability of scientific studies. Particularly in the last decade, key results of many scientific studies in the social and life sciences have been difficult or impossible to replicate. Researchers have had trouble replicating their own work and this of others – this phenomenon is also known as the replication crisis (Baker 2016).

Auspurg and Brüderl (2019) have studied replications in sociology and identify four main sources of error in the social sciences that prevent science from producing valid, robust and

replicable knowledge. First, "bad" measurement such as inadequate question wording, second, invalid and/or unreliable measurement such as inadequate question type, third, bad research design and analysis strategy such as inadequate statistical analysis, and finally "bad" researchers (errors). This includes "biased" researchers who conduct, for instance, non-objective research. Examples of "biased" research includes fraud, collection and falsification of data and/or analysis or questionable research practices as well as "p-hacking". One further reason for the replication crisis is the 40 percentage points lower publication probability of null results in journals compared to a highly statistically significant result and a 60 percentage point lower probability of documenting the results and writing them down (Auspurg and Brüderl 2019).

One consequence of the replication crisis is the renewed focus on evidence-based research practices (EBP) in many social science disciplines (Shaw and D'Intino 2017; Thyer 2004). The goal of the EBP approach is to emphasize the practical application of the best available research procedures. This means that for practical interventions, only those scientific studies are referred to that use the best available research design and analysis strategy (e.g., the usage of randomized controlled trials versus observational studies) (Popper 1956). In the life sciences the EBP movement started in the early 1990s (Zimerman 2013) and other research disciplines followed. This led to scientific movements such as evidence-based education (Pring and Thomas 2004), evidence-based management (Rousseau 2012), evidence-based criminology (Farrington et al. 2003), evidence-based software engineering (Dyba, Kitchenham, and Jorgensen 2005), and finally, evidence-based psychology (APA 2006).

A cornerstone of the EBP approach is the hierarchical system of classifying the degree of evidence in the evidence pyramid (Figure 1.1). The evidence pyramid is a ranking system that describes the reliability of the results measured in research studies (Haynes et al. 1997). The higher the hierarchy of the study design in the evidence pyramid, the stricter the methodology and thus the higher the likelihood that study design can minimize the impact of bias on study results (Paul and Leibovici 2014). Most versions of the pyramid clearly represent a hierarchy of internal validity (risk of distortion). There exist different versions of the evidence pyramid, but all focus on showing weaker study designs in the lower range (expert opinion and case studies), followed by case control and cohort studies in the middle, then randomized controlled

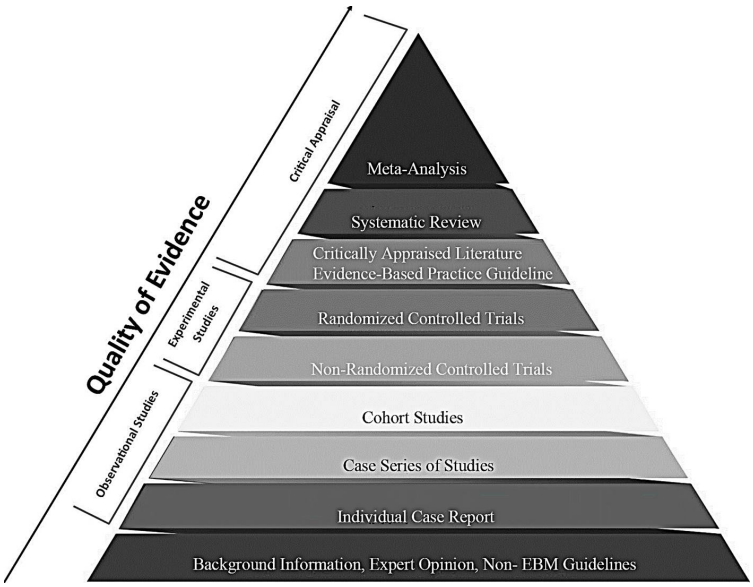


Figure 1.1: The Evidence pyramid (Figure adapted from Haynes et al. (1997)).

trials (RCTs) and at the top, systematic reviews and meta-analyses (Hoffmann, Bennett, and Del Mar 2013, pp.44).

Since there are various definitions of meta-analysis, this dissertation adopts the definition of systematic review and meta-analysis from Green et al. (2008) which reads as follows "A systematic review intended to appraise and synthesize the best available evidence on a defined research question by collecting and summarizing all empirical evidence that fits pre-specified eligibility criteria. A meta-analysis is the use of statistical methods to summarize the results of the included evidence." The description of the evidence pyramid is intuitive and probably correct in many cases (Paul and Leibovici 2014). Some approaches have challenged the placement of systematic reviews and meta-analyses at the top of the pyramid, as for instance, heterogeneity of the included primary research studies (methodological or statistical) is an inherent limitation of meta-analyses that can be minimized or explained but never eliminated (Dechartres et al. 2014).

Meta-analysis can be described as a set of statistical methods for aggregating, summarizing and drawing conclusions from sets of thematically related studies (Bosnjak 2018). In particular, the increasing number of varying research results (Larsen and Von Ins 2010; Michels and Schmoch 2012) can lead to the concealment of true relationships and so aggregating the research results in meta-analyses can help to draw a conclusion. Furthermore, the accumulation

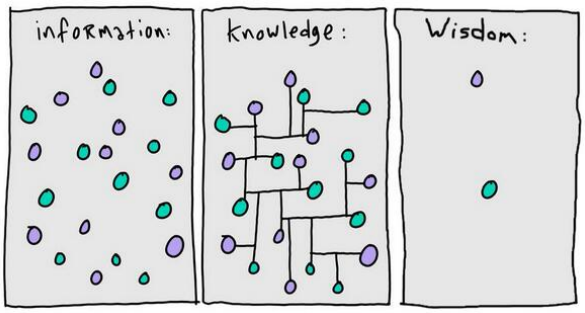


Figure 1.2: From information to wisdom (Figure from Marshall (2013)).

of individual pieces of information could create knowledge and this knowledge could ultimately be reduced to fundamentals to draw best practice recommendations (see Figure 1.2). A meta-analysis accumulates the evidence of individual primary studies, cancels out sampling errors associated with individual studies, calculates an overall effect size, permits to draw conclusions on research gaps and research quality, estimates the heterogeneity of the studies and tries to explain this heterogeneity using moderators. Several types of study designs can be included in a meta-analysis and the strongest types of empirical evidence are meta-analyses of randomized experiments (see Figure 1.1). However, not every social science discipline can rely on enough comparable studies and particularly on randomized controlled trials.

Survey methodological research is still a rather young research discipline, and as in other research disciplines, both knowledge and contradictory findings accumulate (Bosnjak 2018). Therefore, the establishment of evidence-based methodology and the implementation of systematic reviews and meta-analyses is only a logical consequence. A key concept of methodological survey research is to generate knowledge that supports survey operations in designing and implementing survey projects (Bosnjak 2018). In this context, survey methodology is structurally similar to other disciplines that are committed to generate the best empirical evidence and use it to guide (survey) operational actions (Bosnjak 2018). This dissertation is dedicated to high level evidence-based research in terms of the evidence-pyramid and therefore, focuses on the implementation of randomized controlled trials (RCT) and meta-analyses in the field of survey methodology. The next section will give an overview of these two approaches in survey methodology.

1.2 Evidence-based survey methodology - A review of the literature

A major advantage of survey methodology research is that for instance, in contrast to many sociological studies (Weiß and Wagner 2008; Wagner et al. 2019), experiments are often possible. Surprisingly, however, this possibility of causal research is only used in about 40 percent of the published studies. This is illustrated by an analysis of 426 survey methodological articles from four survey methodological journals (*Journal of Survey Statistics and Methodology, Social Science Computer Review, Sociological Methods & Research, International Journal of Public Opinion Research*) in 2016-2019 (see Figure 1.3). The rather small share of studies that are experimental – 40 percent – in survey methodology is surprising, but the planning and implementing of experiments is strongly linked with the opportunity of primary data collections and our analysis shows that in about 64 percent of the studies no primary data collection took place. In addition, large (longitudinal) survey programs are often reluctant to implement methodological experiments because they fear for the longitudinal comparability of their data and systematic biases though some panels establish other venues to test experimental designs, e.g., the Socio-Economic Panel Study (SOEP) innovation panel (Richter and Schupp 2012). Furthermore, experimental designs are not possible in all areas of survey methodology due to organizational and cost restrictions (e.g. randomization of interviewer and interviewee characteristics), but an area as applied as survey methodology research certainly allows further possibilities for establishing experimental designs and thus high-class evidence.

However, not only for randomized controlled trials, but also for meta-analyses and systematic reviews, the potential in survey methods research still seems far from exhausted. My analysis shows that between 2016 and 2019 only nine percent (11 studies) of the published studies were systematic reviews and meta-analyses in the four journals I examined (see Figure 1.3). Admittedly psychology is a much larger field, but Borman and Grigg (2009) identified 60 (26%) meta-analyses in the journal *Psychological Bulletin* alone in the years 2000 to 2005. Nevertheless, the presence of meta-analyses in survey methodology seems to be slowly increas-

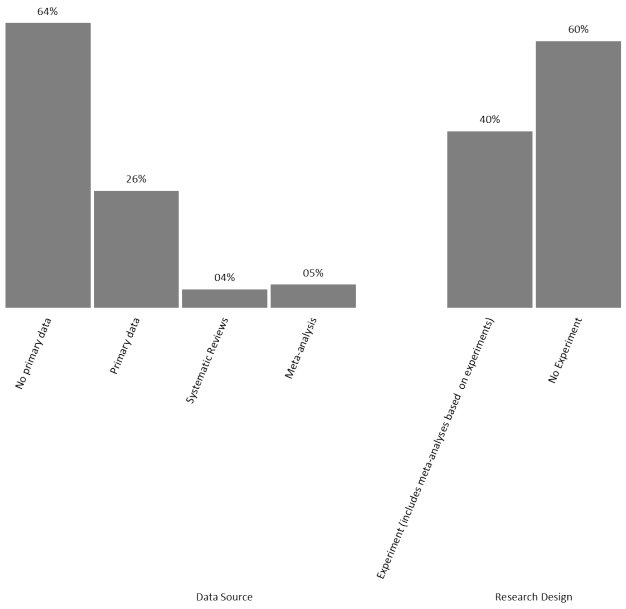


Figure 1.3: Data source and research design overview of publications in four journals (2016-2019) - *based on own calculations*

ing. Among the meta-analyses identified by Cehovin, Bosnjak, and Lozar Manfreda (2018) in their systematic review, the number of such publications remained low until 1998 with a total of 12 published manuscripts, then slowly increased between 1999 and 2008 (18 published meta-analyses), and, finally 24 meta-analyses have been published after 2009. Considering the expanding number of publications in the field of survey methodology, we can assume a slight increase.

Furthermore, Cehovin, Bosnjak, and Lozar Manfreda (2018) identified obvious meta-analytic research gaps along the total survey error framework. In particular, research questions on nonresponse and measurement errors have been meta-analytically addressed, while sampling, coverage and processing errors have never been addressed. Moreover, almost none of the survey design characteristics that are not under the control of the researcher (i.e. social environment, interaction between researcher and respondent, and respondent characteristics) were analyzed. Consequently, the methods of systematic reviews and meta-analyses (of experiments) still offer numerous unexplored research potentials, the benefits and disadvantages of which I will examine in more detail in the next section and illustrate in the course of this dissertation.

1.3 Why this dissertation?

Survey methodological research as a discipline is surprisingly behind compared to other disciplines in the implementation of quantitative instruments for the systematic synthesizing of evidence (Bosnjak 2018). Admittedly, there are many advantages but also some challenges in practising evidence-based research and especially conducting meta-analyses (Green et al. 2008; Borenstein et al. 2009).

There are some challenges in the implementation of experimental variation in the field of survey methodology, I will start with naming a few. First, there are many situations in which the application of experiments is impossible, such as a random allocation of certain respondents' attitudes and characteristics. Second, even if randomisation is possible, such as for the allocation to a survey mode, respondents always have the opportunity to refuse the participation at all. Third, there are still few opportunities in survey research to carry out experimental interventions with population representative (longitudinal) data. The reason for this is that study initiators often fear limitations in the comparability of responses. In the following, I will describe the challenges of using meta analyses in survey method research.

First, an often addressed criticism of meta-analyses is the confusion of "apples and oranges" (Borenstein et al. 2009, pp.357, p.379). This means that differences between individual studies may be lost during the accumulation of evidence in order to collect enough data for the analysis. This would result in heterogeneous and incomparable studies. However, in most cases the heterogeneity of the studies can be addressed with the help of sub-analysis or moderator analysis. The heterogeneity of the studies, in the sense of different research designs, can often be used to explain heterogeneous outcomes and helps to draw conclusions about key covariates (also known as moderators in meta-analytical literature (Borenstein et al. 2009, p.187)).

Second, another point of concern with meta-analyses is summarized under the term "garbage in, garbage out" (Borenstein et al. 2009, pp.380). This point of concern addresses the fact that the accumulation of qualitatively lower outcomes can only result in qualitatively inferior overall findings, i.e. meta-analysis can be understood as a process of "waste management". This

criticism can be prevented with either quality-focused eligibility criteria (e.g. only randomized controlled trials) or by introducing the quality weighting mechanism such as coding the quality of studies afterwards to include as moderators/ independent variables (e.g. student sample vs. general population sample) in the model. As a consequence of the latter approach, heterogeneity between studies can be explained.

Third, missing data and publication bias is a common criticism of meta-analyses (Borenstein et al. 2009; Lipsey and Wilson 2001, pp.263, 378). It is based on the assumption that significant results that agree with theoretical approaches have a higher probability of being published and thus meta-analyses are subject to a so-called publication bias, because they represent especially significant findings. In an attempt to asses this publication bias, the meta-analytical toolkit offers a whole set of tools such as the illustration of a possible bias through plotting the effect sizes from individual studies against the sample size in scatter plots (so called "funnel plot") and statistical correction.

Fourth, one of the most challenging issues in meta-analytic research is the synthesis of multivariate outcomes. While bi-variate outcomes can be accumulated without major problems, meta-analyses quickly reach their limits when accumulating evidence from regression models or structural equation models (Borenstein et al. 2009).

Researchers should not expect to produce a conclusive, debate-ending result by conducting a meta-analysis on an existing literature. Instead, meta-analyses may serve best to draw attention to the existing strengths and/or weaknesses in results and can therefore inspire a careful reexamination of methodology and theory followed by, if necessary, large-scale, preregistered replication efforts (Carter et al. 2017).

Systematic review work is also needed to give recommendations for fieldwork and ensure that the field does not replicate the same research questions repeatedly (Cehovin, Bosnjak, and Lozar Manfreda 2018). This "freed" capacity would then open up new research potential on such topics as behavioral data, mobile data collection, and social media, just to name a few.

Building on this rationale, this dissertation will apply evidence-based methods such as random-

ized controlled trials and meta-analyses (of randomized controlled trials) in survey methodological research.

By doing this, this dissertation has two objectives, first to derive recommendations for survey implementation from evidence-based practice in the context of meta-analyses and randomized controlled trials, and second to demonstrate the applicability of randomized controlled trials and meta-analyses in survey methodology research.

The next three chapters of this dissertation focus on survey mode effects and the fourth study on interviewer training. The first study is a randomized controlled trial and addresses data quality in filter questions on PC versus smartphone devices. The second study is a meta-analysis of randomized controlled trials in which response rates of web surveys are compared with those of other survey modes. Furthermore, this study assesses survey characteristics and their effect on the response rate of web surveys. In the third study, a meta-analysis is presented which examines web response rates in a cross-country context and identifies cross-country factors that influence web response rates. This chapter summarizes in which countries web response rates are comparatively high and why. While study one addresses measurement error, study two and three address nonresponse error, and the fourth study addresses both – nonresponse and measurement error. In this fourth study, the effect of interviewer training on nonresponse and measurement error is examined meta-analytically, investigating which training characteristics influence this effect. The final chapter summarizes the findings and provides a discussion of the application of evidence-based methods in survey methodology.

The next section will now provide a detailed summary on each of the four chapters of this dissertation.

1.4 Summary of chapters

The study in chapter 2 ("Motivated Underreporting in Smartphone Surveys") reports on a classical example for a randomized control trial in survey methodological research and deals with the implementation of filter questions in smartphone surveys. Filter questions are a pop-

ular survey design instrument as they allow in-depth questioning via follow-ups and shorten the questionnaire for respondents who have nothing to report. Filter questions can be asked in either in the *interleafed* (follow-ups immediately after the filter for a given item) or the *grouped* (follow-ups after filter question block) format. Tests of the underlying underreporting (underreporting means the misreporting of facts, especially the non-disclosure of facts) mechanism has shown that motivated underreporting arises from respondents' desires to reduce the burden of the survey. Since conducting a survey on the smartphone may be more burdensome than on a PC due to the smaller screen size, longer page loading times, and more distraction, I expect that motivated underreporting is more pronounced on smartphones. Furthermore, there is only sparse knowledge on data quality in the follow-up questions to filter questions. Since respondents in the interleafed format might know after answering the first question affirmatively that every affirmative answer triggers follow-up questions, I expect respondents in the interleafed format to provide higher data quality in the follow-ups. In addition, I hypothesize this effect to be further enlarged by the mobile device usage, as it is much more burdensome to answer the follow-up questions on the smartphone. I randomly assigned 3,517 respondents of a German online access panel to take the survey either on the PC or the smartphone. My results show that mobile respondents do trigger the same number of filter questions than PC respondents. However, I found that mobile respondents provide lower data quality in terms of more item-nonresponse, heaping (heaping means the usage of rounded values in open questions), and middle category responses in the follow-ups, especially in the grouped format. Furthermore, I found that respondents in the interleafed format provide better data quality compared to the grouped format. I conclude with recommendations for web survey designers.

Chapter 3 ("Web Versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates") also focuses on web surveys; however it addresses the participation decision and includes a meta-analysis of randomized controlled trials of mode studies of web survey response rate comparisons. First, I focus on whether web surveys still yield lower response rates compared to other survey modes. To answer this question, I replicated and extended a previous meta-analysis by Lozar Manfreda et al. (2008) which found that, based on 45 experimental comparisons, web surveys had an 11 percentage point lower response rate compared to

other survey modes. Since the publication of this initial meta-analysis, fundamental changes in Internet accessibility and use suggest that web survey participation propensities have changed considerably. However, in my replication and extension encompassing 114 experimental mode comparisons, I found almost no changes: Web surveys still yield lower response rates than other modes (a 12 percentage point response rate difference). Second, I found that prenotifications, the sample recruitment strategy, the survey's invitation mode, the type of target population, the number of contact attempts, and the country in which the survey was conducted moderated the magnitude of the response rate differences. I conclude with substantial implications for both survey methodology and survey operations involving web surveys.

The 4th chapter ("Which Country-Level Factors Are Associated With Web Survey Response Rates? A Meta-Analysis") assesses web surveys from a cross-cultural perspective. A major challenge in web-based cross-cultural data collection is variation in response rates, which can result in low data quality and nonresponse bias. Country-specific social, economic, and technological factors as well as the willingness of the population to participate in surveys may affect web response rates. This study attempts to evaluate web survey response behavior with meta-analytical methods based on more than 100 experimental studies from seven countries. Three effect sizes (web response rate, response rate of the comparison mode, and response rate difference) are used. Three country-specific factors had an impact on the performance of web survey response rates. Specifically, web surveys achieve high response rates in countries with a high population growth, high internet coverage, and a high survey participation propensity, whereas they are at a disadvantage in countries with a high population age and smartphone coverage. The chapter concludes with practical implications for cross cultural survey research.

The 5th chapter of this dissertation ("How to Conduct Effective Interviewer Training: A Meta-Analysis") turns away from survey mode experiments and deals meta-analytically with interviewer training. Although interviewer training is part of every interviewer-administered study, this topic has so far been addressed surprisingly sparsely. Interviewer training can improve the performance of interviewers, and, thus, also the quality of survey data. However, the question how effective interviewer training is for data quality and, more importantly, which determinants make a training successful remains open. This research uses meta-analytical methods

to evaluate both the improvements in data quality caused by interviewer training and which training determinants are successful in improving the interviewer's performance. In this fifth chapter I refer to various aspects of data quality, namely unit nonresponse, item nonresponse, and correctly administered, read, probed and recorded questions and answers. In 66 experimental comparisons, I find that advanced interviewer training reduces unit and item nonresponse, increases correct probing, administration, reading, and recording of items with up to 40 percentage points. I also find that using a broad variety of training methods, such as blended learning, exercise and feedback sessions, interviewer monitoring and supplementary training material reinforces this effect.

This dissertation concludes with a discussion on the application of evidence-based methods in survey methodology.

References

APA (2006). "Evidence-based practice in psychology". In: *The American Psychologist* 61.4, p. 271.

Auspurg, Katrin and Josef Brüderl (2019). *Is there a credibility crisis in sociology? And if yes, what can be done?* MZES Open Social Science Conference 2019. Conference Presentation. URL: https://www.mzes.uni-mannheim.de/openscience/wp-content/uploads/2019/01/Auspurg_Br%5C%C3%5C%BCderl-Credible-Sociology-Mannheim.pptx (visited on 03/26/2019).

Baker, Monya (2016). "1,500 scientists lift the lid on reproducibility". In: *Nature News* 533.7604, p. 452.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein (2009). *Introduction to meta-analysis*. John Wiley and Sons. 457 pp. ISBN: 9780470057247.

Borman, Geoffrey D and Jeffrey A Grigg (2009). *Visual and narrative interpretation*. Russell Sage Foundation.

Bosnjak, Michael (2018). "Evidence-based survey operations: Choosing and mixing modes". In: *The Palgrave Handbook of Survey Research*. Springer, pp. 319–330.

Carter, Evan, Felix Schönbrodt, Will M Gervais, and Joseph Hilgard (2017). *Correcting for bias in psychology: A comparison of meta-analytic methods*. PsyArXiv.

Cehovin, Gregor, Michael Bosnjak, and Katja Lozar Manfreda (2018). "Meta-analyses in survey methodology: A systematic review". In: *Public Opinion Quarterly* 82.4, pp. 641–660. ISSN: 0033-362X. DOI: 10.1093/poq/nfy042. URL: https://dx.doi.org/10.1093/poq/nfy042.

Dechartres, Agnes, Douglas G Altman, Ludovic Trinquart, Isabelle Boutron, and Philippe Ravaud (2014). "Association between analytic strategy and estimates of treatment outcomes in meta-analyses". In: *Jama* 312.6, pp. 623–630.

Dyba, Tore, Barbara A Kitchenham, and Magne Jorgensen (2005). "Evidence-based software engineering for practitioners". In: *IEEE software* 22.1, pp. 58–65.

Farrington, David P, Doris Layton MacKenzie, Lawrence W Sherman, and Brandon C Welsh (2003). *Evidence-based crime prevention*. Routledge.

Green, Sally, Julian PT Higgins, Philip Alderson, Mike Clarke, Cynthia D Mulrow, Andrew D Oxman, et al. (2008). *Cochrane handbook for systematic reviews of interventions*. John Wiley & Sons Chichester, UK.

Haynes, R Brian, David L Sackett, W Scott Richardson, William Rosenberg, and G Ross Langley (1997). "Evidence-based medicine: How to practice & teach EBM". In: *Canadian Medical Association. Journal* 157.6, p. 788.

Hoffmann, Tammy, Sally Bennett, and Christopher Del Mar (2013). *Evidence-based practice across the health professions-e-Book*. Elsevier Health Sciences.

Larsen, Peder and Markus Von Ins (2010). "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index". In: *Scientometrics* 84.3, pp. 575–603.

Lipsey, Mark W and David B Wilson (2001). "Analysis issues and strategies". In: *Practical Meta-Analysis*. Ed. by MW Lipsey and DB Wilson. Thousand Oaks, CA: SAGE Publications, Inc, pp. 105–128.

Lozar Manfreda, Katja, Michael Bosnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar (2008). "Web surveys versus other survey modes: A meta-analysis comparing response rates". In: *Journal of the Market Research Society* 50.1, p. 79. ISSN: 0025-3618.

Marshall, Bob (2013). *Information, knowledge, wisdom: pic.twitter.com/CvrJrBm29g*. microblog. URL: https://twitter.com/flowchainsensei/status/408167162344648704/photo/1 (visited on 01/20/2014).

Michels, Carolin and Ulrich Schmoch (2012). "The growth of science and database coverage". In: *Scientometrics* 93.3, pp. 831–846.

Paul, M and L Leibovici (2014). "Systematic review or meta-analysis? Their place in the evidence hierarchy". In: *Clinical Microbiology and Infection* 20.2, pp. 97–100.

Popper, Karl (1956). *The logic of scientific discovery*. Routledge.

Pring, Richard and Gary Thomas (2004). *Evidence-based practice in education*. McGraw-Hill Education (UK).

Richter, David and Jürgen Schupp (2012). *SOEP Innovation Sample (SOEP-IS)—Description, structure and documentation*. SOEPpaper.

Rousseau, Denise M (2012). *The Oxford handbook of evidence-based management*. Oxford University Press.

Shaw, Steven R and Joseph D'Intino (2017). "Evidence-based practice and the reproducibility crisis in psychology". In: *Communique, 45 (5)*, pp. 1–21.

Thyer, Bruce A (2004). "What is evidence-based practice?" In: *Brief treatment and crisis intervention* 4.2, p. 167.

Wagner, Michael, Clara H. Mulder, Bernd Weiß, and Sandra Krapf (2019). "The transition from living apart together to a coresidential partnership". In: *Advances in Life Course Research* 39, pp. 77–86. ISSN: 1040-2608. DOI: https://doi.org/10.1016/j.alcr.2018.12.002. URL: http://www.sciencedirect.com/science/article/pii/S104026081830073X.

Weiß, Bernd and Michael Wagner (2008). "Potentiale und Probleme von Meta-Analysen in der Soziologie". In: *Sozialer Fortschritt* 57.10, pp. 250–256.

Zimerman, Ariel L (2013). "Evidence-based medicine: A short history of a modern medical movement". In: *AMA Journal of Ethics* 15.1, pp. 71–76.

Chapter 2

Motivated Underreporting in Smartphone Surveys

2.1 Abstract

Filter questions are a popular survey instrument as they allow in-depth questioning by follow-ups and shorten the questionnaire for respondents who have nothing to report. Basically, filter questions can be asked in either the *interleafed* (follow-ups immediately after the filter) or the *grouped* (follow-ups after filter question block) format. Testing for the underlying underreporting mechanism has shown that motivated underreporting arises from respondents' desire to reduce the burden of the survey. Since conducting a survey on the smartphone is more burdensome than on the PC due to the smaller screen size, longer page loading times, and more distraction, we expect that motivated underreporting is more pronounced on smartphones. Furthermore, in the filter question literature there is only sparse knowledge on data quality in the follow-up questions to filter questions. Since respondents in the interleafed format know after answering the first question affirmatively that every affirmative answer triggers follow-up questions, we expect respondents in interleafed format to provide higher data quality in the follow-ups. In addition, we hypothesize this effect to be further enhanced by the device used. We randomly assigned 3,517 respondents of a German online access panel to either the PC or

the smartphone. Our results show that mobile respondents do not trigger fewer filter questions than PC respondents and respondents in the interleafed format provide better data quality in the follow-ups compared to the grouped format. However, we found that mobile respondents provide lower data quality in terms of more item nonresponse and heaping in the follow-ups, especially in the grouped format. We conclude with recommendations for web survey designers.

2.2 Introduction

Many surveys use eligibility questions to ask respondents only those questions that apply to them. For example, asking unemployed respondents about working hours or salary is meaningless as these follow-up questions do not apply to those respondents. Instead, asking irrelevant questions increases response burden and may leave respondents confused and less willing to complete the rest of the survey. Response burden means the degree to which a respondent perceives participation in a survey as difficult, time consuming, or emotionally stressful (Lavrakas 2008). With the usage of filter questions all eligible respondents are presented with the follow-up questions, while ineligible respondents are routed around those questions and continue with the rest of the survey.

Evidence-based recommendations such as preferring the grouped filter question format have frequently been investigated for face-to-face, telephone, mail and web surveys in the past (Eckman and Kreuter 2018; Kreuter et al. 2011). However, it has not yet been determined whether the previous recommendations can also be applied to mobile web surveys as responding on smartphones is in general more burdensome and skipping follow-up questions would reduce survey burden dramatically. Furthermore, there is only sparse evidence even for traditional modes what the consequences would be for the data quality of the follow-up questions.

In this paper, we capture this research gap and will first examine whether we can replicate the already found effects of the question format for the filter and follow-up questions, secondly whether misreporting in filter and follow-up questions is more pronounced on smartphones, and thirdly whether there is an interaction of question format and device for filter and follow-up

questions. We begin by describing the literature on misreporting and data quality in filter and follow-up questions, mobile devices, and the interaction of format and device. In the following we describe our data, methods and data quality indicators and perform the analyses. Finally, we provide field recommendations for the usage of filter questions in PC and smartphone surveys.

2.2.1 Misreporting in filter and follow-up questions

While numerous studies have already investigated data quality in filter questions, the findings on follow-ups are rather rare. This section summarizes the findings for both filter and follow-up questions and provides our first two hypotheses.

Response behavior to filter questions

Filter questions are often asked to route respondents around follow-up questions that do not apply to them. Such filter questions are found in the US Consumer Expenditure Survey[1] asking for clothing purchases and, if applicable, follow-up questions about the clothes bought, or in the US National Crime and Victimization Survey asking whether the respondent was the victim of a crime and, if applicable, details about the crime. In Germany, filter questions are used in surveys such as the household panel study "Labor Market and Social Security" asking for children and, if applicable, follow-up questions on each child (Kosyakova, Skopek, and Eckman 2014).

While filter and other forms of eligibility questions such as screening questions arguably improve survey designs and reduce response burden, they can also increase measurement error. If asked in certain formats, the structure of filter and follow-up questions allows respondents to foresee that triggering a filter will result in additional questions and increase survey burden. As a result, some respondents reduce the burden of the survey by misreporting to filter questions to avoid the follow-up questions (Eckman et al. 2014).

[1] e.g., https://www.bls.gov/cex/capi/2017/2017-CEQ-CAPI-instrument-specifications.pdf

Several studies have demonstrated such motivated misreporting by comparing responses to filter questions asked in two formats (e.g. Kessler et al. 1998; Duan et al. 2007; Kreuter et al. 2011; Eckman et al. 2014; Bach and Eckman 2018; Kreuter, Eckman, and Tourangeau 2019; Bach, Eckman, and Daikeler 2019). The *interleafed* format asks a filter question with the follow-up items (if applicable) following immediately. The *grouped* format asks all filter questions first before asking the follow-up questions that apply (for an illustration, see Table 2.1). In the interleafed format, respondents can learn that triggering a filter results in additional questions, while it is not possible to foresee the follow-up questions in the grouped format. Comparing the two formats, has shown that respondents trigger, on average, fewer filters in the interleafed format than in the grouped format. Moreover, a comparative study that used administrative records for validation demonstrated that the differences in reporting between the two formats are in fact due to respondents underreporting in the interleafed format in order to reduce the burden of the survey (Eckman et al. 2014). That is, respondents intentionally misreport filters in order to skip follow-up questions and reduce the burden of the survey.

A theoretical approach why respondents would like to reduce the burden of a survey might involve "optimizing" and "satisficing" (Krosnick 1991). Answering a survey requires respondents to invest a substantial cognitive effort in little or no reward, so respondents consider strategies to reduce and optimize survey effort, these strategies are known as "optimizing" and "satisficing" (Krosnick 1991). One way to do this in filter questions is to avoid triggering filter questions in order to bypass follow-up questions and shorten the survey. Since respondents in interleafed format quickly learn how follow-up questions can be bypassed, we expect to replicate the already well-know effect of less triggered filter questions in interleafed question format.

H1: Respondents in the interleafed format trigger fewer filter questions.

Response behavior to follow-up questions

Researchers who rely on survey data are not only interested in responses to filter questions, but also in responses to the follow-up questions. Therefore, our second research goal is to compare the data quality in the follow-up questions for PC and smartphones. We are aware of only

Table 2.1: Example of filter questions in interleafed vs. grouped format

Interleafed version	**Grouped version**
In the past 3 months, have you purchased a coat?	In the past 3 months, have you purchased a coat?
Please briefly describe the most recent coat you purchased.	In the past 3 months, have you purchased a shirt?
For whom was it purchased?	In the past 3 months, have you purchased pants?
In what month did you purchase it?	In the past 3 months, have you purchased a suit?
How much did it cost?	In the past 3 months, have you purchased a dress?
In the past 3 months, have you purchased a shirt?	FOR EACH YES
Please briefly describe the most recent shirt you purchased.	Please briefly describe the most recent [item] you purchased.
[...]	For whom was it purchased?
In the past 3 months, have you purchased a suit?	In what month did you purchase it?
[...]	How much did it cost?
[...]	

Note: Table adapted from Kreuter et al. (2011)

one study (Kreuter et al. 2011) that examined data quality in follow-ups so far. In this study, Kreuter et al. (2011) compared item nonresponse to follow-up questions in a telephone survey between the grouped and interleafed format and found more item nonresponse in the grouped format. That is, respondents in the grouped format trigger more filter questions, but then respond to fewer follow-up questions. We expect to replicate this effect for other data quality indicators than item nonresponse

H2: Respondents in the interleafed format provide better data quality in the follow-up questions than respondents in the grouped format.

2.2.2 Response behavior in PC and smartphone surveys

The desire to reduce the burden of the survey seems to be especially relevant in the context of web surveys as the reduction of burden is quite simple because no interviewer is involved. Respondents use a variety of device types to participate in web surveys (e.g., desktop PCs, laptops, tablets, or smartphones) and the usage of a specific device is known to influence both response burden and behavior (Antoun, Couper, and Conrad 2017; Keusch and Yan 2017). Methodological research that compares the various devices has shown that the response behavior is relatively similar when respondents complete the survey on their PCs, laptops, or

tablets. Taking a survey on smartphones, however, can sometimes lead to some more differences in response behavior (e.g., de Bruijne and Wijnant 2013; Gummer and Rossmann 2015; Antoun, Couper, and Conrad 2017; Schlosser and Mays 2018; Tourangeau et al. 2018).

Mobile respondents were at least as likely to provide conscientious and thoughtful answers and to disclose sensitive information on smartphones as on PCs (Antoun, Couper, and Conrad 2017). They provided no substantial data quality differences in terms of item nonresponse, straightlining, scale reliability, and validity (Tourangeau et al. 2018). Furthermore, mobile respondents did not differ in break-off rate during the survey, item nonresponse, and length of responses to open-ended questions (Schlosser and Mays 2018).

Mobile respondents do not seem to perceive the length of surveys as long as PC respondents (de Bruijne and Wijnant 2013), however it takes them longer to answer a questionnaire (Keusch and Yan 2017; Schlosser and Mays 2018). Moreover, smartphone respondents have more problems in executing survey tasks such as using small sliders and date-picker wheels (Antoun, Couper, and Conrad 2017).

H3: Smartphone respondents trigger fewer filter questions than PC respondents.

H4: Smartphone respondents provide lower data-quality in the follow-up questions than PC respondents.

2.2.3 Response behavior in different questions formats and devices

While there are already some findings on the effect of the filter question format and some findings on response quality on different devices, the effect of the two filter question formats - interleafed and grouped, has never been analyzed for different devices. We aim to close this research gap on misreporting for different devices and expect the least triggered filter questions for mobile respondents in the interleafed question format, as they have a higher burden responding on the smartphone and can easily bypass follow-ups in the interleafed question

format. Furthermore, for the follow-up questions we expect to find an interaction - the lowest data quality for smartphone respondents allocated to the grouped format.

H5: Smartphone respondents in the interleafed filter question format trigger fewer filter questions than respondents in the grouped format or PC respondents.

H6: Smartphone respondents in the grouped question format provide lower data quality compared to respondents in the interleafed format or PC respondents.

2.3 Data and methods

To test our hypotheses, we conducted a web survey where we experimentally varied both filter question format and device. In the following subsections, we describe our data and our data quality indicators.

2.3.1 Check of randomizations

Random allocation of respondents to device and format was intended to remove all differences between the groups, so that any resulting response differences would be due to the experimental manipulation, and not to the characteristics of the respondents. To check that the randomization worked as intended, we applied logistic regression and predicted the format (interleafed vs.grouped) and the device used (PC vs. smartphone). As independent variables, we selected all of our socio-demographic information given for our respondents as well as two paradata measures - survey duration and invitation date. We could not use other variables as they were influenced by the various survey methodological experiments. We selected duration to exclude the risk of slower respondents self-selecting into a particular device and thus differing from faster respondents; the same applies to the participation date for late versus early respondents.

The results of these models are shown in Table 2.2. In the first column, we see that the randomization of the question format worked well: across all respondent characteristics, we

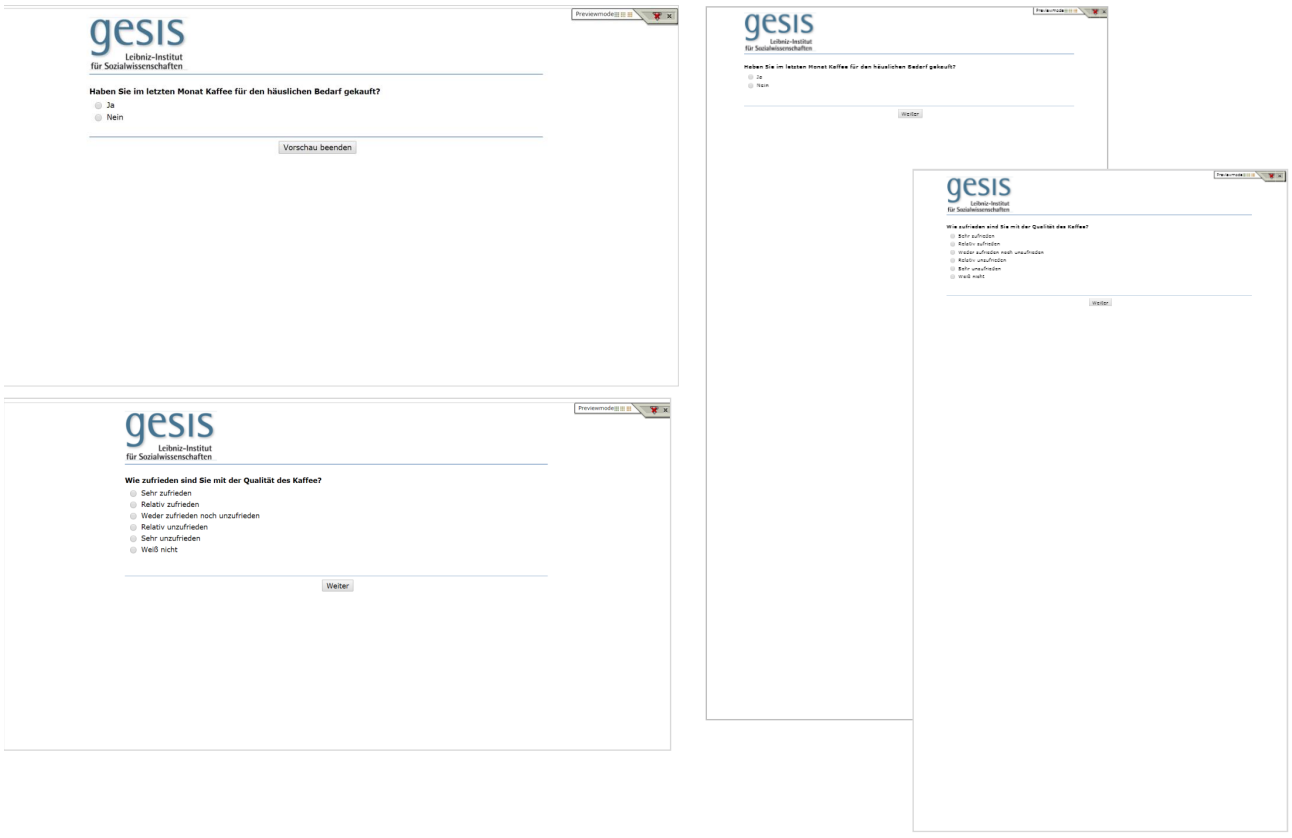


Figure 2.1: Display of filter and follow-up questions on devices

see no significant differences between respondents completing the survey in the two formats. These results reassure us that there were no substantial differences in drop out between the two formats. As shown in Column 2, however, there are systematic differences in the types of respondents who completed the survey on the two devices. It turns out that less educated, low-income, and rural respondents were hard to recruit for the mobile group. One reason for this may be that these groups are less likely to participate in online access panels anyway and were, in addition, not very experienced with smartphones (Kongaut and Bohlin 2016; Puspitasari and Ishii 2016).

To address the imbalance between the respondents completing the smartphone and PC versions of the survey, we applied entropy balance weighting. This approach derives weights to balance the observable characteristics of the PC and smartphone respondents (Hainmueller 2012; Hainmueller and Xu 2011) and has been used for similar purposes before (Eckman and Haas 2017).

Table 2.2: Randomization to format and device

	Format Interleafed (Ref. Grouped) Before Weighting	Format Interleafed (Ref. Grouped) After Weighting	Device Mobile (Ref. PC) Before Weighting	Device Mobile (Ref. PC) After Weighting
Sex	n.s.	n.s.	n.s.	n.s.
Net income	n.s.	n.s.	-.005***	n.s.
Education	n.s.	n.s.	-.012***	n.s.
Housing situation	n.s.	n.s.	n.s.	n.s.
Population size	n.s.	n.s.	-.025***	n.s.
Survey completion date	n.s.	n.s.	n.s.	n.s.
Age	n.s.	n.s.	n.s.	n.s.
Employment status	n.s.	n.s.	n.s.	n.s.
Federal state	n.s.	n.s.	n.s.	n.s.
Duration of the survey	n.s.	n.s.	n.s.	n.s.
Survey invitation date	n.s.	n.s.	n.s.	n.s.
n.s $\geq$ .05				

For example, in our sample more highly educated people have participated via smartphones (see Table 2.2); the entropy balancing method creates case-level weights that adjust the mean of the education variable of the PC respondents (control group) to match the smartphone respondents (treatment group). The method solves for the weights that simultaneously match all of the variables shown in Table 2.2.

After weighting with the entropy balance weights, no significant, observable differences remained between the smartphone and PC survey respondents (see Table 2.2, 4th column). However, we can only weight for observed characteristics and not for other unobserved characteristics that distinguish the two groups, such as a respondent's motivation to participate in the survey. Anyway, we used these weights in all analyses to remove the imbalances between the two devices and make the two groups of respondents comparable.

2.3.2 Data quality indicators

We built four indicators of poor data quality to test hypotheses 2, 4 and 6 referring to data quality in the follow-ups, summarized in Table 2.3. Following Antoun, Couper, and Conrad

(2017), we used heaping as the first indicator. This indicator was built from the follow-up question about how much the product costs. When the reported cost was divisible by ten, the indicator is 1 ("heaping"), and 0 ("no heaping") otherwise. Heaping is an indicator of poor data quality because it takes less cognitive effort to give an approximate price than to remember the exact one and furthermore it is easier to dial rounded values without decimals on the keyboard. For the question format effect, we expect respondents in the grouped format to tend more to heaping because they are surprised and might even be annoyed that each affirmative answer in the filter questions has triggered the follow-ups. For mobile respondents, we expect more heaping, as smartphones have a smaller keyboard and dialing different numbers is more difficult than on a PC. Furthermore, smartphone respondents might be more distracted by their environment and tend to just give an approximate rounded value rather than remember the exact value. In the first bar of figure 2.2, we see the average percent of filter questions with heaping and their standard deviation in percent. In 36% of the triggered price questions respondents provided heaped responses.

Table 2.3: Definition of data-quality indicators

Indicator	Definition	
Heaping	Definition:	Reported value is divisible by 10 / binary
	Other papers using:	Antoun, Couper, and Conrad (2017)
	Follow-Ups Used:	How much did it cost?
Categories not selected	Definition:	Number of categories (not) selected / metric 1-5
	Other papers using:	Lugtig and Toepoel (2016)
	Follow-Ups Used:	For whom was it purchased?
Middle category selected	Definition:	Middle category "neither nor" was selected / metric 1-5
	Other papers using:	Krosnick (1991)
	Follow-Ups Used:	How satisfied are you with the quality of the [product]?
Item nonresponse	Definition:	Item nonresponse or don't know
	Other papers using:	Lugtig and Toepoel (2016) Antoun, Couper, and Conrad (2017)
	Follow-Ups Used:	All

The second indicator was the number of categories not selected in multiple choice items similar as used in Lugtig and Toepoel (2016). Lugtig and Toepoel (2016) used the number of categories

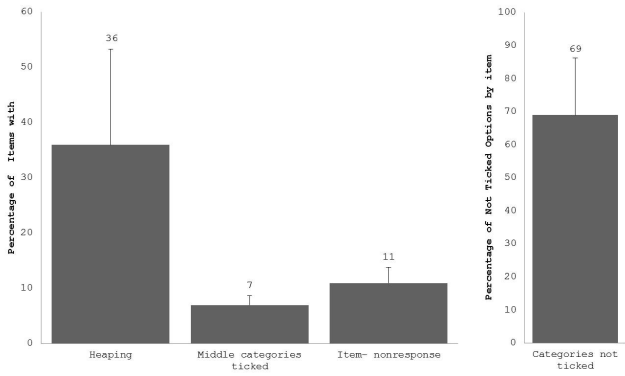


Figure 2.2: Mean percentage and standard deviations in percent of data quality indicators in follow-up questions

selected, however we use the somewhat unusual category items not selected to keep the direction of all of our data quality indicators in the same direction. For this indicator, we used the follow-up question for whom the product was purchased. It might be more difficult or burdensome for smartphone respondents to select more than one of the categories (for myself, another household member, someone else, don't know) on the small display and respondents in the grouped format, as explained in the last section, might be more annoyed by the follow-up questions. So they might answer the follow-up questions with minimal effort and therefore "quickly" select only one category. We therefore expected fewer selected categories for smartphone respondents. Across devices for the triggered filters, 69% of the categories were not selected. This resulted, on average, in 1.24 ticks out of 4 for each triggered filter. To keep the direction of our data quality indicators consistent, we named this indicator "categories not selected".

The third indicator followed studies such as Krosnick (1991) and referred to whether a respondents selected the middle category. For the same reasons as explained above, we expect more middle category responses for respondents in grouped format. Smartphone respondents might be exposed to more distractions and multitasking because the smartphone is always with them and so the survey environment was not fixed. This made it more difficult to concentrate and therefore less likely to be able to made an adequate decision. Bypassing the response decision process by selecting the middle category reduced the burden for respondents. In conclusion, we expected smartphone respondents to be more likely to select the middle category. The middle category was only selected in 7% of the items across both devices.

For respondents in the grouped format, we again expect, for the same reasons as outlined above, more item nonresponse than for respondents in the interleafed format. In addition, following Antoun, Couper, and Conrad (2017) and Lugtig and Toepoel (2016), we expected smartphone respondents to provide more item nonresponse for similar reasons as being more likely to choose the middle categories - distraction and multitasking might lead to lack of concentration. On average, respondents had 11% missing items in the follow-ups.

2.3.3 Analysis plan

In our analysis we will consecutively address our six hypotheses. Therefore, we calculate the mean values for the triggered filter questions separately by question format and device. We test these with t-test for significance. Then we test the interaction of question format and device as applied by Bach, Eckman, and Daikeler (2019) before. We apply this procedure, then for each of our four data quality indicators. We have chosen this procedure to obtain exact values for the filters triggered and compare data quality values. As an alternative, classic regressions are also used, the results of which can be found in the appendix section 2.6.4.

2.4 Results

In the following, we test our six hypotheses: respondents in the interleafed format trigger fewer filter questions (H1); the grouped format leads to lower data quality in the follow-ups than the interleafed format (H2); smartphone respondents trigger fewer filter questions than PC respondents (H3); smartphone respondents provide lower data-quality in the follow-up questions than PC respondents (H4); smartphone respondents in the interleafed filter question format trigger fewer filter questions than respondents in the grouped format or PC respondents (H5) and smartphone respondents in the grouped question format provide lowest data quality in the follow-ups compared to respondents in the interleafed format or PC respondents (H6).

2.4.1 Triggered filter questions and follow-up data quality by question format (H1) and (H2)

First of all, for our first hypotheses, we note that our results replicate the format effect reported in the literature (e.g. Kreuter et al. 2011; Bach, Eckman, and Daikeler 2019): On average, respondents in the grouped format give about one more affirmative answer to the 11 filter questions. Figure 2.3 illustrates this result, the markers show the average number of triggered filter questions in the two formats for smartphone and PC respondents. The slopes represent the interaction and non-overlapping confidence intervals indicate a significant difference. We cannot observe overlapping confidence intervals between the two formats, this indicates the difference between the two formats is statistically significant ($p < 0.001$; see also column 1 of Table 2.4) and respondents in the grouped format trigger more filter questions.

Hypothesis 2 states that data quality in the follow-up questions is lower in the grouped format than in the interleafed format. To test this hypothesis, we developed four indicators of data quality in the follow-up questions (see Section 2.3). Table 2.5 provides an overview of the regression results on the question format and device effects for each of the four data quality indicators. The first part of the table refers to the format effect. The second column ("Interleafed (ref. Grouped)") tests if respondents in the *interleafed* format provide worse data quality. In columns 3 and 4, we investigate this effect separated by PC (column 3) and mobile (column 4) respondents. The lower part of the table investigates in column 2 the device effect and examines this effect then separately for the two formats in columns 3 and 4.

The results indicate (upper part of table 2.5 & 1st row) better data quality in the interleafed format for two of the four data quality indicators. Better data quality, in our case, means that the indicators are significantly *lower* in the interleafed format (recall that our four data quality variables, defined in Table 2.3, are each indicators of *poor* data quality). Respondents in the interleafed format are less likely to use the middle category and and to provide item nonresponse (column 1, rows 3 & 4). The results of the corresponding full regression model can be found in the appendix section 2.6.4.

Table 2.4: Mean number of triggered filter questions by question format and device

Format	Overall	PC	Mobile	N	T-Statistic (p)
Overall	5.2	5.1	5.3	3158	-1.9(.19)
Interleafed	4.7	4.6	4.7	1576	-.5 (.60)
Grouped	5.7	5.7	5.8	1582	-1.4 (.17)
N	3158	1868	1290	3158	
T-value (p)	12.2(¡.001)	10.7 (¡.001)	9.7 (¡.001)		

2.4.2 Triggered filter questions and follow-up data quality by device (H3) and (H4)

Contrary to our expectations in the third hypothesis, we do not find evidence that smartphone respondents engage in more motivated underreporting when responding to filter questions. Comparisons between the devices (rows two and three of Table 2.4) show that there is no difference in triggered filter between smartphone and PC respondents. Regardless of the device, respondents in the grouped format trigger more filter questions, on average, than those who responded in the interleafed format.

In line with our assumptions we find lower data quality for two out of four data quality indicators in the follow-ups for smartphone respondents compared to PC respondents. Each panel of figure 2.4 corresponds to one of the four data quality indicators. In each panel, the markers show the average data quality indicator in the two formats for smartphone and PC respondents. The lines represent format effects, just as they did in figure 2.3. In the top left panel, corresponding to the heaping data quality measure, we see significant differences between PC and smartphone respondents in the grouped format, but not in the interleafed format (confidence interval does overlap). Smartphone respondents show lower data quality (more heaping) than PC respondents in the grouped format. In the top right panel (numbers of not selected categories), there are no significant differences in either format with respect to data quality. In the two bottom panels, number of selected middle categories and item nonresponse, we see higher values (lower data quality) for smartphone respondents in the grouped format. However, this effect is only significant for item nonresponse in the grouped format.

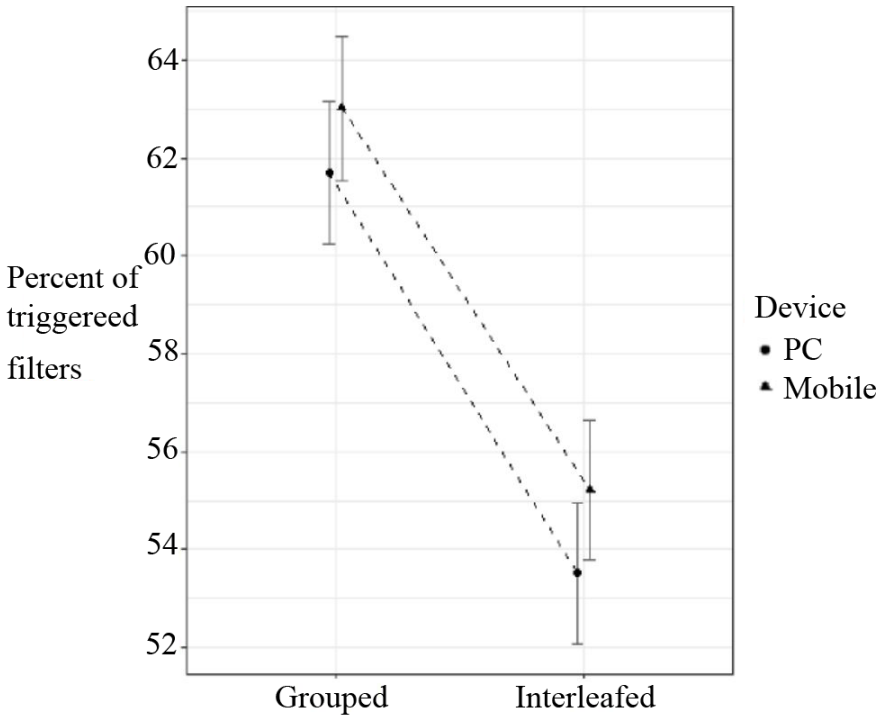


Figure 2.3: Triggered filter questions by format and device (in %)

2.4.3 Triggered filter questions and follow-up data quality in an interaction of device and format (H5) and (H6)

The 5th hypothesis states an interaction between format and device in the filter questions. Figure 2.3 illustrates the interaction between triggered filter questions by format and device as outlined. Similar as in Table 2.4 the y-axis shows the percent of triggered filter questions (instead of the number of filters triggered). In the grouped format, PC respondents trigger almost 62% of the filters and smartphone respondents one percentage point more. In the interleafed format, smartphone respondents also trigger on average one percent more filters however, this difference is, judging by the confidence intervals, not significant. Since the slopes of both effects are almost perfectly parallel, there is no interaction between question format and device (compare also appendix section 2.6.4).

Our last hypothesis states an interaction effect between format and device for the quality of responses to the follow-ups. We investigate hypothesis 6 in figure 2.4. The dashed lines between the two sets of point estimates represent the format effect: the difference between the grouped and interleafed formats. The dashed lines are parallel, which we interpret as evidence that the

Table 2.5: Indicators of data quality in follow-ups, by device and format

	Data quality by format		
Indicator	Interleafed (ref. Grouped)	Interleafed (ref. Grouped) PC	Smartphone
Heaping	no effect	no effect	no effect
Categories not selected	no effect	no effect	no effect
Middle category selected	better*	no effect	better*
Item nonresponse	better*	no effect	better*

	Data quality by device		
Indicator	Smartphone (ref. PC)	Smartphone (ref: PC) Interleafed	Grouped
Heaping	worse*	no effect	worse *
Categories not selected	no effect	no effect	no effect
Middle category selected	no effect	no effect	worse*
Item nonresponse	worse*	no effect	worse*

* $p \leq 0.05$

format effect is the same on the two devices. If the lines were not parallel, we would conclude that the format effect worked differently on the two devices. We tested the full interaction – whether the format effect is significantly different for PC and smartphone respondents (that is, whether slopes of the lines in each panel for each of the four data quality indicators are different). We test this by comparing the slopes by device via a t-test. None of these results are significant (see also appendix table 2.7 in section 2.6.4 for the full regression results).

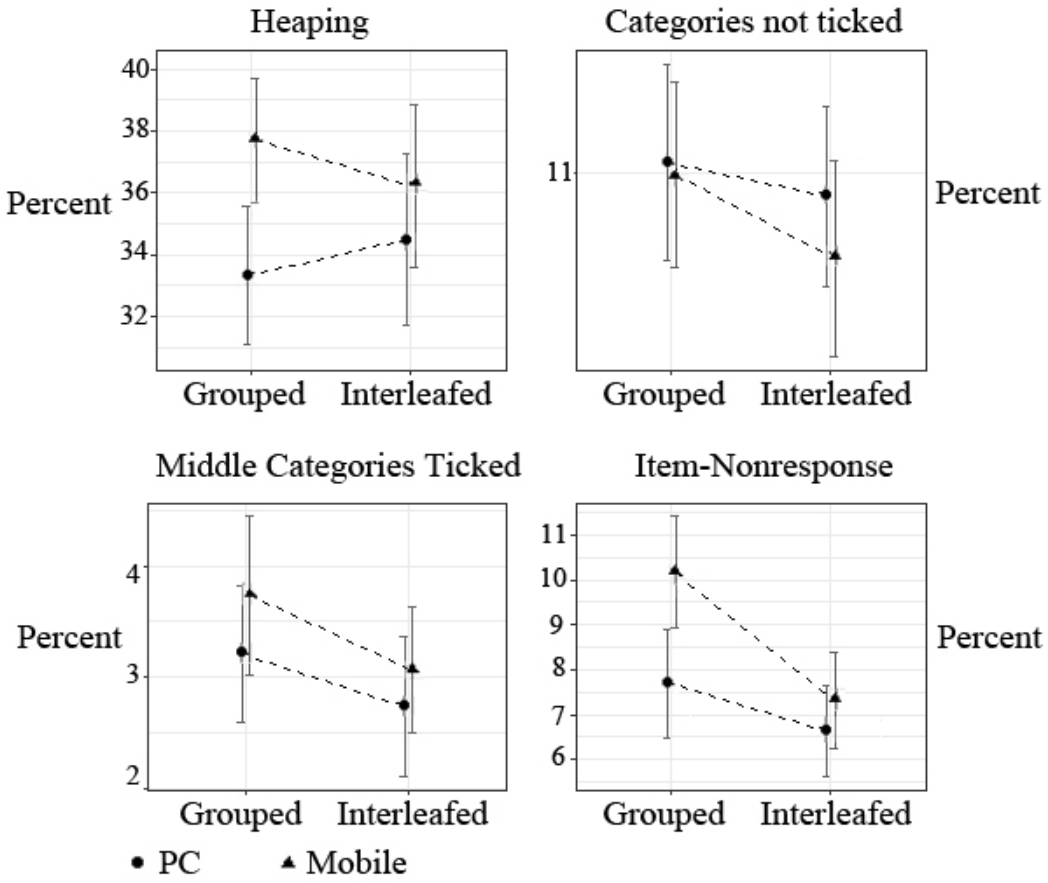


Figure 2.4: Data quality in follow-up questions

2.5 Discussion

This study randomly assigned web survey respondents to two experimental conditions: filter format (interleafed and grouped) and device (PC and smartphone). The data allow us to test six hypotheses about the performance of filter questions by format and device. Our results let us replicate the format effect that is by now well known: respondents in the grouped format trigger

more follow-up questions than those in the interleafed format (H1). However, we did not find support for our third hypothesis that the format effect would be stronger among smartphone respondents.

The second and fourth hypotheses relate to data quality in the follow-up questions rather than responses to the filter questions themselves. The results supported hypothesis 2: on two of our four measures of data quality, the grouped format produced lower data quality to the follow-ups than the interleafed format. This results suggests that the grouped format has two somewhat contradictory effects on data quality: it collects more YES responses to the filter questions but lower data quality to the follow-ups. Thus, the net effect of the filter question format on data quality is more complex than that suggested by previous studies. However, these results in fact hold only for respondents on smartphones (hypothesis 4). Among PC respondents, data quality in the follow-ups did not differ for interleafed and grouped respondents (hypothesis 6).

The study encountered some difficulties in compliance with the device assignment, which we attempted to fix using entropy balance weighting. This approach uses weights that balance the treatment and control groups (here assigned-to-mobile and assigned-to-PC). However, it is possible that there are other differences between the groups that we have not controlled for, they could bias our results. Explicitly, there might be differences in the motivation of the respondent, which influence the self-selection effect into the two devices. If the mobile respondents in particular are more motivated and therefore cooperate to participate (also) on the smartphone, this could explain the null effects on the number of filter questions triggered. Clearly, more evidence is needed on the issue of device effects when answering filter questions and follow-up questions. Specifically, true random assignment to device is difficult, because respondents always have the option not to participate in the survey if they do not like the mode and device they are assigned.

Despite this shortcoming, the results presented above should concern all researchers using filter questions, especially in web surveys. There is mounting evidence that the format in which filters and follow-ups are asked effects responses in various question types. Researchers should think carefully about whether the responses to the filters or the follow-ups are most important

in their research. The interleafed format allows to collect higher quality data with respect to the filters themselves (Eckman et al. 2014), but the grouped format allows to collect higher quality data in the follow-ups. Eckman and Kreuter (2018) argue that the grouped format may be preferable, because the missing data in the follow-ups is more visable to analysts, and imputation can be used to fill in missing values. However, this study shows that the harm to data quality in the grouped format does not always take the form of missing data. When respondents give a response to a follow-up item, and that response is not correct, the problem is not clear to analysts, and it can not as easily be fixed through imputation. Furthermore, this study shows for mixed device studies that mobile respondents do not provide lower data quality in the filter questions but in the follow-up questions and that this effect applies to both question formats. Since this effect occurs particularly with heaping and item nonresponse, we recommend to further optimize the survey design of the follow-up questions for mobile surveys (e.g. by automatically adjusting the font size or the usage of voice recordings) as well as to implement as few open questions as possible to provide exact figures. We also recommend to replicate this study with a population representative sample and in particular to control the motivation of the respondents, or to use a laboratory experiment in which the threat of self-selection is reduced compared to a field experiment.

2.6 Appendix

2.6.1 Survey invitation PC and smartphone

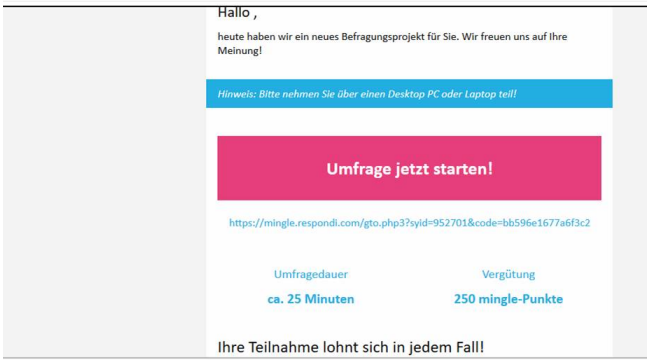


Figure 2.5: Survey invitation PC

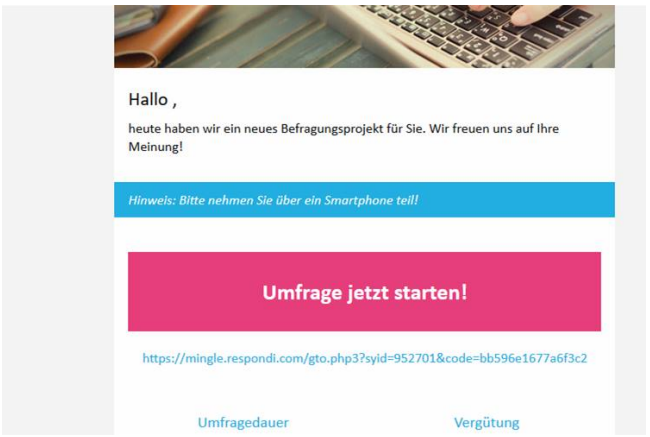


Figure 2.6: Survey invitation smartphone

2.6.2 Text of filter and follow-up questions

- In the past month, have you purchased coffee for consumption at home?
- In the past month, have you purchased beer or wine for consumption at home?
- In the past month, have you purchased tobacco?
- In the past month, have you purchased children's clothing or shoes?
- In the past month, have you purchased clothing or shoes for yourself?

- In the past month, have you purchased chocolate?
- In the past month, have you purchased medication?
- In the past month, have you purchased flowers?
- In the past month, have you purchased pet supplies?
- In the past month, have you purchased movies on DVD or VHS?
- In the past month, have you purchased music on CD or as MP3s (or other digital formats)?
- In the past month, have you purchased a ticket for a concert, theater performance or a movie?
- In the past month, have you purchased any cleaning supplies for your home?

follow-up questions

For each affirmative answer to the above filter questions:

Thinking about your most recent purchase of (fill: item)

- How satisfied are you with the quality of the (fill: item)?
  - a. very satisfied
  - b. somewhat satisfied
  - c. neither nor
  - d. partly Satisfied
  - e. not at all satisfied
- How much did it cost?
  - (Open ended response in Euros)
  - a. Don't know

  - b. Refused

- For whom was it purchased?
  - a. self
  - b. another household member
  - c. someone else
  - d. Don't know
  - e. Refused

2.6.3 Data quality indicators

Table 2.6: Summary statistics for data-quality indicators

Indicator	Mean (in %)	sd	Median	Min	Max
Heaping	36	0.48	0	0	1
Categories not selected	69	0.87	75 (3 cat.)	25 (1 cat.)	100 (4 cat.)
Middle category selected	7	0.25	0	0	1
Item nonresponse	11	0.26	0	0	1

2.6.4 Format, filter and interaction effects

Table 2.7: Regression outcomes

	Triggered Filter Questions	***Heaping***	***Item non-response***	***Categories not ticked***	***Usage of Middle Category***
	Poisson	*Logistic*	*Logistic*	*Poisson*	*Logistic*
	b (SE)	**b (SE)**	**b (SE)**	**b (SE)**	**b (SE)**
Interleafed Format	-.0874***	.05253	.1637*	.2394	.03913*
	(.01067)	(.08598)	(.1231)	(.4826)	(.1605)
Smartphone	.01685	.1914*	.2981**	-.08411	.1175
	(.01069)	(.07941)	(.1149)	(.4879)	(.1449)
Interleafed* Smartphone	.00262	-.1162	-.1927	-.354	-.0458
	(.01498)	(.1163)	(.1631)	(.6871)	(.2153)
constant	.600***	-.6938***	-2.482***	68.83***	-2.66***
	(.00762)	(.05949)	(.08965)	(.3534)	(.1026)
(Pseudo) R$\hat{2}$	0.0051	0.001	0.0042	0.0005	0.0003
N	22363	11135	24396	11135	11131

References

Antoun, Christopher, Mick P Couper, and Frederick G Conrad (2017). "Effects of mobile versus PC web on survey response quality: A crossover experiment in a probability web panel". In: *Public Opinion Quarterly* 81.S1, pp. 280–306.

Bach, Ruben L and Stephanie Eckman (2018). "Motivated misreporting in Web panels". In: *Journal of Survey Statistics and Methodology* 6.3, pp. 418–430.

Bach, Ruben, Stephanie Eckman, and Jessica Daikeler (2019). "Misreporting among reluctant respondents". In: *Journal of Survey Statistics and Methodology* forthcoming.

De Bruijne, M. and A. Wijnant (2013). "Comparing survey results obtained via mobile devices and computers: An experiment with a mobile web survey on a heterogeneous group of mobile devices versus a computerassisted web survey". In: *Social Science Computer Review* 31.4, pp. 482–504.

Duan, Naihua, Margarita Alegria, Glorisa Canino, Thomas McGuire, and David Takeuchi (2007). "Survey conditioning in self-reported mental health service use: Randomized comparison of alternative instrument formats". In: *Health Research and Educational Trust* 42.2, pp. 890–907.

Eckman, Stephanie and Georg-Christoph Haas (2017). "Does granting linkage consent in the beginning of the questionnaire affect data quality?" In: *Journal of Survey Statistics and Methodology* 5.4, pp. 535–551.

Eckman, Stephanie and Frauke Kreuter (2018). "Misreporting to looping questions in surveys: Recall, motivation and burden". In: *Survey Research Methods* 12.1, pp. 59–74.

Eckman, Stephanie, Frauke Kreuter, Antje Kirchner, Annette Jäckle, Roger Tourangeau, and Stanley Presser (2014). "Assessing the mechanisms of misreporting to filter questions in surveys". In: *Public Opinion Quarterly* 78.3, pp. 721–733.

Gummer, T. and J. Rossmann (2015). "Explaining interview duration in web surveys: A multilevel approach". In: *Social Science Computer Review* 33.2, pp. 217–234.

Hainmueller, Jens (2012). "Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies". In: *Political Analysis* 20.1, pp. 25–46.

Hainmueller, Jens and Yiqing Xu (2011). "Ebalance: A Stata package for entropy balancing". In: *Journal of Statistical*.

Kessler, Ronald C., Hans-Ulrich Wittchen, Jamie M. Abelson, Katherine McGonagle, Norbert Schwarz, Kenneth S. Kendler, Bärbel Knäuper, and Shanyang Zhao (1998). "Methodological studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey (NCS)". In: *International Journal of Methods in Psychiatric Research* 7.1, pp. 33–55.

Keusch, Florian and Ting Yan (2017). "Web versus mobile web: An experimental study of device effects and self-selection effects". In: *Social Science Computer Review* 35.6, pp. 751–769.

Kongaut, Chatchai and Erik Bohlin (2016). "Investigating mobile broadband adoption and usage: A case of smartphones in Sweden". In: *Telematics and Informatics* 33.3, pp. 742–752.

Kosyakova, Yuliya, Jan Skopek, and Stephanie Eckman (2014). "Do interviewers manipulate responses to filter questions? Evidence from a multilevel approach". In: *International journal of public opinion research* 27.3, pp. 417–431.

Kreuter, Frauke, Stephanie Eckman, and Roger Tourangeau (2019). "Salience of survey burden and its effects on response behavior to skip questions. Experimental results from telephone and web-surveys". In: *Advances in Questionnaire Design, Development, Evaluation and Testing*. Ed. by P Beatty, D Collins, L Kaye, J Padilla, G Willis, and A. Wilmot. Hoboken: Wiley.

Kreuter, Frauke, Susan McCulloch, Stanley Presser, and Roger Tourangeau (2011). "The effects of asking filter questions in interleafed versus grouped format". In: *Sociological Methods & Research* 40.1, pp. 88–104.

Krosnick, Jon A. (1991). "Response strategies for coping with the cognitive demands of attitude measures in surveys". In: *Applied Cognitive Psychology* 5.3, pp. 213–236. DOI: 10.1002/

acp.2350050305. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2350050305.

Lavrakas, Paul J (2008). *Encyclopedia of survey research methods*. Sage Publications.

Lugtig, Peter and Vera Toepoel (2016). "The use of PCs, smartphones, and tablets in a probability-based panel survey: Effects on survey measurement error". In: *Social Science Computer Review* 34.1, pp. 78–94.

Puspitasari, Lia and Kenichi Ishii (2016). "Digital divides and mobile Internet in Indonesia: Impact of smartphones". In: *Telematics and Informatics* 33.2, pp. 472–483.

Schlosser, S. and A. Mays (2018). "Mobile and dirty: Does using mobile devices affect the data quality and the response process of online surveys?" In: *Social Science Computer Review* 35.3, pp. 212–230.

Tourangeau, Roger, Hanyu Sun, Ting Yan, Aaron Maitland, Gonzalo Rivero, and Douglas Williams (2018). "Web surveys by smartphones and tablets: Effects on data quality". In: *Social Science Computer Review* 36.5, pp. 542–556. DOI: 10.1177/0894439317719438.

Chapter 3

Web Versus Other Survey Modes

An Updated and Extended Meta-Analysis Comparing Response Rates

3.1 Abstract

In press for: Journal of Survey Statistics and Methodology

Do web surveys still yield lower response rates compared to other survey modes? To answer this question, we replicated and extended a meta-analysis by Lozar Manfreda et al. (2008), who found that, based on 45 experimental comparisons, web surveys had an 11 percentage points lower response rate compared to other survey modes. Fundamental changes in Internet accessibility and use since the publication of the original meta-analysis would suggest that people's propensity to participate in web surveys has changed considerably in the meantime. However, in our replication and extension study, which comprised 114 experimental comparisons between web and other survey modes, we found almost no change: Web surveys still yielded lower response rates than other modes (a 12 percentage points difference in response rates). Furthermore, we found that prenotifications, the sample recruitment strategy, the survey's so-

licitation mode, the type of target population, the number of contact attempts, and the country in which the survey was conducted moderated the magnitude of the response rate differences. These findings have substantial implications for web survey methodology and operations.

3.2 Introduction

The use of online surveys is on the rise; in 2007, for the first time, online surveys constituted the majority of all quantitative survey modes implemented worldwide. According to ESOMAR's latest Global Market Research Report (ESOMAR 2018, p. 139), web survey use has more than doubled compared to 2007. Underlying this widespread growth is the transformation of the web surveys from an initially novel to a well-established mode of survey implementation. The broad discussion on online data quality has pointed out, on the one hand, positive data quality aspects of the web mode, for example, an increased level of reporting of sensitive information (Kreuter, Presser, and Tourangeau 2008; Sakshaug, Yan, and Tourangeau 2010) and time-sensitive aspects (Chang and Krosnick 2009). On the other hand, it has also revealed several shortcomings of web surveys such as question skipping, speeding, response inconsistency, and satisficing (Heerwegh and Loosveldt 2008; Kim et al. 2018), as well as representativeness issues (Cornesse and Bosnjak 2018). Web surveys are especially useful when surveying specific populations with high Internet coverage such as students, customers, and employees with email addresses (Cernat, Couper, and Ofstedal 2016; Patrick et al. 2017). For these populations, the coverage bias problem is usually low. For the general population, however, Internet users and non-Internet users are not randomly distributed (Chang and Krosnick 2009; Blom et al. 2017) and this thus presents a challenge to many online surveys. Although the quality aspects of web surveys that deserve further attention are numerous, the present study limits the discussion to response rates as an indicator of nonresponse error.

3.3 Background

Experimental studies comparing the response rates of web surveys with those of other survey modes have reported higher response rates for traditional survey modes (Fricker et al. 2005; Kirchner and Felderer 2016). By contrast, a substantial body of literature has emphasized the advantages of web survey over traditional modes (Greene, Speizer, and Wiitala 2008; Boyle et al. 2016). Whereas these are individual experimental studies, several systematic reviews of response rate comparisons have also been conducted. For instance, Shih and Fan (2008) carried out a meta-analysis comparing only the response rates of web surveys and mail surveys and found, on average, that mail surveys had higher response rates than web surveys. However, the most comprehensive research synthesis to date on the response rate difference between web and other survey modes was conducted by Lozar Manfreda et al. (2008). On average, the authors found an 11 percentage points lower response rate for web surveys than for other survey modes. Moreover, and even more importantly, they examined the study characteristics, also known as moderators, to determine which ones significantly influence this response rate difference. Their results revealed the following moderators of this difference: the sample recruitment base (a smaller response rate difference between web and other survey modes in the case of panel members as compared to one-time respondents); the solicitation or invitation mode chosen for web surveys (a higher response rate difference for postal mail solicitation compared to email solicitation); and the number of times contact is made with respondents (the more contacts made, the larger the response rate difference between modes).

We designed the present study as a replication and extension of Lozar Manfreda et al. (2008) previous research for two main reasons. First, we wanted to identify the benefits and limitations for web response rates compared to other survey modes; second, we wanted to determine whether Lozar Manfreda et al. (2008) findings are still applicable today. Several years have gone by since Lozar Manfreda et al. (2008) finalized their literature search in 2005, and during this time the web survey field has faced many changes. Some of the limitations of web surveys have multiplied. First, there is greater sensitivity with respect to data security nowadays (Callegaro, Lozar Manfreda, and Vehovar 2015, pp. 125); second, there has been an increase in

the diversity of Internet browsers, (mobile) devices, and operating systems, which has caused problems of technical incompatibility (Couper and Peterson 2017); third, there has been an increase in online over-surveying and spam emailing (e.g., Callegaro, Lozar Manfreda, and Vehovar 2015, p. 171); and fourth, there might be a lower legitimacy of researchers who may carry out impersonal and quick web surveys (e.g. Callegaro, Lozar Manfreda, and Vehovar 2015; Groves et al. 2011, p.171,149). On the other hand, new opportunities for web surveys have been developed due to (1) the increased web literacy of web respondents, which reduces technical limitations (Eshet-Alkalai and Chajut 2010); (2) higher Internet coverage rates (e.g., World Bank 2017); (3) the availability of a variety of increasingly user-friendly devices with which to access the Internet (e.g., touchscreens, Wi-Fi connections) (e.g., Al-Razgan et al. 2012); (4) changes in Internet access payments (from pay-per-minute to flat rates) (e.g., Aichele et al. 2006); (5) the fact that contacting people via other modes of communication has become more difficult due, for example, to the increasing number of households without landline telephones (e.g., Dillman, Smyth, and Christian 2014, p. 10). Our second research objective – to determine whether Lozar Manfreda et al.'s (2008) findings are still applicable – is prompted by Shojania et al. (2007), who addressed in their research the question of how quickly systematic reviews go out of date and demonstrated that the median survival time was only 5.5 years. As a consequence, they recommended the regular updating of systematic reviews. Accordingly, this study aims to answer the following research questions (RQs):

RQ 1: Do web surveys yield lower, higher, or the same response rates as other survey modes?

To answer this question, we update the meta-analysis performed by Lozar Manfreda et al. (2008) with respect to possible changes over time. In addition, we aim to explore whether new studies have increased the explanatory power of the variables presumed to explain the variability of the response rate differences between web and other surveys modes, and to determine whether any other moderators also have an impact. In the original meta-analysis performed by Lozar Manfreda et al. (2008), certain survey characteristics such as the number of contact attempts had an influence on the response rate differences between web and other survey modes. Therefore, the response rate differences were heterogeneous and moderator explanation was reasonable.

In our replication and extension of this study, we explore whether and to what extent the mean response rate difference varies and what moderator variables explain this variation. Hence, our second research question asks:

RQ 2: Is the mean response rate difference heterogeneous?

The success of a survey, and thus the response rate, depends strongly on the survey settings and characteristics (Groves and Peytcheva 2008). We expect deviating effects, depending on the modes to which web surveys were compared (e.g., mail, telephone, face-to-face, interactive voice response (IVR), touch tone). A paper-based questionnaire usually remains within reach of the respondent for a period of time and can therefore act as a reminder (Dillman, Smyth, and Christian 2014, p. 382)). In telephone surveys, the time of day that the call is placed plays a crucial role in whether the potential respondent is busy or not available to take the call at all (Tourangeau et al. 2017). Email invitations and reminders for web questionnaires are more likely to be (un)intentionally overlooked (Petrovčič, Petrič, and Manfreda 2016). Incentives in online surveys can be confused with advertising and not taken seriously, especially if the survey sponsor is not a university or governmental organization. Additional effort must be made by researchers using modes other than the web for their surveys—for instance, mailing letters, making telephone calls, or even paying the respondent a personal visit. These additional efforts, if appreciated by respondents, may account for some of the greater legitimacy of these surveys compared to self-administered web surveys, and may therefore lead to higher response rates compared to email invitations or web questionnaires (Millar et al. 2011). Participation might also be higher if it is requested personally (via telephone or face-to-face), as potential respondents might find such personal requests harder to disregard. Moreover, compared to immediately answering survey questions on the phone, respondents have to be much more active when answering a web survey, especially if no email invitation is provided, (Fricker et al. 2005; Greenlaw and Brown-Welty 2009). Nevertheless, surveys for specific target populations with certain characteristics (e.g., higher Internet penetration, engagement with the survey topic) might work better online than surveys for the general population. Furthermore, with the success of the Internet, the attitude of the population toward web surveys has changed over time. This change, reflected in the response rates, is why we expect an effect on the year of

publication. Therefore, our research addresses whether study design or study circumstances have an effect on the response rate difference. These deliberations lead to our third research question:

RQ3: Do the sample recruitment base, solicitation mode, number of contacts, compared mode, type of target population, type of sponsorship, use of incentives, and the year the studies were published impact the variation in the response rate difference?

Whereas our first and second aims in the present study are to update and to increase the statistical power of Lozar Manfreda et al. (2008) meta-analytical findings, which addressed the moderators listed in RQ3, our third aim is to extend their meta-analysis. Therefore, we consider three additional moderators: survey topic, prenotification (i.e., an advance contact with respondents to announce the survey), and survey country. These additions are possible due to the larger number of primary studies included. With regard to the survey topic, it can be assumed that some types of survey topics work better than others on the web. Specifically, web respondents are more likely to provide answers to sensitive questions (Kreuter, Presser, and Tourangeau 2008). In addition, experimental evidence suggests that providing respondents with prenotifications has a consistently positive effect on response rates (Fan and Yan 2010). Here, we seek to determine whether prenotifications exert differential effects on the response rates of web surveys versus other survey modes. Receiving an email request to participate in a web survey may seem less legitimate to respondents, as sending such a request entails less effort on the part of researchers compared to requests via other channels, for example telephone or postal mail. Legitimacy is further undermined by the high number of web surveys currently being conducted and the low level of trust in the online world (Dillman, Smyth, and Christian 2014, p.450). Therefore, we assume that the use of prenotifications in web surveys is less advantageous than in other survey modes, and we postulate that prenotification should increase the response rate difference. Interestingly, and to the best of our knowledge, no meta-analyses on response rates have included cross-national factors. This lack is all the more surprising because country specificities and cultural factors—for example, a country's Internet coverage, mode-specific survey-taking climate, over-surveying, and openness to new

technologies—play a role in the acceptance and conducting of web surveys (Lyberg and Dean 1992; Couper, De Leeuw, et al. 2003). Thus, we hypothesize that a variation in response rate differences between web and other survey modes exists across countries. These deliberations give rise to our fourth research question:

RQ4: Is the response rate difference influenced by (1) the use of prenotifications, (2) the survey topic, and (3) the country in which the survey is conducted?

Because we want to isolate the impact of the survey mode from other causes, we include in our meta-analysis only primary studies with experiments that compare web response rates to the response rates of other survey modes. The next section describes our research method. This is followed by the results section, in which we address the mean difference in response rates in web surveys versus other survey modes and the robustness of this difference, as well as providing an analysis of the moderators. The paper concludes with a discussion of our findings and the limitations of our study.

3.4 Method

To ensure a proper replication of the original study, response rate differences between web surveys and other survey modes were examined using meta-analytic techniques that closely followed those used in Lozar Manfreda et al. (2008). The present section briefly describes the meta-analytic methods, the eligibility criteria and search strategy, the coding of primary studies, and the statistical procedures.

Our systematic review and meta-analysis comprised four steps. First, we conducted a comprehensive literature search using specific search terms derived from a set of study eligibility criteria. Second, we reviewed the manuscripts identified by this literature search and screened out those that did not comply with our eligibility criteria. In the third step, we coded pertinent data in order to compute response rates, and we used the information on potential moderators to calculate effect sizes and perform the moderator analyses. In the final step, we carried out

the meta-analytic statistical analyses. These four steps are explained in detail in the following sections.

3.4.1 Eligibility criteria and search strategy

For our meta-analysis, we employed the same eligibility criteria as those used by Lozar Manfreda et al. (2008), as close adherence to these criteria was an important precondition for mapping possible changes over time. Eligible studies had to meet the following criteria: (1) One of the survey modes used had to be a web-based survey (i.e., a survey in which a web questionnaire was used to gather responses from respondents online using various devices. (2) The web-based survey had to be compared to data from one or more other survey modes (e.g., email, mail, telephone, face-to-face, telefax). (3) Data had to be available on response rates of the web and other survey mode(s). (4) A split-sample experimental design had to have been employed with subjects from the same population who were randomly assigned to different modes. In other words, the eligible studies included a study design in which each respondent was randomly assigned to either the web mode or the compared mode. (5) Subjects had to remain in the mode to which they were randomly assigned; in other words, studies in which subjects were permitted to switch modes were not eligible for inclusion. (6) The implementation of the compared survey modes had to be identical, with the only difference being the mode used to answer the survey questionnaire. Hence, for example, comparisons of surveys that used unequal incentives were excluded.

There is only one difference between the present criteria and those used in Lozar Manfreda et al.'s (2008) meta-analysis. In the original meta-analysis, primary studies that had the same number of contact attempts (regardless of the type of contact) were considered to be identical and were thus included in the meta-analysis. By contrast, we excluded experimental comparisons in which only one survey mode used prenotification (although the overall number of contact attempts might have been the same). This was not an option for the original meta-analysis because the number of studies was much smaller, and taking this approach would have led to a loss of statistical power. In addition, having a larger number of studies at our disposal

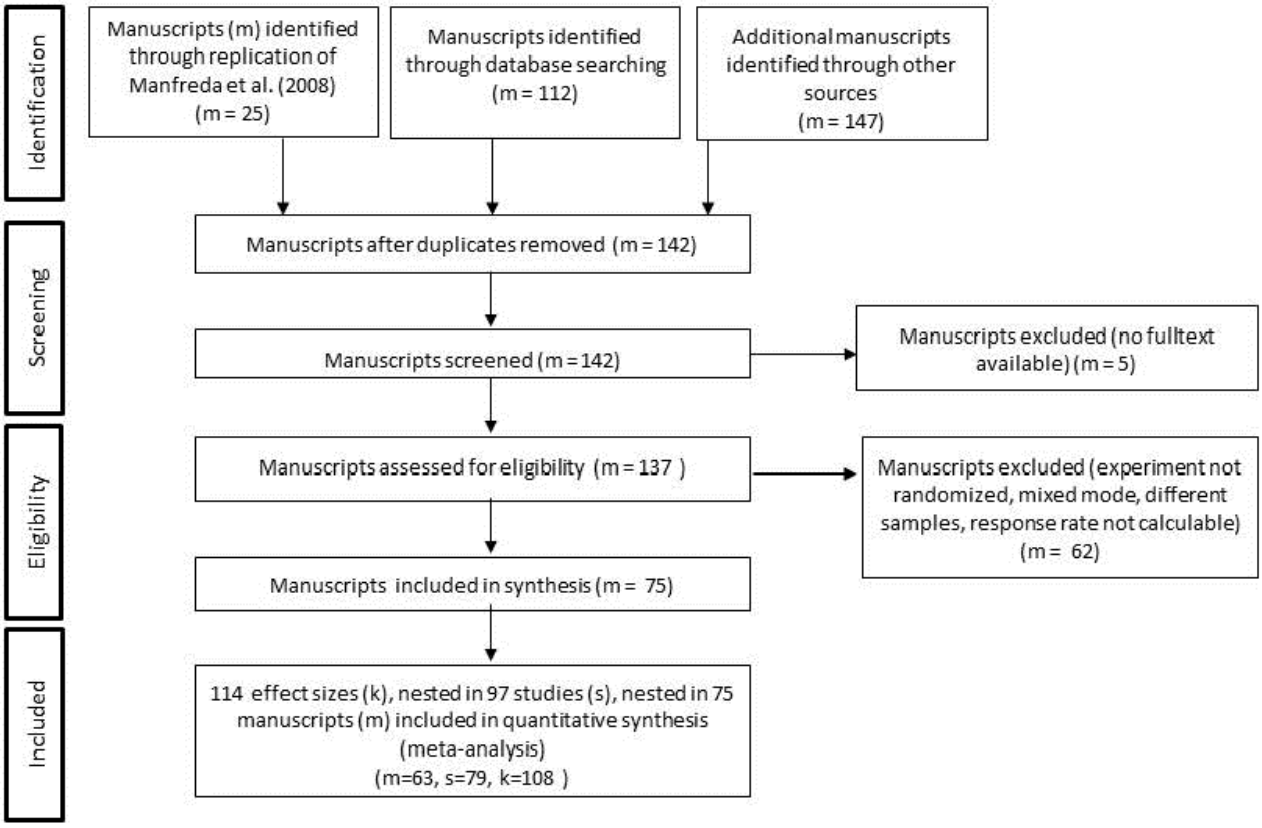
allowed us to examine prenotification as a separate moderator. Consequently, we excluded seven effect sizes (Kaplowitz, Hadlock, and Levine 2001; Miller et al. 2002; Cole 2005), and in this respect, our study is not an exact replication of Lozar Manfreda et al. (2008) meta-analysis. Given the small number of excluded studies, we still consider this a valid approach to determining change over time. In addition, like Lozar Manfreda et al. (2008), we imposed no participant population, time period, or geographical restrictions.

As a first important step to ensure the quality of our meta-analysis, we performed a comprehensive literature search, applying the same search terms as those used in Lozar Manfreda et al.'s (2008) study (see appendix Table 3.5). To overcome the publication bias (Rosenthal 1979) problem, we employed several techniques. With the aid of a snowballing technique, we inspected the reference lists of the selected publications. However, to explicitly collect grey literature, we examined conference abstracts (see appendix table 3.6) from the years 2005 to 2016. The PRISMA flow diagram (Moher et al. 2009) in figure 3.1 provides an overview of our search strategy, which was restricted to the literature in English. Finally, we included over 100 effect sizes in our meta-analysis (indicated by a * in reference section).

3.4.2 Coding procedures

Coding was performed by two independent coders using the coding sheet (see appendix table 2.3). The solicitation mode used in the web mode was the only moderator coded for the web mode; all other moderators are applicable to both modes. The second coder was instructed by the first; coding samples were provided. The second coder coded a random sample of one-third of the manuscripts, and the intercoder reliability showed a Krippendorff's alpha (Krippendorff 2004) of .92, indicating almost a 92 percent agreement between the two coders. As Krippendorff (2004) recommended an alpha value of .80 or higher, this is an excellent value.

Figure 3.1: PRISMA literature search flow diagram



Notes: m-manuscripts, s- studies, k-effect sizes, adapted from Moher et al. (2009)

3.4.3 Statistical method

In line with the original meta-analysis by Lozar Manfreda et al. (2008), we calculated the response rate difference, which is our effect size, using raw frequency. Accordingly, we used the number of invited and eligible subjects compared to the number of actual respondents per mode. In most of the included studies, the effective initial sample size was calculated as the initial sample size minus undeliverable and non-eligible units. However, raw frequencies are essential for calculating the confidence interval for each effect size. In those cases with insufficient data, we used the authors' definition of the response rate and calculated the raw frequencies. As the authors used the same response rate calculation logic and our effect size was the response rate difference between the two modes rather than the raw response rate, using the authors' definition of response rate and respondent was found to be adequate. In addition, although different survey projects may use different definitions of usable respondents (e.g., those who answered 90 percent of the items, 50 percent of the items, etc.), we relied on the authors' definition of usable respondents, and assumed that they used the same criteria for both modes under comparison. As we were interested only in differences, we found this strategy appropriate. We built a dummy variable based on whether the authors provided the raw frequencies or the response rates only. It showed no significant effect in moderating the average response rate difference.

Our effect size is the response rate difference (RD) between the web mode and the compared mode, which was calculated as follows:

$$RD = \frac{N\ respondents\ web\ mode}{N\ invited\ and\ eligible\ subjects\ web\ mode} - \frac{N\ respondents\ other\ mode}{N\ invited\ and\ eligible\ subjects\ other\ mode}$$

Thus, a positive RD indicates a higher response rate for the web mode, and a negative RD indicates a lower response rate for the web mode compared to the other survey mode. In general, our statistical analysis comprised five steps (Lipsey and Wilson 2001). First, we computed the weighted mean response rate difference across all studies by weighting each effect size by

the inverse of its variance. This variance component consisted of the study-level sampling error variance as well as an estimate of between-study variance (Borenstein et al. 2009a). The appendix section 3.7.5 provides a description and interpretation of typical meta-analytic measures as well as further references. As inference should be made for a population of studies larger than the set of observed studies (Hedges and Vevea 1998), we used a random effects analysis. In the next step, we calculated the confidence interval for the mean effect size to indicate the degree of precision of the estimate and whether the mean effect size was statistically significant. In the third step, we performed a homogeneity analysis to assess whether the effect sizes came from the same population (random effects assumption). In the fourth step, we checked the robustness and quality of our findings by using a sensitivity analysis, an outlier analysis, and a publication bias check. The sensitivity analysis involved first calculating the effect size in a multilevel model by nesting the effect sizes in publications and then calculating the effect size separately for the old and the new studies. In the final analysis step, we conducted a mixed-effect model analysis for each moderator separately to determine which moderators had a significant influence on response rate differences. We used the R package "metafor" (version 1.9-9) for the analyses (Viechtbauer 2010).

3.5 Results

3.5.1 Study characteristics

Following our search strategy and eligibility criteria outlined above, we identified 75 manuscripts (24 from the previous study and 51 new manuscripts) that compared the response rates of web and other survey modes using split-sample randomized experimental designs. Because some of these manuscripts contained more than one response rate comparison, 114 response rate comparisons (k) (44 from the previous study and 70 new) were included in our study.

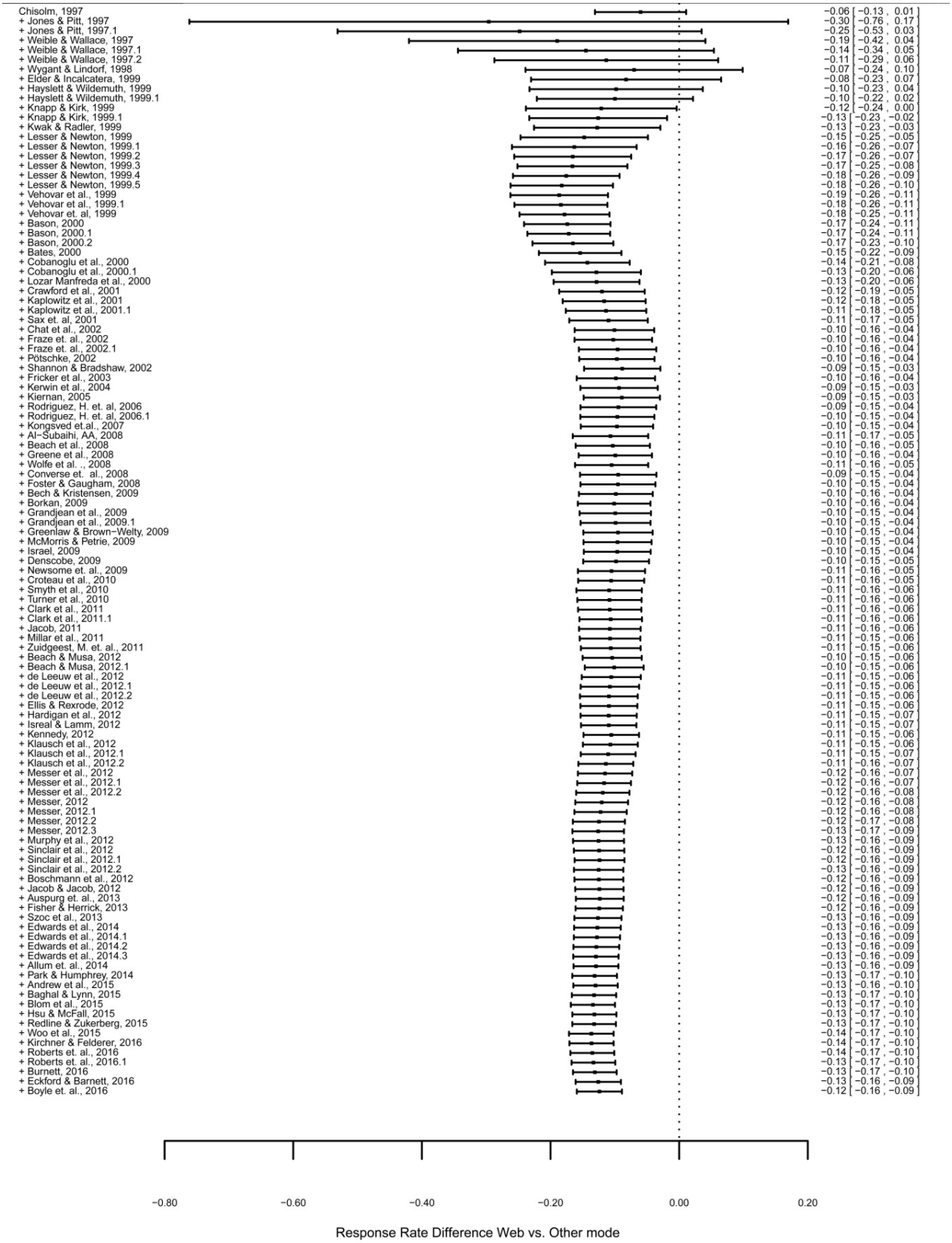
3.5.2 Mean response rate difference: Web surveys versus other survey modes

The sampling-error-weighted mean effect size estimate, computed across all 114 effect sizes under a random effects assumption, was -0.12 (95% CI = 0.16/0.09), which favors other survey modes over the web mode (Table 3.1, first line). This result indicates that web surveys yielded, on average, a 12 percentage points lower response rate compared to other survey modes.

How did the response rate difference develop over time? The response rate difference in the Lozar Manfreda et al. (2008) study was about 11 percentage points (95% CI = 0.15/0.06) (see table 3.1) lower for web surveys, and this value increased slightly in the present analysis to 12 percentage points (95% CI = 0.16/0.09). This result emphasizes the tendency of a basically stagnant, albeit slightly increased, response rate difference over time, which is also depicted by a cumulative forest plot (see figure 3.2) that chronologically describes the accumulation of evidence. The cumulative forest plot reveals two trends: First, the effect size becomes more precise over time (confidence intervals for the overall effect become smaller), which indicates a robust, time-invariant estimate that consistently favors other survey modes in terms of response rate differences. Second, the cumulative effect sizes have a slight tendency to the left, which indicates a rising response rate difference. Furthermore, when examining only the new effect sizes (2005–2016), we detected, on average, a 15 percentage points lower response rate for web surveys (see appendix table 3.8). Thus, to answer the first research question, our results indicate that, overall, the response rate difference remained constant over time, with the tendency to increase non-substantially in favor of other survey modes.

Is the effect size heterogeneous? A homogeneity analysis for all effect sizes reveals a significant Q-score of 7501 ($df = 114, p =\leq .0001$), which indicates the heterogeneity of the effect size distribution under the random effects assumption. This finding called for a moderator analysis to investigate whether moderators influenced the response rate difference (see next section). Before conducting this analysis, we addressed two questions regarding the validity of the findings: publication bias and robustness. Publication bias refers to the problem that significant results have a higher probability of being published and may distort the results.

Figure 3.2: Cumulative forest plot



*Note: * Fieldwork year as given*

Sensitivity analyses did not identify publication bias in our data. We also performed several robustness checks such as excluding the outliers, performing separate analyses for the old and the newer studies, and applying a multilevel approach for effect sizes nested in papers. All the mean response rate differences pointed in the same direction, and no significant differences could be detected. This suggests a robust overall effect size in terms of magnitude and direction. A detailed description of the validity testing is provided in section 3.7.5 of the appendix.

Table 3.1: Meta-analytic summary statistics - random effects model without moderators

n	**Mean r (95 % CI)**	**95% CI**	**T^2 (se)**	**Q_e (df/p)**	**I^2**	**H^2**
113	-0.11 (-0.14/ -0.08)	-0.1468/ -0.0820	0.03 (0.004)	7446.23 (112/ $\leq$ 0001)	99	119

3.5.3 Moderator analysis: Replication

This section presents the results for the moderators. First, the response rate difference was regressed on the survey mode to which web surveys were compared, the sample recruitment strategy, the target population, the type of sponsorship, the solicitation mode, the use of incentives, and the number of contacts – which are all the moderators included in the first meta-analysis (Lozar Manfreda et al. 2008). Second, we extended the original analysis by adding three new moderators: survey topic, prenotification, and survey country. Table 3.2 and Table 3.3 provide the results of the separate analyses that investigated the influence of moderators on the response rate difference between web and other survey modes. As indicated in the last column of Table 3.2, three of the six moderators—sample recruitment strategy, solicitation mode, and number of contacts—significantly explain the response rate difference ($p \leq 0.05$). All three moderators produced significant effects in the original meta-analysis as well. The quality statistics of the random effects models can be found in appendix Table 3.9. The average response rate difference for panel members or respondents from an existing list was nine percentage points lower for the web mode. This difference increased to 21 percentage points for one-time respondents (see Figure 3.3). A second influential moderator was the solicitation

mode: If participation was initially requested by a mode other than email, the response rate for web surveys was at least 14 percentage points lower. However, if respondents were asked by email to participate, this difference shrank, on average, to six percentage points difference. The final study characteristic that significantly influenced the response rate difference was the number of contact attempts. The results suggest that the larger the number of contacts was, the larger the response rate difference became (three percentage points difference for each contact attempt). This result suggests that contact attempts are less effective in the web survey mode. With regard to the target population, our findings indicate that specific populations showed only a small difference in response rates (students and employees: eight percentage points; business respondents: 12 percentage points), whereas the difference between the web mode and the compared mode in surveys of the general population increased distinctly ($p \leq 0.1$).

To sum up, and to answer the second research question, as in the original (Lozar Manfreda et al. 2008) meta-analysis, the sample recruitment base, solicitation mode, and number of contacts were found to have a significant effect on explaining the response rate difference between web and other survey modes. Contrary to the original study, the type of target population was significant on the 10 percent level. However, the compared mode, the use of incentives, the type of sponsorship, and the year the studies were conducted did not significantly explain the response rate difference.

3.5.4 Moderator analysis: Extension

This section presents an extension of the moderator analysis by including three new moderators that were not assessed in Lozar Manfreda et al. (2008) previous meta-analysis (table 3.3). Significant effects were observed for two of these three moderators. First, when prenotifications were used, this strategy was more effective in other survey modes than in web surveys. The use of prenotifications increased the response difference to 15 percentage points (see figure 3.3). This result suggests that survey prenotifications are more effective in any mode other than the web survey mode. This result is in line with our expectations that an email prenotification for a web survey is perceived by target persons to be less important because it involves minor

Table 3.2: Meta-analytic summary statistics - random effects model - replication

Moderator Variable	**Meta-analytic summary statistics (random effect model)**		
	Categories and Number of Cases	**Mean r (95 % CI)**	**p**
Type of Mode Compared to	E-Mail (10)	-0.13 (-0.26/-0.01)	.95
	Mail (70)	-0.12 (-0.16/-0.07)	
	Telephone (20)	-0.14 (-.24/-0.04)	
	Other (14)	-0.14 (-0.24/-0.04)	
Sample recruitment strategy	Panel/ pre-recruited list (10)	-0.09 (-0.21/0.01)	.01
	One-Time Recruitment (34)	-0.21 (-0.27/-0.14)	
	Existing List (70)	-0.09 (-0.13/-0.05)	
Target Population	Students (21)	-0.08 (-0.16/-0.00)	.09
	Employees/Members of Associations (34)	-0.8 (-0.15/-0.02)	
	Business Respondents (11)	-0.12 (-0.24/-0.01)	
	General Population (48)	-0.173 (-0.23/-0.12)	
Type of Sponsorship	Academic (78)	-0.13 (-0.17/-0.09)	.28
	Governmental (27)	-0.12 (-0.19/-0.05)	
	Commercial (9)	-0.04 (-0.17/0.09)	
Solicitation Mode	Mail (61)	-0.16 (-0.21/-0.12)	.02
	E-Mail (40)	-0.06 (-0.1/-0.03)	
	Other (13)	-0.14 (-0.23/-0.04)	
Incentive	Both modes used incentives (40)	-0.15 (-0.21/-0.09)	.02
	No mode used incentives (69)	-0.10 (-0.14/-.06)	
Number of Contacts (Cat.)	0-1 Contact Attempts (20)	-0.06 (-0.1/0.03)	.04
	2-4 Contact Attempts (70)	-0.15 (-0.18/-0.08)	
	5 or more Contact Attempts (5)	-0.03 (-0.13/0.07)	

effort on the part of the researcher. The second significant new moderator is the country in which the survey was conducted. Response rates for web surveys in the United States were, on average, only nine percentage points lower than for other survey modes; this figure rose to 16 percentage points for the United Kingdom and the Netherlands. We had to exclude other countries (Australia, Canada, Germany, Slovenia, and Sweden) from the analysis because less than five experiments in these countries were included in our meta-analysis, and the results would therefore have had little informative value. Providing a summary of the countries in geographical groups made little sense to us at this point, as attitudes to the World Wide Web cannot necessarily be delimited by geographical or continental borders. However, we tested geographically related and value-related (e.g. Hofstede 2016) categories, and the effects turned out to be very robust. As a result, the mode decision in the US should favor web surveys, whereas a much higher response rate difference is to be expected in the UK and the Netherlands. Nevertheless, it is important to point out that a low response rate difference can result from a particularly good performance of the web mode or from a low performance of the comparison mode on the other.

To answer our fourth research question, our findings show that the use of prenotifications and the country in which the survey is conducted significantly impacted survey response rates, whereas the survey topic did not. For the latter, it should be noted that we could not classify survey topics on the basis of their sensitivity. Following the relevant literature, this classification would have been particularly useful, as online respondents have been found to be more willing to disclose information on sensitive topics (Kreuter, Presser, and Tourangeau 2008). Table 3.4 provides an overview of all survey design characteristics and their development over time observed in the original metal-analysis and in our replication study.

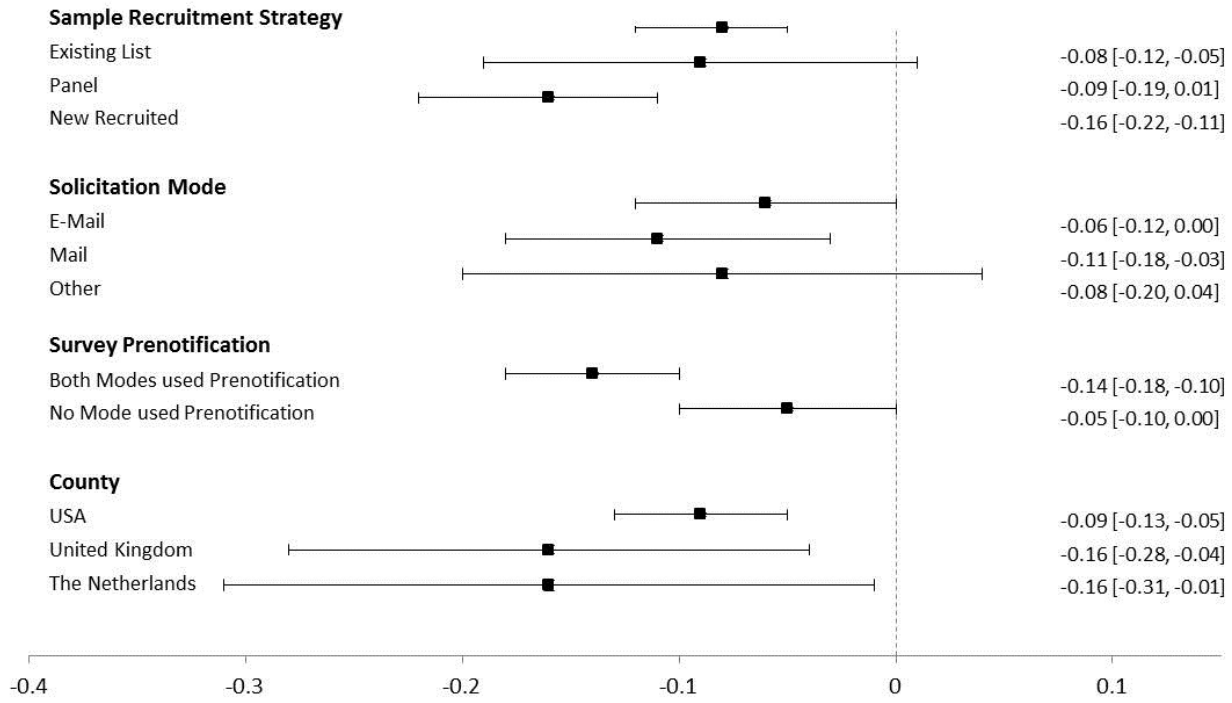
3.6 Discussion

Prior to the present study, the last meta-analysis on response rate differences between web surveys and other survey modes was conducted more than a decade ago (Lozar Manfreda et

Table 3.3: Meta-analytic summary statistics - random effects Model - extension

Moderator Variable	**Meta-analytic summary statistics (random effect model)**		
	Categories and Number of Cases	**Mean r (95 % CI)**	**p**
Survey Topic	Public Opinion (17)	-0.20 (-0.29/-0.11)	.39
	Professional Issue (Job) (32)	-0.12 (-0.19/-0.07)	
	Technology (12)	-0.14 (-0.21/-0.02)	
	Lifestyle (19)	-0.14 (-0.24/-0.03)	
	Other (33)	-0.09 (-0.15/-0.02)	
Prenotification for Study	Both Modes (63)	-0.14 (-0.18/-0.10)	.03
	No Mode (45)	-0.05 (-0.10/0.00)	
Survey Country	US (80)	-0.09 (-.013/-.05)	.01
	UK (10)	-0.16 (-0.28/-0.04)	
	NL (6)	-0.16 (-0.31/-0.09)	

Figure 3.3: Forest plot of significant categorical moderators



Response rate difference in percentage points

Notes: Number of contact attempts (continuous variable) was also significant.

Table 3.4: Overview of study design characteristics

Moderator Variable	**Moderator had a significant influence on the response rate difference in ...**	
	2008	**2017**
Type of Mode Compared to	not significant	not significant
Sample recruitment strategy	significant	significant
Target Population	not significant	not significant
Type of Sponsorship	not significant	not significant
Solicitation Mode	significant	significant
Incentive	not significant	significant
Number of Contacts (Cat.)	significant	significant
Publication Year	not significant	not significant
Region the Survey was conducted	-	significant
Prenotification for Study	-	significant
Survey Topic	-	not significant

al. 2008). Since then, the status and relevance of the web mode has changed. We examined these changes by including in our meta-analysis over 100 experiments related to response rate differences between web surveys and other modes. Overall, we found a basically stagnant heterogeneous mean response rate difference of 12 percentage points. Consequently, by choosing a web survey mode, researchers run the risk of achieving lower response rates than in traditional modes. Two groups of reasons can be used to explain this finding: long-term generic and contextual. The first group includes the lower perceived legitimacy of web surveys. Respondents may consider researchers' efforts to be less substantial—for example, "merely" sending an email compared to more time-consuming contact by telephone, where a researcher calls respondents once or even several times. The greater effort on the part of the researcher and the personal contact make it more difficult for the respondent to refuse to participate. Furthermore, the literature suggests that respondents perceive web surveys to be less mandatory, and web survey requests via email are often overlooked or routed to spam filters before they are read (Dillman, Smyth, and Christian 2014, pp. 419). Contextual reasons include increased web over-surveying. Because web surveys are quicker and cheaper, they are often used for surveys with limited resources. Furthermore, they now constitute the most popular survey mode worldwide (ESOMAR 2018). Moreover, respondents may receive a large amount of spam emails and find it difficult to distinguish between those that are relevant and those that are not. Other contextual reasons may be the greater sensitivity about security and privacy on the Internet (Marreiros, Tonin, and Vlassopoulos 2016) – especially in Europe since the General Data Protection Regulation became applicable in May 2018 (European Comission 2018) – and the great diversity of Internet browsers, devices (including mobile), and operating systems that can cause technical incompatibility problems.

In addition to studying change over time, the second and third aims of the present study were to increase the statistical power of the moderator analyses and to identify further influencing factors, especially as the mean effect size is heterogeneous. In the original study, Lozar Manfreda et al. (2008) demonstrated how the sample recruitment base, solicitation mode, and number of contacts significantly influenced the response rate difference. Our research corroborates these findings and revealed other significant moderators. More specifically, we found that using a

prenotification was more effective in all survey modes except the web mode, which confirms our assumption in this regard. People are more likely to overlook a prenotification via email than via traditional communication channels (Crawford, Couper, and Lamias 2001). One can argue that, in traditional survey modes, a researcher's investment in multiple contacts is perceived by the respondents to be an indication of the importance and legitimacy of the survey (Tuten 1997; Evans and Mathur 2005). Considerably more work is necessary to fully understand this phenomenon. Another significant predictor of the response rate difference is the survey country. Surveys conducted in the US produce higher web response rates or lower response rates in other modes, which results in a lower response rate difference overall. This suggests that the nonresponse problem in web surveys is lowest in the US. More research is needed to better understand the significant differences across countries and to determine the specific factors responsible for response rate differences at country level.

The present findings have substantial implications for the choice of survey mode. They offer cumulative evidence about the survey-environment factors that improve response rates in web surveys. To narrow the gap between response rates in web surveys and other survey modes, we therefore recommend forgoing the prenotification of web surveys, and instead using email solicitation and between one and two contact attempts. In an ideal case, the sample consists of panel respondents from a specific population in the US.

3.6.1 Limitations and further research

Changes in the web, and particularly in mobile technology, suggest that further meta-analyses should take into account different devices used to answer web questionnaires and the way in which they may be affecting response rates and, even more importantly, differences in nonresponse bias. Thus replications of the present cumulative meta-analysis to further track changes over time should include a mobile devices dimension.

The second limitation of the present study is that it does not account for the absolute response rate level. Although the response rate difference is small, it still ignores whether the absolute response rate was high (or low) in general across all modes. To gain further evidence about the

absolute web response level and its moderators, we strongly recommend that meta-analytical research be carried out in this regard.

The third limitation of this meta-analysis is that we estimated a large number of moderator models. Our findings could therefore be affected by the possibility of capitalizing on chance (rejecting a true null hypothesis). This means that some of the moderators in this meta-analysis may have shown significant results only by chance. Although a Bonferroni correction could remedy this, it is not recommended for power reasons (Schmidt and Hunter 2014; Polanin and Pigott 2015, p.83).

The fourth limitation of our meta-analysis is the fact that we could not address the critical issue of breakoff rates, because the breakoff rates in the web surveys and the compared modes were only occasionally reported, and a meta-analytical consideration of this topic was thus not possible. As the literature indicates that breakoff rates in academic or governmental web surveys are higher than in other response modes and range between 14 and 35 percent (Musch and Reips 2000; Lozar Manfreda and Vehovar 2002; Peytchev 2009; McGonagle 2013), one cause for the response rate gap could be different break-off probabilities. However, more research is needed on this issue.

Fifth, our study does not address whether the nonresponse rate is an indicator of nonresponse error and nonresponse bias. A low response rate does not necessarily lead to high nonresponse error, as the latter refers to the differences in the statistics between respondents and nonrespondents. Nonresponse error occurs if the nonrespondents – if they had responded – would have provided different answers than the actual respondents. Several studies have actually shown that low response rates do not necessarily indicate large nonresponse error (Keeter et al. 2000; Groves and Peytcheva 2008). However, a high response rate usually minimizes the probability that nonrespondents affect survey results, which is why we believe that the present study adds new knowledge of relevance for understanding nonresponse error.

However, further meta-analytic research needs to be done to establish whether these findings hold for other measures of web survey data quality – namely, on the one hand, representation-related errors and biases (as representativeness indicators) and, on the other hand, measurement-

related quality indicators (such as item nonresponse, consistency of answers, richness of responses to open-ended questions, speed of answering, acquiescence, social desirability, break-off, and conditioning effects). With respect to data quality, if it can be shown that responses from web survey modes are comparable to the responses from other survey modes, the problem of lower response rates in web surveys would not be as critical, particularly when one takes into account that fewer resources are needed to conduct web surveys. The present study did not consider that web surveys are usually cheaper to conduct compared to traditional modes. One could argue that the money saved by conducting a web survey can be used to produce better data quality and reduce the response rate difference, for example, by incentivizing reluctant respondents.

Related to this, it should be emphasized that inspection of the cumulative forest plot (figure 3.2) reveals that, starting from the year 2002, the response rate difference did not change substantially. Therefore, experiments that simply compared the response rate difference across different survey modes could have stopped then. Instead, more effort should have been invested in exploring the mechanisms that induce web survey participation. Further research should thus focus primarily on the value of the web mode: How can the value of online surveys be increased taking account of a variety of data quality indicators as well as the latest developments in mobile web surveys?

3.7 Appendix

3.7.1 Search strategy

	Lozar Manfreda et al. (2008)	Daikeler et al. (2018)
Search terms	web survey, Internet survey, online survey, web-based survey, Internet-based survey, electronic survey; supplemented by response rate, return rate, participation rate, and nonresponse rate	web survey, Internet survey, online survey, web-based survey, Internet-based survey, electronic survey; supplemented by response rate, return rate, participation rate, and nonresponse rate
Search engines	ScienceDirect , ISI Web of Knowledge, Directory of Open Access Journals, EBSCOhost, Emerald, Ingenta select, LookSmart's FindArticles, The Internet Public Library, Kluwer Online Journals, Proquest	Web of Science, Scopus, Proquest (ERIC, PsycINFO, Sociological Abstracts), Science Direct, Emerald Insight, Wiley Online Library, EconLit, PubMed, Business Source Premier, DOAJ, EconBiz, BASE, ipl.org, WebSM, Springerlink, Ebsco, (Google Scholar).

Table 3.5: Comparison of search terms and search engines

Source	Time Period
American Association for Public Opinion Research (AAPOR) Conference	2006–2015
General Online Research (GOR)	2006–2015
Joint Statistical Meetings	2005–2016
Other Conferences Listed at WebSM.org	2005–2016

Table 3.6: Conference overview

3.7.2 Variable overview

	Variable	Description	Scale/Categories
General	**Author(s)**	Name of author	nominal

	Title	Title of record	nominal
Effect Size	**Web Mode: Units Contacted**	Number of individuals randomly assigned to the web mode and eligible as per author's definition	counts
	Web Mode: Respondents	Number of respondents in the web mode as per author's definition	continuous
	Compared Mode: Units Contacted	Number of individuals randomly assigned to the compared mode and eligible as per author's definition	continuous
	Compared Mode: Respondents	Number of respondents in the compared mode as per author's definition	counts
Moderators	**Use of Incentives**	If incentives (pre- or postpaid) were used	ordinal/ both modes used incentives; no incentives used
	Number of Contacts	Maximum of contact attempts (incl. prenotification, main contact, follow ups)	ordinal/0–1 contact attempts; 2–4 contact attempts; 5 and more contact attempts
	Prenotification	If the study used a prenotification before the questionnaire was sent	ordinal/ for both mode; no prenotification
	Publication Year	Year the study was published	continuous

Survey Country	The country in which the survey was conducted	ordinal: US; UK; The Netherlands
Sample Recruitment Strategy	How the sample was recruited	ordinal/ panel[1]; existing list, e.g., membership list; one-time recruitment
Sponsorship	Who sponsored the survey	ordinal/ academic; governmental; commercial
Survey Topic	The topic of the questionnaire	ordinal/public opinion; professional issue, e.g., job; technology; lifestyle; other
Comparison Mode	The type of mode web was compared to	ordinal/email; mail; telephone; other
Type of Target Population	The target population for the survey	ordinal/students; employees or members of associations; business respondents; general population
Web Contact Mode	Which solicitation mode used the web survey	ordinal/ mail; email; other

Table 3.7: Variable and moderator overview

[1]In all studies, panel respondent refers to those respondents who are already participating in a panel and not to respondents who are asked to participate in a panel.

3.7.3 Robustness checks

For all studies	-0.12 (95% CI -0.16/0.09)
For new studies 2005-2016	-0.15 (95% CI -0.19/ -0.11)
For old studies 1997-2005	-0.08 (95% CI= -0.14/-0.02)
Averaged for Multi-effect sizes- Manuscripts	-0.11 (95% CI =-0.16/-0.08)
Without outlier	-0.10 (95% CI =-0.13 /-0.07)
Multilevel (effect size nested in authors)	-0.10 (95%CI =-0.15/-0.06), σ^2=0.04

Table 3.8: Sampling error weighted mean response rate difference overview

3.7.4 Measurement overview

This section describes in more detail some of the measures used in this contribution. For a detailed explanation, see Borenstein et al. (2009b).

T^2 - The proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies (T^2) is rather low in our model (0.03). This indicates that the observed variance is low and/or the variance within-studies is large Borenstein et al. 2009b, p. 115. The consideration of the effect sizes in the forest plot allows conclusions to be drawn about the latter.

Q_e, H^2 & I^2, are estimators describing the study heterogeneity. Q_e is used to determine the total amount of study-to-study variation Borenstein et al. 2009b, p. 115. H^2 can be interpreted as a standardized Q statistic. The I^2 statistic can be interpreted as the proportion of total variation in the estimates of treatment effect that is due to heterogeneity between studies (Higgins and Thompson 2002). All three measures indicate heterogeneous effect sizes; therefore, the explanation of heterogeneity with moderators is preferable.

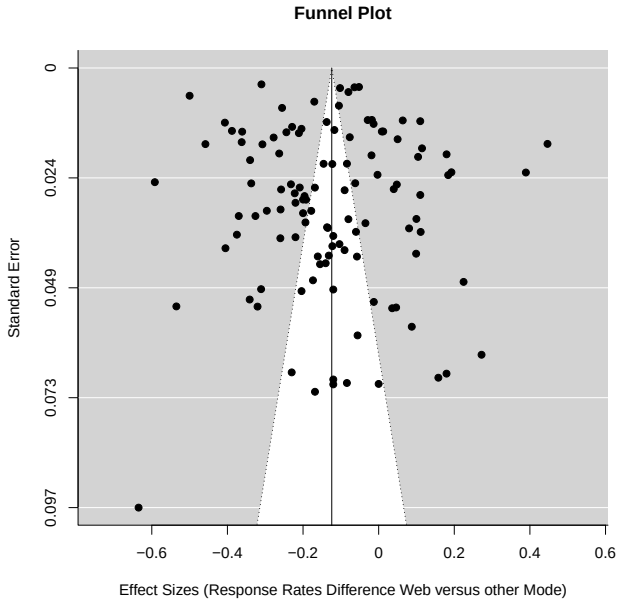


Figure 3.4: Funnel plot

3.7.5 Publication bias and sensitivity analysis

In a next step, we examine whether the mean response rate difference is influenced by publication bias. The funnel plot of Light and Pillemer (1984) in 3.4 is a visual method used to inspect publication bias. It shows the individual observed effect sizes on the x-axis against the corresponding standard errors. It is important that the point cloud on both sides of the line is approximately equal in number and distribution, indicating that both published and unpublished findings have comparable effect sizes and significance levels, and they hold true for our analysis. This result is emphasized by the Egger's regression test, which tests the asymmetry of the funnel plot. The result of this test is nonsignificant, which means that the funnel plot is not asymmetric and there is no evidence for a publication bias problem.

In addition, as proposed by Wang and Bushman (1998), plotting the quantiles of the effect size distribution against the quantiles of the normal distribution in a normal quantile plot does not give rise to concerns regarding a possible publication bias (see 3.5). The cases did neither deviate substantially from linearity nor have suspicious gaps.

In our sensitivity analysis, we performed several robustness checks. First, we calculated the average mean effect size separately for the old and new studies. Second, we excluded the

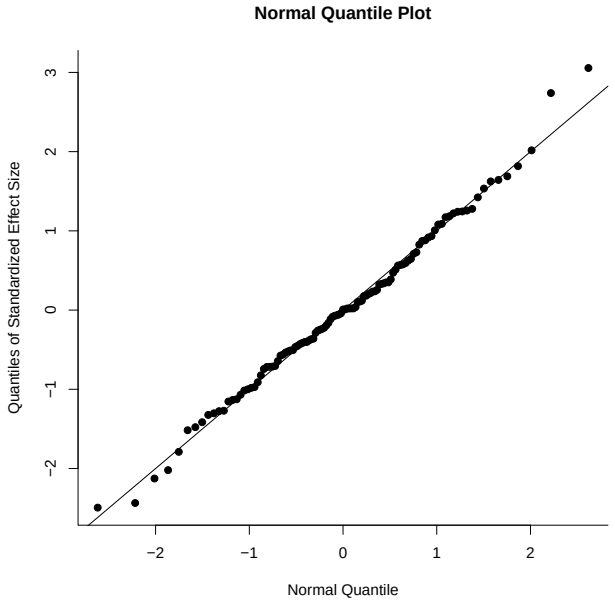


Figure 3.5: Normal quantile plot

five outliers (Jones and Pitt 1999; Al-Subaihi 2008; Converse, Wolfe, and Oswald 2008; Woo, Kim, and Couper 2015; Eckford and Barnett 2016b) with a larger Cook's distance than .04 (see Viechtbauer 2010). Then, with respect to the manuscripts with several effect sizes, we averaged the dependent effect sizes into one single mean response rate difference. Last, we conducted a multilevel random effect model analysis to model the fact that some effect sizes are nested within the examined studies, and the residual variance (σ^2) at author-level accounted for 4%. 3.8 provides an overview of the robustness analyses. All the mean response rate differences point to the same direction, and no significant differences could be detected. This suggests a robust overall effect size in terms of magnitude and direction.

3.7.6 Summary statistics of moderators

Table 3.9: Quality statistics for moderator analysis

Moderator Variable	**Heterogeneity estimators**			**Mixed - Effect Meta Regression**	
	T^2 (se)	**Q_e total (df/ p)**	**I^2**	**H^2**	**Model fit R^2**	**Q_m (df)**
Type of Mode Compared to	0.04 (0.005)	12181.12 (109/≤.001)	99.31	144.54	0.00	0.34 (3)
Sample recruitment strategy	0.03 (0.005)	14404.06 (110/≤.001)	99.30	142.04	7.14	9.95 (2)
Target Population	0.03 (0.005)	13201.72 (110/≤.001)	99.26	135.83	3.42	6.48 (3)
Type of Sponsorship	0.03 (0.005)	12169.78 (110/≤.001)	99.32	147.23	0.00	1.90 (3)
Solicitation Mode	0.03 (0.00)	13508.66 (110/≤.001)	99.34	5.40	5.40	7.86 (2)
Incentives	0.04 (0.00)	12276.50 (109/≤.001)	99.28	138.12	0.50	1.43 (1)
Number of Contacts	0.03 (0.005)	10726.04 (99/≤.001)	98.31	144.66	4.87	7.00 (2)
Survey Topic	0.03 (0.01)	10839.21 (107≤.001)	99.3	138.7	0.14	4.08 (4)
Prenotification for Study	0.03 (0.03)	11607.04 (105/≤.001)	99.3	136.5	3.45	4.68 (1)
Survey Country	0.03 (0.01)	11461.94 (96/≤.001)	99.1	112.2	8.42	10.45 (2)

References

Aichele, Corinna, Rob Flickenger, Carlo Fonda, Jim Forster, Ian Howard, Thomas Krag, and Marco Zennaro (2006). *Wireless networking in the developing world*.

*Al Baghal, Tarek and Peter Lynn (2015). "Using motivational statements in web-instrument design to reduce item-missing rates in a mixed-mode context". In: *Public Opinion Quarterly* 79.2, pp. 568–579. ISSN: 1537-5331.

*Auspurg, Katrin, Jonathan Burton, Carl Cullinane, Adeline Delavande, Laura Fumagelli, Maria Iacovou, Annette Jäckle, Olena Kaminska, Peter Lynn, Paul Mathews, et al. (2013). *Understanding Society Innovation Panel Wave 5: Results from methodological experiments*.

*Bales, Gregory T, Courtney MP Hollowell, Rajesh V Patel, and Glenn S Gerber (2000). "Internet and postal survey of endourologic practice patterns among American urologists". In: *The Journal of urology* 163.6, pp. 1779–1782. ISSN: 0022-5347.

*Bason, James (2000). *Comparison of telephone, mail, web, and IVR surveys of drug and alcohol use among University of Georgia students*. Conference Paper.

*Beach, Scott, Donald Musa, Patricia Beeson, and Carrie Sparks (2008). *Mode effects and non-response bias in an undergraduate student satisfaction survey: Results from a randomized experiment comparing telephone and web administration*. Conference Paper.

*Bech, Mickael and Morten Bo Kristensen (2009). "Differential response rates in postal and web-based surveys among older respondents". In: *Survey Research Methods* 3.1, pp. 1–6. ISSN: 1864-3361.

Blom, Annelies G, Jessica ME Herzing, Carina Cornesse, Joseph W Sakshaug, Ulrich Krieger, and Dayana Bossert (2017). "Does the recruitment of offline households increase the sample representativeness of probability-based online panels? Evidence from the German internet panel". In: *Social Science Computer Review* 35.4, pp. 498–520.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein (2009a). *Introduction to meta-analysis*. John Wiley and Sons. 457 pp. ISBN: 9780470057247.

Borenstein, Michael, Larry V Hedges, Julian Higgins, and Hannah R Rothstein (2009b). *Introduction to meta-analysis*. Hoboken, CA: Wiley Online Library. ISBN: 0470743387.

*Borkan, Bengue (2009). "The mode effect in mixed-mode surveys: Mail and web surveys". In: *Social Science Computer Review* 28.3, pp. 371–380. DOI: 10.1177/0894439309350698. URL: http://ssc.sagepub.com/content/28/3/371.full.pdf.

*Boschman, Julitta S., Henk F. van der Molen, Monique H. W. Frings-Dresen, and Judith K. Sluiter (2012). "Response rate of bricklayers and supervisors on an internet or a paper-and-pencil questionnaire". In: *International Journal of Industrial Ergonomics* 42.1, pp. 178–182. ISSN: 0169-8141. DOI: http://dx.doi.org/10.1016/j.ergon.2011.11.007. URL: http://www.sciencedirect.com/science/article/pii/S0169814111001326%20https://ac.els-cdn.com/S0169814111001326/1-s2.0-S0169814111001326-main.pdf?_tid=b0215cb4-b11b-4498-b699-6295ddfc0c88&acdnat=1538068579_f43afa6251d41abdc2fe4fb810037e3d.

*Boyle, Kevin J., Mark Morrison, Darla Hatton MacDonald, Roderick Duncan, and John Rose (2016). "Investigating Internet and mail implementation of stated-preference surveys while controlling for differences in sample frames". In: *Environmental and Resource Economics* 64.3, pp. 401–419. ISSN: 1573-1502. DOI: 10.1007/s10640-015-9876-2.

*Burnett, Craig M (2016). "Exploring the difference in participants' factual knowledge between online and in-person survey modes". In: *Research Politics* 3.2, pp. 1–7. ISSN: 2053-1680. DOI: 10.1177/2053168016654326.

Callegaro, Mario, Katja Lozar Manfreda, and Vasja Vehovar (2015). *Web survey methodology*. London, UK: Sage. ISBN: 1473927307.

*Cernat, Alexandru, Mick P. Couper, and Mary Beth Ofstedal (2016). "Estimation of mode effects in the health and retirement study using measurement models". In: *Journal of Survey Statistics and Methodology*, pp. 501–524.

Chang, Linchiat and Jon A Krosnick (2009). "National surveys via RDD telephone interviewing versus the Internet: Comparing sample representativeness and response quality". In: *Public Opinion Quarterly* 73.4, pp. 641–678.

*Chatt, Cindy, Michael Dennis, Rick Li, Alicia Motta-Stanko, and Paul Pulliam (2005). *Data collection mode effects controlling for sample origins in a panel survey: Telephone versus internet*. Conference Presentation. (Visited on 03/26/2019).

*Chisholm, John (1998). "Using the internet to measure and increase customer satisfaction and loyalty". In: *The Worldwide Internet Seminar*. Ed. by ESOMAR.

*Clark, Melissa, Michelle Rogers, and Andrew Foster (2011). "A randomized trial of the impact of survey design characteristics on response rates among nursing home providers". In: *Evaluation Health Prof.* 34, pp. 464–486.

*Cobanoglu, Cihan, Bill Warde, and Patrick J. Moreo (2001). "A comparison of mail, fax and web-based survey methods". In: *International Journal of Market Research* 43, pp. 405–410.

*Cole, Shu-Tian (2005). "Comparing mail and web-based survey distribution methods: Results of surveys to leisure travel retailers". In: *Journal of Travel Research* 43, pp. 422–430. DOI: 10.1177/0047287505274655. URL: http://jtr.sagepub.com/content/43/4/422.full.pdf.

*Converse, P. D., E. W. Wolfe, and F. L. Oswald (2008). "Response rates for mixed-mode surveys using mail and e-mail/Web". In: *American Journal of Evaluation*. DOI: 10.1177/1098214007313228. URL: http://aje.sagepub.com/content/29/1/99.full.pdf.

Cornesse, Carina and Michael Bosnjak (2018). "Is there an association between survey characteristics and representativeness? A meta-analysis". In: *Survey Research Methods* 12.1, pp. 1–13. ISSN: 1864-3361.

Couper, Mick P, Edith D De Leeuw, et al. (2003). "Nonresponse in cross-cultural and cross-national surveys". In: *Cross-cultural survey methods*, pp. 157–177.

Crawford, Scott D, Mick P Couper, and Mark J Lamias (2001). "Web surveys: Perceptions of burden". In: *Social science computer review* 19.2, pp. 146–162.

*Crawford, Scott, Sean McCabe, Mick Couper, and Carol Boyd (2002). "From mail to Web: Improving response rates and data collection efficiencies". In: pp. 25–28.

*Croteau, Anne-Marie, Linda Dyer, and Marco Miguel (2010). "Employee reactions to paper and electronic surveys: An experimental comparison". In: *IEEE Transactions on Professional Communication*. DOI: 10.1109/TPC.2010.2052852. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5556399%20http://ieeexplore.ieee.org/ielx5/47/5556396/05556399.pdf?tp=&arnumber=5556399&isnumber=5556396.

*De Leeuw, Edith, Gerry Nicolaas, Pamela Campanelli, and Joop Hox (2012). *Question or mode effects in mixed-mode surveys: A cross-cultural study in the Netherlands, Germany, and the UK*. Conference Paper.

*Denscombe, Martyn (2009). "Item non-response rates: A comparison of online and paper questionnaires". In: *International Journal of Social Research Methodology* 12.4, pp. 281–291. ISSN: 1364-5579. DOI: 10.1080/13645570802054706. URL: http://www.tandfonline.com/doi/abs/10.1080/13645570802054706.

Dillman, Don A, Jolene D Smyth, and Leah Melani Christian (2014). *Internet, phone, mail and mixed-mode surveys: The tailored design method*. John Wiley & Sons, pp. 1–528.

*Eckford, Rachel D. and Donell L. Barnett (2016a). "Comparing paper-and-pencil and Internet survey methods conducted in a combat-deployed environment". In: *Military Psychology* 28.4, pp. 209–225. ISSN: 0899-5605 1532-7876. DOI: 10.1037/mil0000118. URL: http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2016-18401-001&site=ehost-live%20Rachel.Eckford@gmail.com.

Eckford, Rachel D and Donell L Barnett (2016b). "Comparing paper-and-pencil and Internet survey methods conducted in a combat-deployed environment". In: *Military Psychology* 28.4, pp. 209–225.

*Edwards, Michelle L, Don A Dillman, and Jolene D Smyth (2014). "An experimental test of the effects of survey sponsorship on internet and mail survey response". In: *Public Opinion Quarterly* 78, pp. 734–750. DOI: 10.1093/poq/nfu027. URL: http://poq.oxfordjournals.org/content/78/3/734.full.pdf.

*Elder, Andrew and Tony Incalcatera (2000). *Pushing the envelope. Moving a major syndicated study to the Web*. Conference Paper.

*Ellis J. M.; Rexrode, D. L. (2012). *Addressed-based sampling – A better sample? Exploring the benefits of using addressed-based sampling in a state-wide targeted sub-population*. Statute. unpublished.

Eshet-Alkalai, Yoram and Eran Chajut (2010). "You can teach old dogs new tricks: The factors that affect changes over time in digital literacy". In: *Journal of Information Technology Education: Research* 9, pp. 173–181. ISSN: 1539-3585.

ESOMAR (2018). "Global market research report 2018: An ESOMAR Industry Report". In: URL: https://www.esomar.org/knowledge-center/library?publication=2898.

European Comission, EU (2018). *2018 Reform of EU data protection rules*. Legal Rule or Regulation. URL: https://ec.europa.eu/commission/priorities/justice-and-fundamental-rights/data-protection/2018-reform-eu-data-protection-rules_en.

Evans, Joel R and Anil Mathur (2005). "The value of online surveys". In: *Internet research* 15.2, pp. 195–219. ISSN: 1066-2243.

Fan, Weimiao and Zheng Yan (2010). "Factors affecting response rates of the web survey: A systematic review". In: *Computers in human behavior* 26.2, pp. 132–139.

*Fisher, S. H. and R. Herrick (2013). "Old versus new: The comparative efficiency of mail and internet surveys of state legislators". In: *State Politics Policy Quarterly*. DOI: 10.1177/1532440012456540. URL: http://spa.sagepub.com/content/13/2/147.full.pdf.

*Foster, Kelly N. and Monica Gaughan (2008). *Examining response rates and patterns in a multimode experiment: A study of department chairs/heads in STEM programs at research intensive universities*. Conference Paper.

*Fraze, Steve, Kelley Hardin, Todd Brashears, Jacqui Haygood, and James Smith (2003). "The effects of delivery mode upon survey response rate and perceived attitudes of Texas agriscience teachers". In: *Journal of Agricultural Education* 44.2, pp. 27–37.

*Fricker, Scott, Mirta Galesic, Roger Tourangeau, and Ting Yan (2005). "An experimental comparison of web and telephone surveys". In: *Public Opinion Quarterly* 69.3, pp. 370–392.

*Grandjean, Burke D., Nanette M. Nelson, and Patricia A. Taylor (2009). *Comparing an internet panel survey to mail and phone surveys on willingness to pay for environmental quality: A national mode test*. Conference Paper.

*Greene, Jessica, Howard Speizer, and Wyndy Wiitala (2008). "Telephone and web: Mixed-mode challenge". In: *Health services research* 43.1, pp. 230–248.

*Greenlaw, C and S Brown-Welty (2009). "A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost". In: *Evaluation Review*.

*Grigorian, Karen, Scott Sederstrom, and Thomas. Hoffer (2004). *Web of intrigue? Evaluating effects on response rates of between web SAQ, CATI, and mail SAQ options in a national panel survey*. Conference Paper.

Groves, Robert M, Floyd J Fowler Jr, Mick P Couper, James M Lepkowski, Eleanor Singer, and Roger Tourangeau (2011). *Survey methodology*. Vol. 561. John Wiley & Sons.

Groves, Robert M. and Emilia Peytcheva (2008). "The impact of nonresponse rates on nonresponse bias: A meta-analysis". In: *Public Opinion Quarterly* 72.2, pp. 167–189. ISSN: 0033-362X. DOI: 10.1093/poq/nfn011. URL: http://poq.oxfordjournals.org/content/72/2/167.abstract.

*Hardigan, Patrick C, Claudia Tammy Succar, and Jay M Fleisher (2012). "An analysis of response rate and economic costs between mail and web-based surveys among practicing dentists: A randomized trial." In: *Journal of community health*. DOI: 10.1007/s10900-011-9455-6.

*Hayslett, Michelle and Barbara Wildemuth (2005). "Pixels or pencils? The relative effectiveness of web-based versus paper surveys". In: *Library Information Science Research* 26.1, pp. 73–93.

Hedges, Larry V and Jack L Vevea (1998). "Fixed-and random-effects models in meta-analysis". In: *Psychological methods* 3.4, p. 486. ISSN: 1939-1463.

*Heerwegh, Dirk and Geert Loosveldt (2008). "Face-to-face versus Web surveying in a high-internet-coverage population: Differences in response quality". In: *Public Opinion Quarterly* 72.5. 10.1093/poq/nfn045, pp. 836–846. ISSN: 0033-362X. DOI: 10.1093/poq/nfn045. URL: http://dx.doi.org/10.1093/poq/nfn045.

Higgins, Julian and Simon Thompson (2002). "Quantifying heterogeneity in a meta-analysis". In: *Statistics in medicine* 21.11, pp. 1539–1558. ISSN: 1097-0258.

Hofstede, Geert (2016). *Cultural dimensions*. Web Page. URL: https://geert-hofstede.com/.

*Israel, GD (2012). *Using mixed-mode contacts to facilitate participation in public agency client surveys*. Conference Paper. URL: http://pdec.ifas.ufl.edu/satisfaction/articles/Using%5C%20Mixed-Mode%5C%20Contacts%5C%20Handout.pdf.

*Jacob, RT (2011). "An experiment to test the feasibility and quality of a web-based questionnaire of teachers". In: *Evaluation review*.

*Jones, Matt, Gary Marsden, Norliza Mohd-Nasir, Kevin Boone, and George Buchanan (1999). "Improving web interaction on small displays". In: *Computer Networks* 31.11, pp. 1129–1137. ISSN: 1389-1286.

*Jones, R and N Pitt (1999). "Health surveys in the workplace: Comparison of postal, email and World Wide Web methods." In: *Occupational medicine (Oxford, England)* 49.8, pp. 556–8. ISSN: 0962-7480. DOI: 10.1093/occmed/49.8.556. URL: http://occmed.oxfordjournals.org/content/49/8/556.full.pdf.

*Kaplowitz, Michael D., Timothy D. Hadlock, and Ralph Levine (2001). "A comparison of web and mail survey response rates". In: *Public Opinion Quarterly* 68, pp. 94–101. ISSN: 0033362X. DOI: 10.1093/poq/nfh006. URL: http://poq.oxfordjournals.org/content/68/1/94.full.pdf.

*Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M Groves, and Stanley Presser (2000). "Consequences of reducing nonresponse in a national telephone survey". In: *Public opinion quarterly* 64.2, pp. 125–148. ISSN: 0033-362X.

*Kerwin, Jeffrey, Pat D. Brick, Kerry Levin, David Cantor, Jennifer O'Brien, Andrew Wang, and Stephen-Shipp Stephanie (2004). *Web, mail, and mixed-mode data collection in a survey of Advanced Technology Program applicants*. Conference Paper.

*Kiernan, N. E. (2005). "Is a web survey as effective as a mail survey? A field experiment among computer users". In: *American Journal of Evaluation* 26.2, pp. 245–252. ISSN: 1098214005. DOI: 10.1177/1098214005275826. URL: http://aje.sagepub.com/content/26/2/245.full.pdf.

*Kim, Yujin, Jennifer Dykema, John Stevenson, Penny Black, and D Paul Moberg (2018). "Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys". In: *Social Science Computer Review*, p. 0894439317752406. ISSN: 0894-4393.

*Kirchner, Antje and Barbara Felderer (2016). "The effect of nonresponse and measurement error on wage regression across survey modes: A validation study". In: *Total Survey Error in Practice*, Ch. 25.

*Knapp, Herschel and Stuart Kirk (2003). "Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter?" In: *Computers in Human Behavior* 19.1, pp. 117–134.

*Kongsved, Sissel Marie, Maja Basnov, Kurt Holm-Christensen, and Niels Henrik Hjollund (2007). "Response rate and completeness of questionnaires: A randomized study of internet versus paper-and-pencil versions". In: *Journal of medical Internet research* 9.3. DOI: 10.2196/jmir.9.3.e25.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau (2008). "Social desirability bias in CATI, IVR, and web surveys". In: *Public Opinion Quarterly* 72, pp. 847–865.

Krippendorff, Klaus (2004). "Reliability in content analysis: Some common misconceptions and recommendations". In: *Human Communication Research* 30.3, pp. 411–433. URL: http://d%20x.doi.org/10.1111/j.1468-2958.2004.tb00738.x.

*Kwak, Nojin and Barry Radler (2002). "A comparison between mail and web surveys: Response pattern, respondent profile, and data quality". In: *Journal of official statistics* 18.2, p. 257.

*Lesser, Virginia and Lydia Newton (2001). "Mail, email and web surveys: A cost and response rate comparison in a study of undergraduate research activity". In: *AAPOR Annual Conference, Montreal, Quebec*.

Light, RJ and DB Pillemer (1984). "Quantitative procedures". In: *Suming up: the science of reviewing research*.

Lipsey, Mark W and David B Wilson (2001). "Analysis issues and strategies". In: *Practical Meta-Analysis*. Ed. by MW Lipsey and DB Wilson. Thousand Oaks, CA: SAGE Publications, Inc, pp. 105–128.

Lozar Manfreda, Katja, Michael Bosnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar (2008). "Web surveys versus other survey modes: A meta-analysis comparing response rates". In: *Journal of the Market Research Society* 50.1, p. 79. ISSN: 0025-3618.

*Lozar Manfreda, Katja and Vasja Vehovar (2002). "Survey design features influencing response rates in web surveys". In: *The International Conference on Improving Surveys Proceedings*. Citeseer, pp. 25–28.

*Lozar Manfreda, Katja, Vasja Vehovar, and Zenel Batagelj (2000). "Web versus mail questionnaire for an institutional survey". In: *The Challenge of the Internet*, pp. 1–11.

Lyberg, Lars and Pat Dean (1992). "Methods for reducing nonresponse rates: A review". In: *Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, FL*.

Marreiros, Helia, Mirco Tonin, and Michael Vlassopoulos (2016). "'Now that you mention it': A survey experiment on information, salience and online privacy". In: *CESifo Working Paper Series No. 5756*. URL: https://ssrn.com/abstract=2747891.

*McGonagle, Katherine A (2013). "Survey breakoffs in a computer-assisted telephone interview". In: *Survey research methods* 7.2, p. 79.

*McMorris, BJ and RS Petrie (2009). "Use of web and in-person survey modes to gather data from young adults on sex and drug use: An evaluation of cost, time, and survey error based on a randomized mixed-mode design". In: *Evaluation review* 33, pp. 138–158. URL: http://erx.sagepub.com/content/33/2/138.full.pdf.

*Messer, Benjamin L (2012). "Pushing households to the web: Experiments of a 'web+mail' methodology for conducting general public surveys". In: PHD work, unpublished.

*Millar, Morgan M, Don A Dillman, Benjamin Messer, Shaun Genter, Meredith Williams, and Thom Allen (2011). "Improving response to web and mixed-mode surveys". In: *Public Opinion Quarterly* 75, pp. 249–269. DOI: 10.1093/poq/nfr003. URL: http://poq.oxfordjournals.org/content/75/2/249.full.pdf.

Miller, Thomas, Michelle Miller Kobayashi, Erin Caldwell, Sarah Thurston, and Ben Collett (2002). "Citizen surveys on the web general population surveys of community opinion". In: *Social Science Computer Review* 20.2, p. 124 136.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Annals of internal medicine* 151.4, pp. 264–269. ISSN: 0003-4819.

Musch, Jochen and Ulf-Dietrich Reips (2000). "A brief history of Web experimenting". In: *Psychological experiments on the Internet*. Elsevier, pp. 61–87.

*Newsome, Jocelyn, Kerry Levin, Pat Dean Brick, Pat Langetieg, Melissa Vigil, and Michael Sebastiani (2009). *Multi-mode survey administration: Does offering multiple modes at once depress response rates?* Conference Paper.

*Park, A. and A. Humphrey (2014). *Mixed-mode surveys of the general population - Results from the European Social Survey mixed-mode experiment*. Conference Paper.

*Patrick, Megan E, Mick P Couper, Virginia B Laetz, John E Schulenberg, Patrick M O'Malley, Lloyd D Johnston, and Richard A Miech (2017). "A sequential mixed-mode experiment in the US National Monitoring the Future Study". In: *Journal of survey statistics and methodology* 6.1, pp. 72–97.

Petrovčič, Andraž, Gregor Petrič, and Katja Lozar Manfreda (2016). "The effect of email invitation elements on response rate in a web survey within an online community". In: *Computers in Human Behavior* 56, pp. 320–329.

Peytchev, Andy (2009). "Survey breakoff". In: *Public Opinion Quarterly* 73.1, pp. 74–97.

Polanin, Joshua R and Terri D Pigott (2015). "The use of meta-analytic statistical significance testing". In: *Research synthesis methods* 6.1, pp. 63–73.

*Al-Razgan, Muna S., Hend S. Al-Khalifa, Mona D. Al-Shahrani, and Hessah H. AlAjmi (2012). "Touch-based mobile phone interface guidelines and design recommendations for elderly people: A survey of the literature". In: Neural Information Processing. Springer Berlin Heidelberg, pp. 568–574. ISBN: 978-3-642-34478-7.

*Roberts, Caroline, Dominique Joye, and Michelle-Ernst Stähli (2016). "Mixing modes of data collection in Swiss social surveys: Methodological report of the LIVES-FORS mixed mode experiment".

*Rodriguez, Hector P, Ted von Glahn, William H Rogers, Hong Chang, Gary Fanjiang, and Dana Gelb Safran (2006). "Evaluating patients' experiences with individual physicians: A randomized trial of mail, internet, and interactive voice response telephone administration of surveys". In: *Medical care* 44.2, pp. 167–174. ISSN: 0025-7079.

Rosenthal, Robert (1979). "The file drawer problem and tolerance for null results". In: *Psychological bulletin* 86.3, p. 638. ISSN: 1939-1455.

Sakshaug, Joseph W., Ting Yan, and Roger Tourangeau (2010). "Nonresponse error, measurment error, and mode of data collection tradeoffs in a multi-mode survey of senstitve and non-senstitve items". In: *Public Opinion Quarterly* 74.5, pp. 907–933. DOI: 10.1093/poq/nfq057.

*Sax, Linda J, Shannon K. Gilmartin, and Alyssa N. Bryant (2001). "Assessing response rates and nonresponse bias in web and paper surveys". In: *Research in higher education* 44.1, pp. 409–432.

Schmidt, Frank L and John E Hunter (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Sage publications. ISBN: 1483324516.

*Shannon, David M. and Carol C. Bradshaw (2002). "A comparison of response rate, response time, and costs of mail and electronic surveys". In: *The Journal of Experimental Education* 70.2, pp. 179–192. ISSN: 00220973, 19400683. URL: http://www.jstor.org/stable/20152675.

Shih, Tse-Hua and Xitao Fan (2008). "Comparing response rates from web and mail surveys: A meta-analysis". In: *Field methods* 20.3, pp. 249–271. ISSN: 1525-822X.

Shojania, Kaveh G, Margaret Sampson, Mohammed T Ansari, Jun Ji, Steve Doucette, and David Moher (2007). "How quickly do systematic reviews go out of date? A survival analysis". In: *Annals of internal medicine* 147.4, pp. 224–233. ISSN: 0003-4819.

*Sinclair, Martha, Joanne O'Toole, Manori Malawaraarachchi, and Karin Leder (2012). "Comparison of response rates and cost-effectiveness for a community-based survey: Postal, internet and telephone modes with generic or personalised recruitment approaches". In: *BMC medical research methodology* 12.1, p. 132. ISSN: 1471-2288.

*Al-Subaihi, Ali A (2008). "Comparison of web and telephone survey response rates in Saudi Arabia". In: *The Electronic Journal of Business Research Methods* 6.2, pp. 123–132.

Tourangeau, Roger, J Michael Brick, Sharon Lohr, and Jane Li (2017). "Adaptive and responsive survey designs: A review and assessment". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.1, pp. 203–223. ISSN: 1467-985X.

Tuten, Tracy L (1997). *Getting a foot in the electronic door: Understanding why people read or delete electronic mail*. Report. ZUMA.

*Vehovar, Vasja, Katja Lozar Manfreda, and Zenel Batagelj (2001). "Sensitivity of electronic commerce measurement to the survey instrument". In: *International Journal of Electronic Commerce* 6, pp. 31–51.

Viechtbauer, Wolfgang (2010). "Conducting meta-analyses in R with the metafor package". In: *Journal of Statistical Software* 36.3, pp. 1–48.

Wang, Morgan C and Brad J Bushman (1998). "Using the normal quantile plot to explore meta-analytic data sets". In: *Psychological Methods* 3.1, p. 46. ISSN: 1939-1463.

*Weible, Rick and John Wallace (1998). "Cyber research: The impact of the Internet on data collection". In: *Marketing Research* 10.3, pp. 18–31.

*Wolfe, Edward W., Patrick D. Converse, and Frederick L. Oswald (2008). "Item-level nonresponse rates in an attitudinal survey of teachers delivered via mail and Web". In: *Journal of Computer-Mediated Communication* 14, pp. 35–66. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2008.01430.x.

*Woo, Youngje, Sunwoong Kim, and Mick P Couper (2015). "Comparing a cell phone survey and a web survey of university students". In: *Social Science Computer Review* 33.3, pp. 399–410. ISSN: 0894-4393. DOI: 10.1177/0894439314544876. URL: %3CGo%20to%20ISI%3E://WOS:000354306600007.

World Bank, ITU (2017). *Individuals using the Internet (% of population)*. URL: https://data.worldbank.org/indicator/it.net.user.zs (visited on 02/15/2019).

*Yeager, David S, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang (2011). "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples". In: *Public Opinion Quarterly* 75.4, pp. 709–747. DOI: 10.1093/poq/nfr020. URL: http://poq.oxfordjournals.org/content/75/4/709.full.pdf.

*Zuidgeest, Marloes, Michelle Hendriks, Laura Koopman, Peter Spreeuwenberg, and Jany Rademakers (2011). "A comparison of a postal survey and mixed-mode survey using a question-

naire on patients' experiences with breast care". In: *Journal of medical Internet research* 13.3, e68.

Chapter 4

Which Country-Level Factors Are Associated With Web Survey Response Rates?

A Meta-Analysis

4.1 Abstract

A major challenge in web-based cross-cultural data collections are the varying response rates, which can result in low data quality and nonresponse bias. Country-specific social, economic, and technological factors as well as the willingness of the population to participate in surveys may affect web response rates. This study attempts to evaluate web survey response behavior with meta-analytical methods based on more than 100 experimental studies from seven countries. Three dependent variables, so called effect sizes (web response rate, response rate of the comparison mode, and response rate difference), are used. Three country-specific factors had an impact on the performance of web survey response rates. Specifically, web surveys achieve high response rates in countries with a high population growth, high internet coverage, and a

high survey participation propensity, whereas they are at a disadvantage in countries with a high population age and mobile phone coverage. We conclude with practical implications for cross-cultural survey research.

4.2 Introduction

In an increasingly globalized world, cross-national research questions and thus cross-national datasets have become more important. Against the background of cost-intensive and inflexible face-to-face surveys, there are international attempts for web-based cross-cultural data collections (e.g., CRONOS Panel[1] , OPPA[2]). One of the major challenges in web-based data collection is nonresponse bias (Groves and Peytcheva 2008; Kreuter 2013; Bethlehem 2010). Noticeably, the nonresponse rate is not equal with nonresponse bias, but those two concepts are strongly related, and their relationship is moderated by survey design features such as the topic of the questionnaire and the survey population (Groves and Peytcheva 2008). Previous research, which aimed at explaining web response rates (e.g., Daikeler, Bosnjak, and Lozar Manfreda 2019; Lozar Manfreda et al. 2008), found that web surveys yield on average 12 percentage points lower response rates than their comparison modes, but could not explain large parts of the response rate heterogeneity. The reason for this might be that only survey design factors were included as explanatory variables in those studies.

However, differences in cross-national and cross-cultural non-survey design factors may also influence response behavior. Groves and Couper (2012) implicitly address this possibility in their nonresponse framework for household surveys by pointing out that the social environment can be a possible source of nonresponse. Blom (2012) explicitly addresses sampling unit macro level factors such as a country's gross domestic product (GDP) or survey culture in her conceptual model of country level contact rates. However, her analysis approach focuses on survey characteristics and she does not test any of the macro level factors.

While many research articles examine survey design factors, research on differences in cross-

[1]https://bit.ly/2U1BYcV
[2]https://openpanelalliance.org/ (Das, Kapteyn, and Bosnjak 2018))

country nonresponse is still in its infancy (Johnson, Lee, and Cho 2010). Recently, Daikeler, Bosnjak, and Lozar Manfreda (2019) replicated and extended a meta-analysis on web response rate differences and included three survey countries (US, UK and the Netherlands) in their moderator analyses. The authors found significant effects for the response rate difference between countries, but did not investigate this heterogeneity across countries in the response rates further. Previous cross-country comparative research on nonresponse has focused in particular on cultural dimensions and the link to nonresponse in interviewer administered surveys (Jans et al. 2019; Johnson, Lee, and Cho 2010) and cross-cultural dimensions of internet consumption (Hermeking 2005). Macroeconomic and country level factors have so far only been addressed in theoretical models, but to our knowledge they have never been examined in the context of web nonresponse.

Our study captures this research gap and has two goals, first to determine in which countries web surveys provide a valuable alternative as a survey mode and second to understand why web surveys work better in some countries than in others and what role indicators such as social, economic as well as technological factors and the country-specific survey participation propensity play for response rates in web surveys compared to other survey modes. Our study uses the strongest methodology available for comparative research, which is a meta-analysis solely based on experimental studies (APA 2006; Vandenbroucke 1998). By doing that, the study is a cross-national extension of previous research by Lozar Manfreda et al. (2008) and Daikeler, Bosnjak, and Lozar Manfreda (2019). Specifically, we focus on experimental web mode comparisons and analyze the respective web response rate, the comparison mode response rate, and the response rate difference between them.

4.3 Country-specific predictors of web survey response rates

The reasons for varying web response rates at the country level can be diverse. We identified four country-specific macro factors from the economic "PEST" (political, economic, socio-

cultural, technological) macro-level approach (Cadle, Paul, and Turner 2014, pp.3). However, we have renamed the socio-cultural factor "survey participation propensity" for our approach. As we want to explain response behavior and we are specifically referring to attitudes towards surveys as to general socio-cultural attitudes. Furthermore, we renamed the political factors to social factors as those indicators refer to social policy issues in the first place. Finally, we come up with the following macro-level factors influencing response behavior: social, economic and technological factors as well as the survey participation propensity (see Figure 4.1).

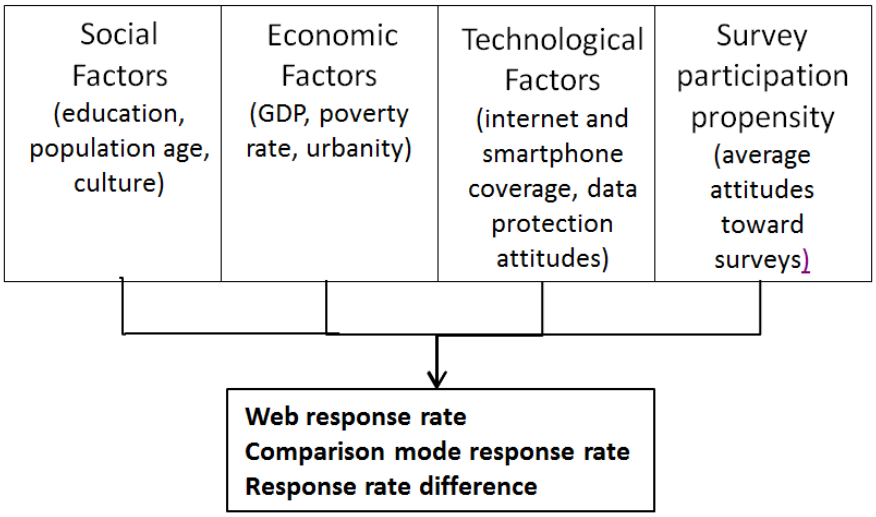


Figure 4.1: Macro-level factors for web response

4.3.1 Social factors

The country-specific social environment affects the acceptance and success of web surveys. How well web surveys perform compared to other modes depends to great degree on the perceived burden of web surveys (Crawford, Couper, and Lamias 2001). The lower the perceived burden for a web survey, the more likely are respondents to participate in web surveys. We expect that education and age are two determining factors as to how burdensome the survey is perceived. With respect to educational attainment, we expect that the higher the level of education in a country, the more accepted web surveys are in that country. Specifically, higher education is often linked to higher cognitive ability (Falch and Sandgren Massih 2011) and this is thought to reduce the effort to operate an online device and thus to complete a survey online. In contrary,

we do not expect a similar effect of education on the response rate of the comparison mode. Therefore, our assumption is that the higher the level of education in a country, the lower is the response rate difference.

Furthermore, we expect web surveys to be less accepted in countries with an ageing population, as older people are often less open-minded about new developments and find it harder to learn new skills (Charness and Boot 2009). Based on that, we expect a relatively good performance of the comparison mode in countries with an older population but low response rates for the web mode, which would result in a large response rate difference.

In addition, we expect countries with high population growth to achieve high web survey response rates compared to other survey modes. In countries with high population growth, there are many young people who are open to new ideas and the Internet is part of their everyday life. Thus, they generally have a lower burden for using the Internet. One example of this is the Arab Spring, in which the Internet played a central role as a communication medium (Howard et al. 2011). Moreover, it can be difficult to reach this young and mobile population with other survey modes due to their high mobility. We are aware that the proportion of older people in a country and its population growth might be correlated, but do not necessarily have to be (Fehr, Jokisch, and Kotlikoff 2008). In summary, we expect that the higher the population growth, the higher the web response rate, and the lower the comparison mode response rate. Consequently, this should minimize the difference between the modes.

4.3.2 Economic factors

The wealth of a country is essential for the performance of web surveys compared to other survey modes. In countries with a high level of wealth, Internet access supposedly is available to all population strata (Van Dijk 2006). This in turn means that the usage of the Internet is socially desirable, everyday life and that large parts of the population have the skills necessary to use it. Furthermore, in countries with a high level of prosperity, residents are often working in companies that use computers and thereby increase the familiarity with technology. In this way, the perceived burden for Internet use is minimal. For countries with a high GDP we,

therefore, expect that web surveys are well accepted by the population and that their response rate difference compared to other survey modes is rather small.

4.3.3 Technological development

Another important factor that may influence the web survey response rate in a specific country is the degree of technological development in that country (e.g. Bosnjak et al. 2005; Couper 2000; Couper et al. 2007; Rookey, Hanway, and Dillman 2008; Silber et al. 2018). The higher the Internet coverage, the easier it is for the population to use the Internet regularly. By using the Internet regularly, the web skills are trained and therefore the burden for participating in a web survey is reduced (Van Deursen and Van Dijk 2011). The higher the willingness to participate in a web survey is, the lower the difference between web and other survey modes. The same applies to the proportion of Internet users in the population. The more popular the Internet is among various social strata, the more likely it is that people will use it in their everyday live (Teo, Lim, and Lai 1999). The more the Internet is used, the lower the burden of using it to answer a survey. This may lead to higher web response rates and a lower response rate difference as comparison mode response rates should not be affected by the number of Internet users.

Finally, increased mobile phone network coverage in a country is expected to have a positive effect on web survey responses. The provision of mobile Internet across a country means that there are no longer any geographical or time limits for responding to web surveys (Wright 2005), while for other survey modes, such as face-to-face and mail surveys, there are more constraints. Therefore, we expect high web response rates and a low response rate difference for countries with broad network coverage.

4.3.4 Survey participation propensity

In the context of decreasing response rates in many countries and the greater than ever challenge to recruit respondents to participate in surveys of any survey mode (Atrostic, Bates, and Sil-

berstein 2001; Brick and Williams 2013; Curtin, Presser, and Singer 2005; Kreuter 2013; Rogers et al. 2004; Williams and Brick 2017), researchers have conducted international comparisons to determine factors that are related to higher and lower response rates, e.g. country-specific survey climate and response propensity (e.g. Barbier, Loosveldt, and Carton 2015; Beullens et al. 2018). One indicator for the acceptance of surveys in a country might be the willingness of citizens to participate in surveys of any mode in that country. A driver for the willingness to participate at country level might be data protection concerns (Gummer and Daikeler 2018), which are determined, for example, by the media but also by the history of a country (e.g., State security (STASI) in the GDR). Following this argumentation, we assume that the higher the willingness of citizens to participate in previous surveys of any mode in a country (influenced by, for instance, low data protection concerns and a positive attitude towards surveys), the higher the participation is for web surveys. One reasons for this might be that people with positive survey attitudes and low data protection concerns may assess web surveys as a more convenient and less burdensome way of participation. Consequently, web surveys may profit somewhat more from a positive survey climate (Loosveldt and Joye 2016). All in all, web surveys in countries with high response rate levels in previous surveys should perform equally well and lower response rate differences between modes can be expected.

4.4 The present study

Since a large part of the effect size heterogeneity remains unexplained in Lozar Manfreda et al. (2008) and Daikeler, Bosnjak, and Lozar Manfreda (2019) and since these two studies primarily focus on characteristics (such as the usage of incentives) of the included studies, this study will address the question of cross-country differences in web survey participation behavior. Using meta-analytic methods, we investigate in which countries web surveys receive high response rates compared to other survey modes and which country level indicators favor this. To do so, we will examine whether social, economic, and technological country-specific factors and the survey participation propensity influence the success of web surveys with the help of three effect sizes: the response rate of the web survey, the response rate of the comparison mode, and

the response rate difference between the two. Our findings might provide helpful information for researchers who aim at evaluating whether a web survey is likely to be a successful mode of data collection.

In the next section, we describe our methods and the operationalization of our moderators. In the following results section, we first give a descriptive overview of the selected experimental studies and then analyze whether there are cross-country differences in the performance of web surveys. Subsequently, we discuss our results in a broader context, limitations of the present study, and implications for future web data collections.

4.5 Method

This work uses the eligible studies of Lozar Manfreda et al. (2008) as a starting point and supplements it with further studies (Daikeler, Bosnjak, and Lozar Manfreda 2019). Our literature search and eligibility criteria are based on the search strategy and eligibility criteria of Lozar Manfreda et al. (2008). This section describes the meta-analytic methods used, the eligibility criteria and search strategy, the coding of the primary studies, and the statistical methods used.

4.5.1 Overview of meta-analytic procedure

Our meta-analysis comprises four steps. First, we conduct a comprehensive literature search for certain search terms. Second, we compare the manuscripts identified by this literature search with our eligibility criteria. Records that do not meet our criteria are excluded. Third, we code relevant data for calculating the response rates as well as the survey country. Based on the survey country, we then add country-specific information to our dataset. This supplementary data is based on the operationalization of social, economic, technological factors, and the survey participation propensity. To reflect the social status of a country, we use the average education, population growth, and the proportion of people over 65 in a country (see Table 4.1). We

operationalize the economic status of a country with the GDP of the respective country. We measure the degree of technical progress and openness toward technology through Internet and mobile phone coverage and the proportion of Internet users per country. For mapping the survey participation propensity in a specific country, we examine a variety of factors. First, we examine whether the response rate difference is significantly influenced by the web response rate or the response rate of the comparison mode. Furthermore, we include the aggregated response rates at the country level from the last five years and lastly we add the response rate of the International Social Survey Programme (ISSP) in the respective publication year to reflect the survey participation propensity of a country. Table 1 gives an overview of the sources of this additional information. Finally, we carry out the meta-analytical statistical analyses. Each of these four steps is explained in the subsequent sections.

4.5.2 Eligibility criteria and search strategy

The eligible studies must meet the following criteria: (1) A split-sample experimental design must have been performed on subjects from the same population who were randomly assigned to different survey modes. (2) One of the survey modes must be a web-based survey (i.e., a survey using a web questionnaire to collect respondents' answers online on a PC or laptop; mobile only studies were excluded). This web-based survey must be compared with data from at least one other survey mode (e.g., mail, telephone, face-to-face, or fax survey). (3) Data on response rates from the web and other survey modes as well as the survey country, which refers to the country in which the survey was conducted, must be available. (4) The subjects must have remained in the mode to which they were randomly assigned, i.e. studies in which the subjects could change modes were not eligible to participate. (5) The implementation of the compared survey modes must be identical. We don't have restrictions regarding population of participants, time period, and geography. This means we include studies in our meta-analysis regardless of which respondent population (e.g. such as student surveys) they use, regardless of when they are conducted and regardless of which country they are performed in. As a first important step to ensure the quality of our meta-analysis, we conduct a comprehensive literature

Table 4.1: Country-specific indicator: Sources

Factor	Variable	Source	Description
Social factors	Education	world bank	Education index
	Annual population growth	world bank	Annual population growth in a country by year and country
	Population ages 65 and over	OECD	The elderly population is defined as the share of people aged 65 and over (versus the working age -15-64 years) population) by year and country.
Economic factors	GDP	OECD	Gross domestic product (GDP) at market prices is the expenditure on final goods and services minus imports by year and country.
Technological factors	Internet coverage	world bank	Individuals using the Internet (% of population) by year and country
	Cellphone coverage	world bank	Mobile cellular subscriptions by year and country
	Internet users in %	world value survey	Using the internet (daily, weekly, monthly, less than monthly, never) by year and country
Survey participation propensity	Web response rate	calculated	Primary study (in %)
	Other mode response rate	calculated	Primary study (in %)
	Country-level aggregated Web response rate	calculated	Country- level 5 year aggregated value of current paper (in %)
	Country-level aggregated other mode response rate	calculated	Country- level 5 year aggregated value of current paper(in %)
	ISSP response rate of publication year	ISSP database	Response rate by publication year and country of the last ISSP round (in %)

Note. See the Online Appendix table 4.3 for a description of the exact sources including website and datasets of each indicator.

search (search terms: web survey, Internet survey, online survey, web-based survey, Internet-based survey, electronic survey; supplemented by response rate, return rate, participation rate, and nonresponse rate).

In order to overcome the problem of publication bias (Rosenthal 1979), we use various techniques. With the help of the snowballing technique, the reference lists of the manuscripts already selected are used. However, in order to collect explicitly unpublished studies, we compile abstracts of conferences (i.e.various scientific meetings from conferences, workshops, congresses, project meetings, invited lectures, among others). Conferences before 2005 are not included, since studies from these conferences are already included in the 25 manuscripts of the Lozar Manfreda et al.'s meta-analysis.

Information on coding strategy and intercoder reliability can be found in Daikeler, Bosnjak, and Lozar Manfreda (2019) and in the previous chapter of this dissertation.

4.5.3 Statistical method and effect sizes

Our effect sizes are the web response rate, the response rate of the comparison mode, and the response rate difference. Accordingly, we have calculated the number of invited and eligible subjects for each mode and compared them for the response rate difference. However, raw frequencies are essential for calculating the confidence interval for each effect size. In cases where insufficient data was provided, we used the authors' definition of the response rate and calculated the raw frequencies. We built a dummy variable on whether the authors provided the raw frequencies or response rates only. In the robustness analysis, the dummy variable did not show a significant moderation effect on the average response rate difference. The three effect sizes were calculated as follows:

$$d_{web} = \frac{I_{web}}{N_{web}}$$

$$d_{ComparisonMode} = \frac{I_{other}}{N_{other}}$$

$$d_{RRD} = \frac{I_{web}}{N_{web}} - \frac{I_{other}}{N_{other}}$$

with N_{web} − Web Respondents; I_{Web} − No of invited/ eligible subjects for web mode; N_{other}−Comparison Mode Respondents; I_{other}− No of invited/ eligible subjects for comparison mode

While the interpretation of the web and comparison mode response rates is intuitive, a positive response rate difference (d_{RRD})) refers to a higher response rate for the web mode compared to the other modes, and a negative response rate difference refers to a lower response rate for the web mode. Our three effect sizes are very closely linked and can be derived from each other. Nevertheless, we decided to report all three effect sizes, as further analyses show that 32 percent of the effect size heterogeneity (σ^2) at the country level (between cluster) can be explained by the response rate of the comparison mode (see appendix Table 4.4). Therefore, it is essential to consider also the performance of the comparison mode in order to assess whether web surveys in a country are a recommendable survey tool. In general, our statistical analysis comprises five steps (Lipsey and Wilson 2001). First, we calculate the weighted average response rate difference and the confidence interval per country by weighting each effect size with the inverse value of its variance. This variance component consists of the variance of sampling errors at the study level and an estimate of the variance between the studies (Borenstein et al. 2009). We use a random effect analysis because we aim at conclusions for a population that is larger than the amount of selected studies (Hedges and Vevea 1998). Due to the limited number of countries, we were not able to perform a multi-level meta-analysis, which would enable us to disentangle the differences between countries level from the characteristics of the studies (study level) (Cheung 2014).

In a second step, we perform a homogeneity analysis at the country level to determine whether the effect variables come from the same population and test if a moderator analysis is appropriate. In the third step, we check the robustness and quality of our results with sensitivity,

an outlier, and a publication bias analysis. In the fourth analysis step, we examine which country-specific factors have a significant influence on the response rate differences. For the analyses, the R-package 'metafor' (version 1.9-9) is used (Viechtbauer 2010). We choose "RD" (risk difference) as the effect size measure. The 'metafor' package automatically transforms a risk difference into the log of the effect size which makes these outcome measures symmetric around zero and enables a distribution of measures that is closer to a normal distribution.

4.6 Results

In this section, we first describe the descriptive characteristics of the studies (sections on study characteristics and cultural differences), and then, examine whether there are cross-cultural differences in response rates between the seven countries and which of the four country-specific factors might moderate those differences (section on country-specific predictors).

4.6.1 Study characteristics and sensitivity

The 110 studies that we identify as eligible are dominated by two characteristics. First, most (63%) of the web surveys are compared with mail surveys and second, most (73%) of the included studies were conducted in the United States (compare Figure 4.2).

In order to prevent distortions in our analyses due to the strong presence of US studies and mail comparisons, we conducted three additional sensitivity analyses to ensure the robustness of the results. Therefore, we replicated our results with subpopulations of the US studies as well as the mail comparison studies. First, we drew two random samples with a selection of US studies. Second, we performed the analysis for the mail comparisons only. Furthermore, our results might be biased because there could be a correlation between the comparison mode and the survey country. For this reason, we did not only calculate the results for mail mode alone, but also for telephone and face-to-face or the other comparison modes. All those analyses replicated the subsequent findings (see appendix table 4.5). In addition, since Daikeler, Bosnjak, and

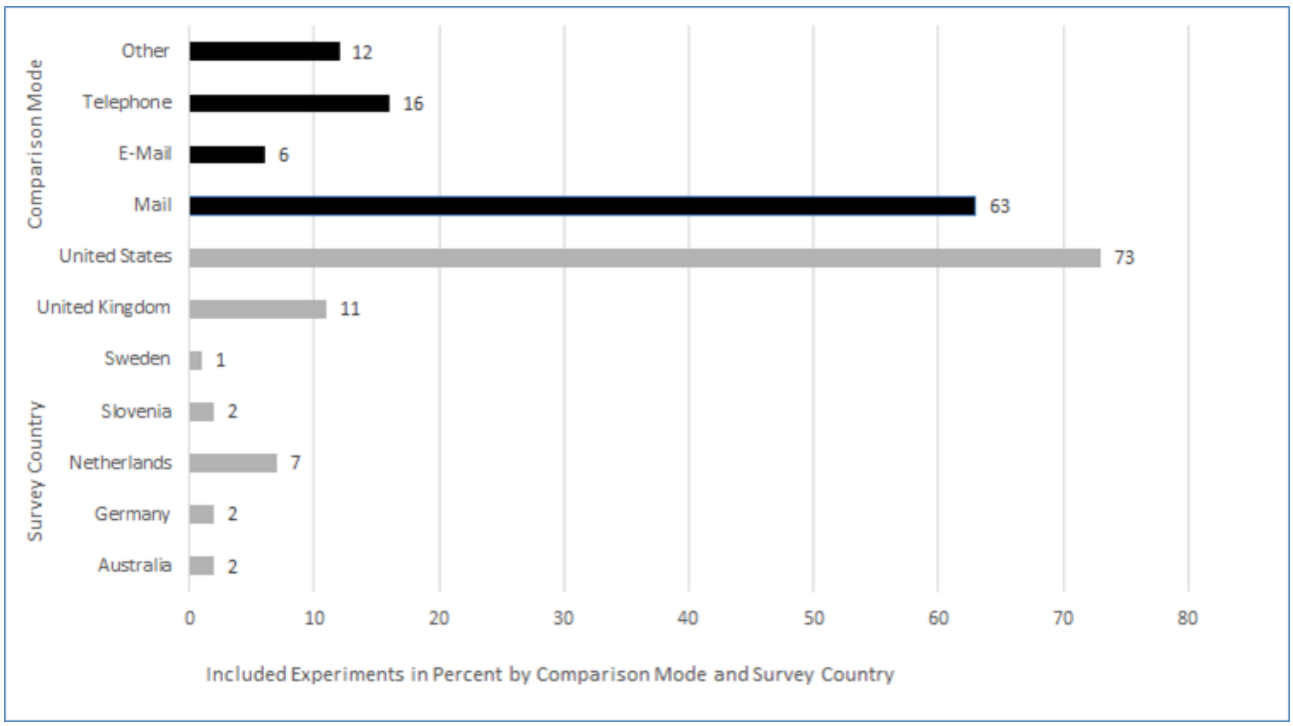


Figure 4.2: Comparison mode and survey country overview

Lozar Manfreda (2019) and Lozar Manfreda et al. (2008) concluded that the contact mode of the survey, the sample population, and the number of contact attempts determine the response rate difference between the web and the comparison mode, we investigated whether those variables correlate with the survey country to avoid pseudo-correlations. The results showed that those three factors correlate only marginally with the survey country (contact mode r=.13 & p=.17; sample population r=.12 & p=.22; number of contact attempts r=. 05 & p=.69).

4.6.2 Cultural differences in web surveys

Across all 110 experiments, the response rate difference between the web and the comparison mode is 12 percentage points (confidence interval: 9%/ 16%). Thereby, web surveys obtain on average a 36 percent response rate and the comparison mode 48 percent (see Figure 4). Notably, the response rate difference has remained stable compared to Lozar Manfreda et al.'s study in 2008 (11 percentage points). Consequently, web surveys are robustly inferior to other survey modes in terms of their response rates.

All three effect sizes are heterogeneous (significant Q-score of 7,501 ($df = 114, p \leq .0001$),

see third chapter, table 1.1). Heterogeneity of an effect size means that the value of the effect size, such as the response rate difference, is not consistent across the studies, but varies significantly. Consequently, a moderator analysis that aims to explain the heterogeneity is advisable (Borenstein et al. 2009). We address this heterogeneity by testing the country itself as a possible moderator, and our results showed significant differences in response rates at the country level for all three effect sizes, indicating that the performance of a web survey depends partly on the specific survey country. Specifically, the country level can account for none of the heterogeneity of the web response rate and for 12 percent of the response rate difference. For comparison mode the country level can explain 32 percent of the heterogeneity (see appendix Table 4.5).

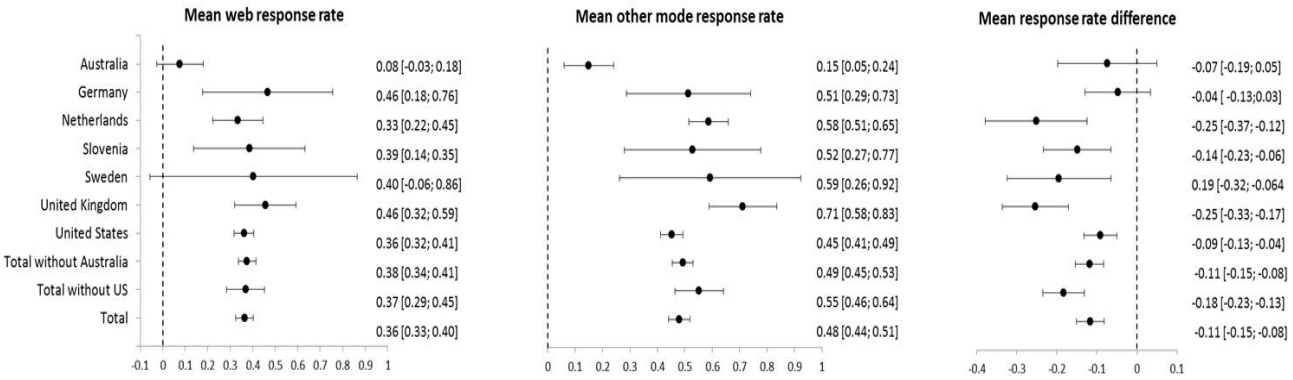


Figure 4.3: Response rate overview across countries

Focusing on the heterogeneity of the average effect sizes of the studies in Figure 4.3, it is visually apparent that the cross-country heterogeneity of the effect sizes of the web mode is smaller than the one of the comparison modes and the response rate difference.

The third forest plot in Figure 4.3 shows that the response rate difference is non-significant in Australia and Germany. Specifically, Australia seems to be an outlier with particularly low response rates. Further investigating this pattern, all three effect sizes do not show a substantial alternation and remain heterogeneous with respect to country differences if we exclude the country with the most studies (US) or the smallest effect sizes (Australia; see Figure 4.3, line 8 and 9). This further emphasizes the robustness of our findings. In the next section, we investigate in detail how this cross-country heterogeneity of effect sizes can be explained with social, economic, and technological factors as well as survey participation

propensity determinants.

4.6.3 Country-specific predictors for the success of web surveys

Considering the country-specific social factors of web survey response rates, especially the country-specific proportion of young and old people influences the success of web surveys (see Table 3). The higher the population growth and the lower the proportion of older people in a country, the smaller the difference in response rates between web surveys and their comparison mode. In our moderator analysis, however, this response rate difference seems to be especially influenced by the success or failure of the comparison mode. The more young people live in a society, the lower the response rates of face-to-face, mail, telephone or other survey modes. The opposite effect can be observed for the proportion of persons over 65 years. The more old people live in a country, the better the comparison modes work. In summary, the success of web surveys depend less on classic social factors as education than on the age of the population.

For economic factors such as the GDP, we expected that the higher the GDP of a country, the higher the prosperity and the more persons are able to access the Internet regularly. Therefore, we expected a high web response rate and a low response rate difference for those countries. When considering the results of our analysis, we do not detect this positive effect of the GDP on the success of web surveys compared to other survey modes (see Table 4.2).

Technological advancements of a country should also be taken into account in researchers' mode decisions. The results show that the higher the Internet and mobile phone coverage in a country, the higher are the response rates for the web and the comparison modes (see Table 4.2). For the response rate difference, our analysis in Table 3 shows that the better the Internet coverage in a country, the lower the response rate difference for these two modes. The opposite is shown for the mobile phone coverage. The higher the mobile phone coverage in a country, the larger the response rate difference between the web and the comparison mode. In other words, the higher the mobile phone coverage, the greater is the disadvantage of web surveys compared to other modes of data collection. The number of actual Internet users did not show a significant effect. To summarize the results for technological advancements, Internet and mobile phone coverage

Table 4.2: Social, economic, technological and survey participation propensity determinants for the success of web surveys

Outcome Measure		Response Rate Difference	Web response rate	Comparison mode response rate
Moderator				
Social factors	Education	n.s.	n.s.	n.s.
	annual population growth	0.1756 ***	n.s.	-0.2645 ***
	Population ages 65 and over	-0.018 **	n.s.	0.0299 ***
Economic factors	GDP	n.s.	n.s.	-0.0000 **
Technological advancement	Internet Coverage	0.0015 *	0.0015 *	0.0023 ***
	Cellphone Coverage	-0.0008 *	0.0010 *	0.0018 ***
	internet users	n.s.	n.s.	n.s.
Survey participation propensity	web response rate (reported in the study)	0.0039 ***	.	0.0061 ***
	other mode response rate (reported in the study)	-0.0038 ***	0.0062 ***	.
	Web response rate (aggregated)	n.s.	0.9969 ***	1.2871 ***
	Other mode response rate (aggregated)	-0.4199 **	0.4578 *	0.8818 ***
	Issp response rate	0.0027 **	n.s.	n.s.
Sig. level:	0.01 $\leq$ ***, 0.05 $\leq$ **, 0.10 $\leq$*			

moderated response rates for web surveys. Specifically, the results show that the better the mobile phone coverage, the larger the web survey response rate but at the same time also the response rate difference.

Finally, as the fourth factor, we examine the influence of the country-specific survey participation propensity on web surveys. Our analysis shows that a positive survey participation propensity leads to higher response rates (see Table 4.2). However, it also reveals that the comparison mode benefits even more from the positive survey participation propensity than the web mode. If the country-specific response rates of the comparison modes are generally high, web surveys also achieve higher response rates and vice versa. The response rate difference is mainly moderated by the response rate of the comparison mode - absolute and aggregated. The higher the response rates of the comparison modes, the larger the response rate difference. This phenomenon is also reflected in the response rate of the ISSP. The higher the response rate of the ISSP in a country, the larger the difference between web surveys and their comparison mode. This means that if the ISSP has a high response rate in a country, a large difference between the web survey mode and the comparison mode is to be expected.

Furthermore, as Cadle, Paul, and Turner (2014) explicitly mentioned socio-cultural factors, we tested Hofstede's individualism and uncertainty avoidance dimensions (Hofstede 2016), the Schwartz values concerning hedonism, conservation, and openness to change (Schwartz and Boehnke 2004), as well as the Internet usage and trust values from the World Value Survey (Inglehart et al. 2014) and did not find any effects. We decided against reporting those additional analyses in the results section because we did not see a plausible theoretical explanation, which links web survey participation behavior to those cultural concepts (Daikeler, Bosnjak, and Lozar Manfreda 2019). The complete results of these additional analyses are reported in the appendix Table 4.6.

4.7 Discussion

Our research addresses the question of whether the success of web surveys depends on the survey country and which country-specific indicators favor high web response rates. We developed this research question from previous meta-analytical research (Daikeler, Bosnjak, and Lozar Manfreda 2019; Lozar Manfreda et al. 2008), which used international data bases but did not further investigate heterogeneity of effect sizes on the country-level as their focus was on the impact of the survey design. In our study, we meta-analyzed a dataset that consisted of more than 100 split sample random experiments. Our results show that the survey country is a source of heterogeneity for each of our three effect sizes (web response rate, response rate of the comparison mode, and response rate difference between both modes). Driven by the question why in some countries cost-effective and time-saving web surveys might be more appropriate, while in others less, this meta-analytical study investigated whether country-specific factors have an impact on the web response rate. To this end, we studied four country-specific factors: social circumstances, economic circumstances, technological development, and the country-specific response propensity.

One of our main finding is that the heterogeneity of the difference in response rates between web and other modes across countries is to a large degree due to the performance of the comparison mode. Web surveys perform more similarly across countries than other survey modes. Given that the response rates of the comparison mode vary considerably across countries, the decision for or against a web survey in a specific country should always take the response rate expectations for alternative modes into account.

When considering country-specific factors that moderate cross-country differences in response rates, the results show that three out of four macro-economic factors have an impact on the web survey response rate. A higher web survey response rate is linked with high population growth, high internet coverage, and high response propensity. However, web surveys are seriously disadvantaged compared to other modes when a country's population is older and there is higher mobile phone coverage.

With respect to social factors, we find as expected, that the higher the population growth and the lower the proportion of older people in a country, the better web surveys work. For countries with a high level of education, we expected a better performance of web surveys, but we do not find this positive effect; possibly due to the very similar educational levels of the included countries.

As a proxy for economic factors, we tested the GDP and expected a positive association. However, this factor does not have an impact on web or comparison mode response rates. Again, a possible reason for this non-significant effect could be that all countries have a similar economic status.

Regarding technological factors, web surveys are more appropriate for countries with high Internet coverage rates. This finding is in line with our expectations as we assumed that high Internet coverage rates enable the population to use the Internet regularly and regular usage reduces the burden. However, our results also show that the better the mobile phone coverage of a country is, the higher the response rates for the comparison modes. Consequently, the response rate difference between the web and the comparison mode is larger in countries with high mobile phone coverage. This finding is surprising from today's perspective, since we often equate mobile phone coverage with mobile Internet coverage, but the analyzed studies are mainly from a time when there was no mobile Internet available.

For the country-based survey response propensity, we find that web surveys work well in countries with high comparison mode and ISSP response rates. This is in line with previous research in a region of Belgium, which was able to show that a low survey participation propensity leads to lower response rates and a higher number of contact attempts (Barbier, Loosveldt, and Carton 2015). Thus, our study enables us to generalize Barbier, Loosveldt, and Carton (2015) result regarding survey participation propensity and response rates across countries. For the survey mode decision, it can be concluded that web surveys will probably work in a country where the comparison mode performs good as well. However, the mode switch might result in a decrease of the response rate but might also bring organizational and financial benefits. Altogether, web surveys are an especially useful alternative to traditional modes when a country

has a young, technology-oriented, and survey-friendly population.

4.7.1 Practical implications

Our findings show that web surveys can be used as an alternative to other modes in all seven countries especially when other modes are not feasible due to survey costs or decreasing response rates. However, the web response rate decreased in almost all countries compared to other modes. It should be emphasized that the performance of web surveys with regard to their response rate compared to other survey modes is less dependent on the web mode itself than on the response rate of the comparison mode. So, the mode selection should always take the performance of alternative modes into account.

However, the decision for or against the web mode should not only be made on the basis of the expected response rate. Rather, the expected response rate is only a part of a complete data quality assessment. Other indicators should be considered at the same time, especially coverage and nonresponse bias (Sax, Gilmartin, and Bryant 2001; Fuchs and Busse 2009). Also, considerations of measurement error as well as expected field time and cost restrictions are further important factors influencing the mode decision (Couper 2000; Silber et al. 2018).

4.7.2 Limitations and further research

First, 73 percent of our studies were conducted in the US and despite our robustness analysis, evidence from other countries is needed to further strengthen our findings. This is especially important because most of the experimental studies were conducted in countries that are considered to have a western-world background. Including more (diverse) countries would also lead to more statistical power and a deeper understanding of the moderating factors. Furthermore, including more countries would enable researchers to use multi-level meta-analytic models, which would allow to separate country-level variance.

Second, we only searched English literature so that we may have excluded published experimental studies by that decision. Including other languages than English search terms would

especially in a cross-cultural context advisable.

Third, some authors (e.g. Rammstedt, Danner, and Bosnjak 2017; Johnson et al. 2018) and theoretical approaches (e.g. Cadle, Paul, and Turner 2014) have included value-oriented concepts such as Hofstede (2016) or Schwartz and Boehnke (2004) in their models. We tested several value concepts (Hofstede's individualism and uncertainty avoidance dimensions (Hofstede 2003), the Schwartz values concerning hedonism, conservation, and openness to change Schwartz and Boehnke (2004), the Internet usage and trust values from the World Value Survey (Inglehart et al. 2014) and did not find any effects. However, the concepts we tested are not optimal because, for example, no Hofstede values are available over time and other concepts, such as "open-mindedness towards new ideas", are not collected at all over time and in a cross-country context.

Fourth, we decided against including survey-based indicators such as incentives, the sample population, or the contact attempts, as Daikeler, Bosnjak, and Lozar Manfreda (2019) showed that those indicators could only explain a very small amount of the heterogeneity of the response rate difference. Nevertheless, the studies are not randomized across countries, comparison modes and survey-based indicators, though our sensitivity analyses separately by mode did not indicate systematic differences across modes.

Fifth, the response rate is only one data quality indicator among many. Nonresponse bias is often more relevant, especially because previous research has shown that low response rates can result in high response bias but do not have to be linked (Groves and Peytcheva 2008).

Lastly, the increasing popularity of mobile web surveys calls for their inclusion in future mode comparisons as well as meta-analyses. This avenue of research seems especially promising in a cross-cultural context as many persons in Asian and African countries access the Internet mainly with their smartphones (Statista 2018).

4.8 Appendix

4.8.1 Data sources

Variables	Source	Description	Scale
Education	Worldbank; http://data.worldbank.org/	Individuals using the Internet (% of population)	percent
GDP	OECD; https://data.oecd.org/gdp/	Gross domestic product (GDP) at market prices is the expenditure on final goods and services minus imports. GDP per capita data are measured in US dollars at current prices and PPPs	Dollar
Annual population growth	Worldbank; http://data.worldbank.org/	annual population growth in a country	percent

Population ages 65 and older	OECD; https://data.oecd.org/pop/	The elderly population is defined as people aged 65 and over. The share of the dependent population is calculated as total elderly and youth population expressed as a ratio of the total population. The elderly dependency rate is defined as the ratio between the elderly population and the working age (15-64 years) population.	percent
Internet coverage	Worldbank; http://data.worldbank.org/	Individuals using the Internet (% of population)	percent
Cellphone coverage	Worldbank; http://data.worldbank.org/	mobile cellular subscriptions (per 100 people)	percent
Internet users	World value survey; http://www.worldvaluessurvey.org		
Web response rate	included study		
Other mode response rate	included study		
Aggregated web response rate	generated from web response rate		

Aggregated other mode response rate	generated form other mode response rate	
Issp response rate	ISSP; http://www.issp.org/	percent

Table 4.3: Data Sources

4.8.2 Heterogeneity of effect sizes

Table 4.4: Heterogeneity differences in web and comparison mode

Response Rate Difference	I^2 (Residual heterogeneity/ unaccounted variability)	H^2 (Unaccounted variability / sampling variability)	R^2 (Amount of heterogeneity accounted for)	Amount of heterogeneity on country level in multilevel model
Web mode response rate aggregated	0.99	141.11	0.00	
Comparison mode response rate aggregated	0.99	134.79	0.06	
Web mode response rate	0.99	119.51	0.20	0.00%
Comparison mode response rate	0.99	119.03	0.18	31.96%

4.8.3 Robustness checks

Mail only

Country	Mean web response rate (p)	Mean other mode response rate	Mean response rate difference
Australia	0.0765 (≤ 0.0001)/n.s.	0.1502 (≤ 0.0001)/ n.s.	-0.0739 (≤0.0001)/ n.s.

Germany	0.4659 (≤0.0001)/ 0.5367 (0.000)	0.5135 (≤ 0.0001)/ 0.5797 (0.000)	-0.0479 (≤0.0001)/ n.s.
Netherlands	0.3347 (≤0.0001)/ 0.4148 (0.000)	0.5861 (≤0.0001)/ 0.5273 (0.000)	-0.2515 (≤0.0001)/ n.s.
Slovenia	0.3855 (≤0.0001)/ 0.5117 (0.000)	0.5288 (≤0.0001)/ 0.6405 (0.000)	-0.1490 (≤0.0001)/ n.s.
Sweden	0.4034 (≤0.0001)/ 0.4013 (0.001)	0.5922 (≤0.0001)/ 0.5912 (0.000)	-0.1945 (≤0.0001)/ n.s.
United Kingdom	0.4567 (≤0.0001)/ 0.3508 (0.001)	0.7118 (≤0.0001)/ 0.6789 (0.000)	-0.2538 (≤0.0001)/ -0.3285 (0.001)
United States	0.3623 (≤0.0001)/ 0.3446 (0.000)	0.4528 (≤0.0001)/ 0.4528 (0.000)	-0.0913 (≤ 0.0001)/ -0.1091 (0.000)

US Sample 1

Country	mean web response rate (p)/	mean other mode response rate	mean response rate difference
Australia	0.0765 (≤0.0001)/ n.s.	0.1502 (≤0.0001)/ n.s.	-0.0739 (≤0.0001)/ n.s.
Germany	0.4659 (≤0.0001)/ 0.4658 (0.000)	0.5135 (≤0.0001)/ 0.5132 (0.000)	-0.0479 (≤0.0001)/ n.s.
Netherlands	0.3347 (≤0.0001)/ 0.3355 (0.000)	0.5861 (≤0.0001)/ 0.5849 (0.000)	-0.2515 (≤0.0001)/ -0.2509 (0.000)
Slovenia	0.3855 (≤0.0001)/ 0.3854 (0.000)	0.5288 (≤0.0001)/ 0.5306 (0.000)	-0.1490 (≤0.0001)/ n.s.

Sweden	0.4034 (≤0.0001)/ 0.4022 (0.001)	0.5922 (≤0.0001)/ 0.5916 (0.000)	-0.1945 (≤0.0001)/ n.s.
United Kingdom	0.4567 (≤0.0001)/ 0.4567 (0.000)	0.7118 (≤0.0001)/ 0.7119 (0.000)	-0.2538 (≤0.0001)/ -0.2547 (0.000)
United States	0.3623 (≤0.0001)/ 0.4355 (0.000)	0.4528 (≤0.0001)/ 0.4131 (0.000)	-0.0913 (≤0.0001)/ n.s.

US Sample 2

Country	mean web response rate (p)/	mean other mode response rate	mean response rate difference
Australia	0.0765 (≤0.0001)/ n.s.	0.1502 (≤0.0001)/ 0.1507 (0.1)	-0.0739 (≤0.0001)/ n.s.
Germany	0.4659 (≤0.0001)/ 0.4658 (0.000)	0.5135 (≤0.0001)/ 0.5132 (0.000)	-0.0479 (≤0.0001)/ n.s.
Netherlands	0.3347 (≤0.0001)/ 0.3355 (0.000)	0.5861 (≤0.0001)/ 0.5849 (0.000)	-0.2515 (≤0.0001)/ -0.2515 (0.000)
Slovenia	0.3855 (≤0.0001)/ 0.3853 (0.000)	0.5288 (≤0.0001)/ 0.5307 (0.000)	-0.1490 (≤0.0001)/ -0.1908 (0.1)
Sweden	0.4034 (≤0.0001)/ 0.4020 (0.001)	0.5922 (≤0.0001)/ 0.5916 (0.000)	-0.1945 (≤0.0001)/ -0.1908 (0.1)
United Kingdom	0.4567 (≤0.0001)/ 0.4568 (0.000)	0.7118 (≤0.0001)/ 0.7120 (0.000)	-0.2538 (≤0.0001)/ -0.2544 (0.000)
United States	0.3623 (≤0.0001)/ 0.3245 (0.000)	0.4528 (≤0.0001)/ 0.3733 (0.000)	-0.0913 (≤0.0001)/ n.s.

Table 4.5: Robustness check by mode and US studies

Table 4.6: Cultural dimension

Value	**Description**	**Response Rate Difference**	**Web Response Rate**
Hofstede Individualism	The high side of this dimension, called Individualism, can be defined as a preference for a loosely-knit social framework in which individuals are expected to take care of only themselves and their immediate families. Its opposite, Collectivism, represents a preference for a tightly-knit framework in society in which individuals can expect their relatives or members of a particular in group to look after them in exchange for unquestioning loyalty. A society's position on this dimension is reflected in whether people's self-image is defined in terms of "I" or "we."	p=0.4532	p=0.5812
Hofstede Uncer-atainty Avoidance	The Uncertainty Avoidance dimension expresses the degree to which the members of a society feel uncomfortable with uncertainty and ambiguity. The fundamental issue here is how a society deals with the fact that the future can never be known: should we try to control the future or just let it happen? Countries exhibiting strong UAI maintain rigid codes of belief and behaviour, and are intolerant of unorthodox behaviour and ideas. Weak UAI societies maintain a more relaxed attitude in which practice counts more than principles.	p=0.4579	p=0.8372
Schwarz Openness	Openness to change: Stimulation, self-direction and some hedonism	p=0.6874	p=0.2134

References

*Al Baghal, Tarek and Peter Lynn (2015). "Using motivational statements in web-instrument design to reduce item-missing rates in a mixed-mode context". In: *Public Opinion Quarterly* 79.2, pp. 568–579. ISSN: 1537-5331.

APA (2006). "Evidence-based practice in psychology". In: *The American Psychologist* 61.4, p. 271.

Atrostic, Barbara K, Nancy Bates, and Adriana Silberstein (2001). "Nonresponse in US government household surveys: Consistent measures, recent trends, and new insights". In: *Journal of Official Statistics* 17.2, p. 209. ISSN: 0282-423X.

*Auspurg, Katrin, Jonathan Burton, Carl Cullinane, Adeline Delavande, Laura Fumagelli, Maria Iacovou, Annette Jäckle, Olena Kaminska, Peter Lynn, Paul Mathews, et al. (2013). *Understanding Society Innovation Panel Wave 5: Results from methodological experiments*.

*Bales, Gregory T, Courtney MP Hollowell, Rajesh V Patel, and Glenn S Gerber (2000). "Internet and postal survey of endourologic practice patterns among American urologists". In: *The Journal of urology* 163.6, pp. 1779–1782. ISSN: 0022-5347.

Barbier, Sara, Geert Loosveldt, and Ann Carton (2015). *The Flemish survey climate: An analysis based on the survey of social-cultural changes in Flanders*.

*Bason, James (2000). *Comparison of telephone, mail, web, and IVR surveys of drug and alcohol use among University of Georgia students*. Conference Paper.

*Beach, Scott, Donald Musa, Patricia Beeson, and Carrie Sparks (2008). *Mode effects and nonresponse bias in an undergraduate student satisfaction survey: Results from a randomized experiment comparing telephone and web administration*. Conference Paper.

*Bech, Mickael and Morten Bo Kristensen (2009). "Differential response rates in postal and web-based surveys among older respondents". In: *Survey Research Methods* 3.1, pp. 1–6. ISSN: 1864-3361.

Bethlehem, Jelke (2010). "Selection bias in web surveys". In: *International Statistical Review* 78.2, pp. 161–188. DOI: doi:10.1111/j.1751-5823.2010.00112.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2010.00112.x.

Beullens, Koen, Caroline Vandenplas, Geert Loosveldt, and Ineke Stoop (2018). "Response rates in the European Social Survey: Increasing, decreasing, or a matter of fieldwork efforts?" In: *Survey Methods: Insights from the Field*. ISSN: 2296-4754.

Blom, Annelies G. (2012). "Explaining cross-country differences in survey contact rates: Application of decomposition methods". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 175.1, pp. 217–242.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein (2009). *Introduction to meta-analysis*. John Wiley and Sons. 457 pp. ISBN: 9780470057247.

*Borkan, Bengue (2009). "The mode effect in mixed-mode surveys: Mail and web surveys". In: *Social Science Computer Review* 28.3, pp. 371–380. DOI: 10.1177/0894439309350698. URL: http://ssc.sagepub.com/content/28/3/371.full.pdf.

*Boschman, Julitta S., Henk F. van der Molen, Monique H. W. Frings-Dresen, and Judith K. Sluiter (2012). "Response rate of bricklayers and supervisors on an internet or a paper-and-pencil questionnaire". In: *International Journal of Industrial Ergonomics* 42.1, pp. 178–182. ISSN: 0169-8141. DOI: http://dx.doi.org/10.1016/j.ergon.2011.11.007. URL: http://www.sciencedirect.com/science/article/pii/S0169814111001326%20https://ac.els-cdn.com/S0169814111001326/1-s2.0-S0169814111001326-main.pdf?_tid=b0215cb4-b11b-4498-b699-6295ddfc0c88&acdnat=1538068579_f43afa6251d41abdc2fe4fb810037e3d.

Bosnjak, Michael, G Forsman, A Isaksson, Katja Lozar Manfreda, M. Schonlau, and T. Tuten (2005). "Preface to JOS special issue on web surveys". In: *Journal of Official Statistics* 22.

*Boyle, Kevin J., Mark Morrison, Darla Hatton MacDonald, Roderick Duncan, and John Rose (2016). "Investigating Internet and mail implementation of stated-preference surveys while controlling for differences in sample frames". In: *Environmental and Resource Economics* 64.3, pp. 401–419. ISSN: 1573-1502. DOI: 10.1007/s10640-015-9876-2.

Brick, J Michael and Douglas Williams (2013). "Explaining rising nonresponse rates in cross-sectional surveys". In: *The ANNALS of the American academy of political and social science* 645.1, pp. 36–59. ISSN: 0002-7162.

*Burnett, Craig M (2016). "Exploring the difference in participants' factual knowledge between online and in-person survey modes". In: *Research Politics* 3.2, pp. 1–7. ISSN: 2053-1680. DOI: 10.1177/2053168016654326.

Cadle, James, Debra Paul, and Paul Turner (2014). *Business analysis techniques*. Chartered Institute for IT.

*Cernat, Alexandru, Mick P. Couper, and Mary Beth Ofstedal (2016). "Estimation of mode effects in the health and retirement study using measurement models". In: *Journal of Survey Statistics and Methodology*, pp. 501–524.

Charness, Neil and Walter R Boot (2009). "Aging and information technology use: Potential and barriers". In: *Current Directions in Psychological Science* 18.5, pp. 253–258.

*Chatt, Cindy, Michael Dennis, Rick Li, Alicia Motta-Stanko, and Paul Pulliam (2005). *Data collection mode effects controlling for sample origins in a panel survey: Telephone versus internet*. Conference Presentation. (Visited on 03/26/2019).

Cheung, Mike W-L (2014). "Modeling dependent effect sizes with three-level meta-analyses: a structural equation modeling approach." In: *Psychological Methods* 19.2, p. 211.

*Chisholm, John (1998). "Using the internet to measure and increase customer satisfaction and loyalty". In: *The Worldwide Internet Seminar*. Ed. by ESOMAR.

*Clark, Melissa, Michelle Rogers, and Andrew Foster (2011). "A randomized trial of the impact of survey design characteristics on response rates among nursing home providers". In: *Evaluation Health Prof.* 34, pp. 464–486.

*Cobanoglu, Cihan, Bill Warde, and Patrick J. Moreo (2001). "A comparison of mail, fax and web-based survey methods". In: *International Journal of Market Research* 43, pp. 405–410.

*Cole, Shu-Tian (2005). "Comparing mail and web-based survey distribution methods: Results of surveys to leisure travel retailers". In: *Journal of Travel Research* 43, pp. 422–430. DOI: 10.1177/0047287505274655. URL: http://jtr.sagepub.com/content/43/4/422.full.pdf.

*Converse, P. D., E. W. Wolfe, and F. L. Oswald (2008). "Response rates for mixed-mode surveys using mail and e-mail/Web". In: *American Journal of Evaluation*. DOI: 10.1177/1098214007313228. URL: http://aje.sagepub.com/content/29/1/99.full.pdf.

Couper, Mick P (2000). "Web surveys: A review of issues and approaches". In: *The Public Opinion Quarterly* 64.4, pp. 464–494. ISSN: 0033-362X.

Couper, Mick P, Arie Kapteyn, Matthias Schonlau, and Joachim Winter (2007). "Noncoverage and nonresponse in an Internet survey". In: *Social Science Research* 36.1, pp. 131–148. ISSN: 0049-089X.

Crawford, Scott D, Mick P Couper, and Mark J Lamias (2001). "Web surveys: Perceptions of burden". In: *Social science computer review* 19.2, pp. 146–162.

*Crawford, Scott, Sean McCabe, Mick Couper, and Carol Boyd (2002). "From mail to Web: Improving response rates and data collection efficiencies". In: pp. 25–28.

*Croteau, Anne-Marie, Linda Dyer, and Marco Miguel (2010). "Employee reactions to paper and electronic surveys: An experimental comparison". In: *IEEE Transactions on Professional Communication*. DOI: 10.1109/TPC.2010.2052852. URL: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=5556399%20http://ieeexplore.ieee.org/ielx5/47/5556396/05556399.pdf?tp=&arnumber=5556399&isnumber=5556396.

Curtin, Richard, Stanley Presser, and Eleanor Singer (2005). "Changes in telephone survey nonresponse over the past quarter century". In: *Public opinion quarterly* 69.1, pp. 87–98. ISSN: 1537-5331.

Daikeler, Jessica, Michael Bosnjak, and Katja Lozar Manfreda (2019). "Web versus other survey modes : An Updated and extended meta-analysis compring response rates". In: *Journal of Survey Statistics and Methodology* forthcoming.

Das, Marcel, Arie Kapteyn, and Michael Bosnjak (2018). "Open probability-based panel infrastructures". In: *The Palgrave Handbook of Survey Research*. Springer, pp. 199–209.

*De Leeuw, Edith, Gerry Nicolaas, Pamela Campanelli, and Joop Hox (2012). *Question or mode effects in mixed-mode surveys: A cross-cultural study in the Netherlands, Germany, and the UK*. Conference Paper.

*Denscombe, Martyn (2009). "Item non-response rates: A comparison of online and paper questionnaires". In: *International Journal of Social Research Methodology* 12.4, pp. 281–291. ISSN: 1364-5579. DOI: 10.1080/13645570802054706. URL: http://www.tandfonline.com/doi/abs/10.1080/13645570802054706.

*Eckford, Rachel D. and Donell L. Barnett (2016). "Comparing paper-and-pencil and Internet survey methods conducted in a combat-deployed environment". In: *Military Psychology* 28.4, pp. 209–225. ISSN: 0899-5605 1532-7876. DOI: 10.1037/mil0000118. URL: http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=2016-18401-001&site=ehost-live%20Rachel.Eckford@gmail.com.

*Edwards, Michelle L, Don A Dillman, and Jolene D Smyth (2014). "An experimental test of the effects of survey sponsorship on internet and mail survey response". In: *Public Opinion Quarterly* 78, pp. 734–750. DOI: 10.1093/poq/nfu027. URL: http://poq.oxfordjournals.org/content/78/3/734.full.pdf.

*Elder, Andrew and Tony Incalcatera (2000). *Pushing the envelope. Moving a major syndicated study to the Web.* Conference Paper.

*Ellis J. M.; Rexrode, D. L. (2012). *Addressed-based sampling – A better sample? Exploring the benefits of using addressed-based sampling in a state-wide targeted sub-population.* Statute. unpublished.

Falch, Torberg and Sofia Sandgren Massih (2011). "The effect of education on cognitive ability". In: *Economic Inquiry* 49.3, pp. 838–856.

Fehr, Hans, Sabine Jokisch, and Laurence J Kotlikoff (2008). "Fertility, mortality and the developed world's demographic transition". In: *Journal of policy modeling* 30.3, pp. 455–473.

*Fisher, S. H. and R. Herrick (2013). "Old versus new: The comparative efficiency of mail and internet surveys of state legislators". In: *State Politics Policy Quarterly*. DOI: 10.1177/1532440012456540. URL: http://spa.sagepub.com/content/13/2/147.full.pdf.

*Foster, Kelly N. and Monica Gaughan (2008). *Examining response rates and patterns in a multimode experiment: A study of department chairs/heads in STEM programs at research intensive universities.* Conference Paper.

*Fraze, Steve, Kelley Hardin, Todd Brashears, Jacqui Haygood, and James Smith (2003). "The effects of delivery mode upon survey response rate and perceived attitudes of Texas agriscience teachers". In: *Journal of Agricultural Education* 44.2, pp. 27–37.

*Fricker, Scott, Mirta Galesic, Roger Tourangeau, and Ting Yan (2005). "An experimental comparison of web and telephone surveys". In: *Public Opinion Quarterly* 69.3, pp. 370–392.

Fuchs, Marek and Britta Busse (2009). "The coverage bias of mobile web surveys across European countries". In: *International Journal of Internet Science* 4.1, pp. 21–33.

*Grandjean, Burke D., Nanette M. Nelson, and Patricia A. Taylor (2009). *Comparing an internet panel survey to mail and phone surveys on willingness to pay for environmental quality: A national mode test*. Conference Paper.

*Greene, Jessica, Howard Speizer, and Wyndy Wiitala (2008). "Telephone and web: Mixed-mode challenge". In: *Health services research* 43.1, pp. 230–248.

*Greenlaw, C and S Brown-Welty (2009). "A comparison of web-based and paper-based survey methods testing assumptions of survey mode and response cost". In: *Evaluation Review*.

*Grigorian, Karen, Scott Sederstrom, and Thomas. Hoffer (2004). *Web of intrigue? Evaluating effects on response rates of between web SAQ, CATI, and mail SAQ options in a national panel survey*. Conference Paper.

Groves, Robert M and Mick P Couper (2012). *Nonresponse in household interview surveys*. John Wiley & Sons.

Groves, Robert M. and Emilia Peytcheva (2008). "The impact of nonresponse rates on nonresponse bias: A meta-analysis". In: *Public Opinion Quarterly* 72.2, pp. 167–189. ISSN: 0033-362X. DOI: 10.1093/poq/nfn011. URL: http://poq.oxfordjournals.org/content/72/2/167.abstract.

Gummer, Tobias and Jessica Daikeler (2018). "A note on how prior survey experience with self-administered panel surveys affects attrition in different modes". In: *Social Science Computer Review*, p. 0894439318816986.

*Hardigan, Patrick C, Claudia Tammy Succar, and Jay M Fleisher (2012). "An analysis of response rate and economic costs between mail and web-based surveys among practicing dentists: A randomized trial." In: *Journal of community health*. DOI: 10.1007/s10900-011-9455-6.

*Hayslett, Michelle and Barbara Wildemuth (2005). "Pixels or pencils? The relative effectiveness of web-based versus paper surveys". In: *Library Information Science Research* 26.1, pp. 73–93.

Hedges, Larry V and Jack L Vevea (1998). "Fixed-and random-effects models in meta-analysis". In: *Psychological methods* 3.4, p. 486. ISSN: 1939-1463.

*Heerwegh, Dirk and Geert Loosveldt (2008). "Face-to-face versus Web surveying in a high-internet-coverage population: Differences in response quality". In: *Public Opinion Quarterly* 72.5. 10.1093/poq/nfn045, pp. 836–846. ISSN: 0033-362X. DOI: 10.1093/poq/nfn045. URL: http://dx.doi.org/10.1093/poq/nfn045.

Hermeking, Marc (2005). "Culture and Internet consumption: Contributions from cross-cultural marketing and advertising research". In: *Journal of Computer-Mediated Communication* 11.1, pp. 192–216.

Hofstede, Geert (2003). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Sage publications. ISBN: 1452207933.

— (2016). *Cultural dimensions*. Web Page. URL: https://geert-hofstede.com/.

Howard, Philip N, Aiden Duffy, Deen Freelon, Muzammil M Hussain, Will Mari, and Marwa Maziad (2011). "Opening closed regimes: what was the role of social media during the Arab Spring?" In:

Inglehart, R, C Haerpfer, A Moreno, C Welzel, K Kizilova, J Diez-Medrano, M Lagos, P Norris, E Ponarin, and B Puranen (2014). *World Values Survey: All rounds—country-pooled datafile version*. Generic.

*Israel, GD (2012). *Using mixed-mode contacts to facilitate participation in public agency client surveys*. Conference Paper. URL: http://pdec.ifas.ufl.edu/satisfaction/articles/Using%5C%20Mixed-Mode%5C%20Contacts%5C%20Handout.pdf.

*Jacob, RT (2011). "An experiment to test the feasibility and quality of a web-based questionnaire of teachers". In: *Evaluation review*.

Jans, M., K. McLaughlin, J. Viana, D. Grant, R. Park, and N. A. Ponce (2019). "Geographic correlates of nonresponse in California". In: *Advances in Comparative Survey Methods*. Ed. by T Johnson, B. Pennell, I. A. Stoop, and Brita Dorer. DOI: doi:10.1002/9781118884997.

ch38. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118884997.ch38.

Johnson, Timothy P, Geon Lee, and Young Ik Cho (2010). "Examining the association between cultural environments and survey nonresponse". In: *Survey Practice* 3.3. ISSN: 2168-0094. URL: http://www.surveypractice.org/index.php/SurveyPractice/article/view/134/html.

Johnson, Timothy P, Beth?Ellen Pennell, Ineke AL Stoop, and Brita Dorer (2018). "The promise and challenge of 3MC research". In: *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, pp. 1–12.

*Jones, Matt, Gary Marsden, Norliza Mohd-Nasir, Kevin Boone, and George Buchanan (1999). "Improving web interaction on small displays". In: *Computer Networks* 31.11, pp. 1129–1137. ISSN: 1389-1286.

*Jones, R and N Pitt (1999). "Health surveys in the workplace: Comparison of postal, email and World Wide Web methods." In: *Occupational medicine (Oxford, England)* 49.8, pp. 556–8. ISSN: 0962-7480. DOI: 10.1093/occmed/49.8.556. URL: http://occmed.oxfordjournals.org/content/49/8/556.full.pdf.

*Kaplowitz, Michael D., Timothy D. Hadlock, and Ralph Levine (2001). "A comparison of web and mail survey response rates". In: *Public Opinion Quarterly* 68, pp. 94–101. ISSN: 0033362X. DOI: 10.1093/poq/nfh006. URL: http://poq.oxfordjournals.org/content/68/1/94.full.pdf.

*Keeter, Scott, Carolyn Miller, Andrew Kohut, Robert M Groves, and Stanley Presser (2000). "Consequences of reducing nonresponse in a national telephone survey". In: *Public opinion quarterly* 64.2, pp. 125–148. ISSN: 0033-362X.

*Kerwin, Jeffrey, Pat D. Brick, Kerry Levin, David Cantor, Jennifer O'Brien, Andrew Wang, and Stephen-Shipp Stephanie (2004). *Web, mail, and mixed-mode data collection in a survey of Advanced Technology Program applicants*. Conference Paper.

*Kiernan, N. E. (2005). "Is a web survey as effective as a mail survey? A field experiment among computer users". In: *American Journal of Evaluation* 26.2, pp. 245–252. ISSN: 1098214005.

DOI: 10.1177/1098214005275826. URL: http://aje.sagepub.com/content/26/2/245.full.pdf.

*Kim, Yujin, Jennifer Dykema, John Stevenson, Penny Black, and D Paul Moberg (2018). "Straightlining: Overview of measurement, comparison of indicators, and effects in mail–web mixed-mode surveys". In: *Social Science Computer Review*, p. 0894439317752406. ISSN: 0894-4393.

*Kirchner, Antje and Barbara Felderer (2016). "The effect of nonresponse and measurement error on wage regression across survey modes: A validation study". In: *Total Survey Error in Practice*, Ch. 25.

*Knapp, Herschel and Stuart Kirk (2003). "Using pencil and paper, Internet and touch-tone phones for self-administered surveys: Does methodology matter?" In: *Computers in Human Behavior* 19.1, pp. 117–134.

*Kongsved, Sissel Marie, Maja Basnov, Kurt Holm-Christensen, and Niels Henrik Hjollund (2007). "Response rate and completeness of questionnaires: A randomized study of internet versus paper-and-pencil versions". In: *Journal of medical Internet research* 9.3. DOI: 10.2196/jmir.9.3.e25.

Kreuter, Frauke (2013). "Facing the nonresponse challenge". In: *The ANNALS of the American Academy of Political and Social Science* 645.1, pp. 23–35. DOI: 10.1177/0002716212456815. URL: http://ann.sagepub.com/content/645/1/23.abstract.

*Kwak, Nojin and Barry Radler (2002). "A comparison between mail and web surveys: Response pattern, respondent profile, and data quality". In: *Journal of official statistics* 18.2, p. 257.

*Lesser, Virginia and Lydia Newton (2001). "Mail, email and web surveys: A cost and response rate comparison in a study of undergraduate research activity". In: *AAPOR Annual Conference, Montreal, Quebec.*

Lipsey, Mark W and David B Wilson (2001). "Analysis issues and strategies". In: *Practical Meta-Analysis*. Ed. by MW Lipsey and DB Wilson. Thousand Oaks, CA: SAGE Publications, Inc, pp. 105–128.

Loosveldt, Geert and Dominique Joye (2016). "Defining and assessing survey climate". In: *The SAGE Handbook of Survey Methodology*, pp. 67–77.

Lozar Manfreda, Katja, Michael Bosnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar (2008). "Web surveys versus other survey modes: A meta-analysis comparing response rates". In: *Journal of the Market Research Society* 50.1, p. 79. ISSN: 0025-3618.

*Lozar Manfreda, Katja and Vasja Vehovar (2002). "Survey design features influencing response rates in web surveys". In: *The International Conference on Improving Surveys Proceedings*. Citeseer, pp. 25–28.

*Lozar Manfreda, Katja, Vasja Vehovar, and Zenel Batagelj (2000). "Web versus mail questionnaire for an institutional survey". In: *The Challenge of the Internet*, pp. 1–11.

*McGonagle, Katherine A (2013). "Survey breakoffs in a computer-assisted telephone interview". In: *Survey research methods* 7.2, p. 79.

*McMorris, BJ and RS Petrie (2009). "Use of web and in-person survey modes to gather data from young adults on sex and drug use: An evaluation of cost, time, and survey error based on a randomized mixed-mode design". In: *Evaluation review* 33, pp. 138–158. URL: http://erx.sagepub.com/content/33/2/138.full.pdf.

*Messer, Benjamin L (2012). "Pushing households to the web: Experiments of a 'web+mail' methodology for conducting general public surveys". In: PHD work, unpublished.

*Millar, Morgan M, Don A Dillman, Benjamin Messer, Shaun Genter, Meredith Williams, and Thom Allen (2011). "Improving response to web and mixed-mode surveys". In: *Public Opinion Quarterly* 75, pp. 249–269. DOI: 10.1093/poq/nfr003. URL: http://poq.oxfordjournals.org/content/75/2/249.full.pdf.

*Newsome, Jocelyn, Kerry Levin, Pat Dean Brick, Pat Langetieg, Melissa Vigil, and Michael Sebastiani (2009). *Multi-mode survey administration: Does offering multiple modes at once depress response rates?* Conference Paper.

*Park, A. and A. Humphrey (2014). *Mixed-mode surveys of the general population - Results from the European Social Survey mixed-mode experiment*. Conference Paper.

*Patrick, Megan E, Mick P Couper, Virginia B Laetz, John E Schulenberg, Patrick M O'Malley, Lloyd D Johnston, and Richard A Miech (2017). "A sequential mixed-mode experiment in the US National Monitoring the Future Study". In: *Journal of survey statistics and methodology* 6.1, pp. 72–97.

Rammstedt, Beatrice, Daniel Danner, and Michael Bosnjak (2017). "Acquiescence response styles: A multilevel model explaining individual-level and country-level differences". In: *Personality and Individual Differences* 107, pp. 190–194. ISSN: 0191-8869.

*Al-Razgan, Muna S., Hend S. Al-Khalifa, Mona D. Al-Shahrani, and Hessah H. AlAjmi (2012). "Touch-based mobile phone interface guidelines and design recommendations for elderly people: A survey of the literature". In: Neural Information Processing. Springer Berlin Heidelberg, pp. 568–574. ISBN: 978-3-642-34478-7.

*Roberts, Caroline, Dominique Joye, and Michelle-Ernst Stähli (2016). "Mixing modes of data collection in Swiss social surveys: Methodological report of the LIVES-FORS mixed mode experiment".

*Rodriguez, Hector P, Ted von Glahn, William H Rogers, Hong Chang, Gary Fanjiang, and Dana Gelb Safran (2006). "Evaluating patients' experiences with individual physicians: A randomized trial of mail, internet, and interactive voice response telephone administration of surveys". In: *Medical care* 44.2, pp. 167–174. ISSN: 0025-7079.

Rogers, A, Maureen A Murtaugh, S Edwards, and ML Slattery (2004). "Contacting controls: Are we working harder for similar response rates, and does it make a difference?" In: *American journal of epidemiology* 160.1, pp. 85–90. ISSN: 1476-6256.

Rookey, Bryan D, Steve Hanway, and Don A Dillman (2008). "Does a probability-based household panel benefit from assignment to postal response as an alternative to Internet-only?" In: *Public Opinion Quarterly* 72.5, pp. 962–984. ISSN: 1537-5331.

Rosenthal, Robert (1979). "The file drawer problem and tolerance for null results". In: *Psychological bulletin* 86.3, p. 638. ISSN: 1939-1455.

*Sax, Linda J, Shannon K. Gilmartin, and Alyssa N. Bryant (2001). "Assessing response rates and nonresponse bias in web and paper surveys". In: *Research in higher education* 44.1, pp. 409–432.

Schwartz, Shalom H and Klaus Boehnke (2004). "Evaluating the structure of human values with confirmatory factor analysis". In: *Journal of research in personality* 38.3, pp. 230–255. ISSN: 0092-6566.

*Shannon, David M. and Carol C. Bradshaw (2002). "A comparison of response rate, response time, and costs of mail and electronic surveys". In: *The Journal of Experimental Education* 70.2, pp. 179–192. ISSN: 00220973, 19400683. URL: http://www.jstor.org/stable/20152675.

Silber, H., J. Daikeler, L. Weidner, and M. Bosnjak (2018). "Web survey". In: *Wiley StatsRef: Statistics Reference Online*. DOI: 10.1002/9781118445112.stat07984.

*Sinclair, Martha, Joanne O'Toole, Manori Malawaraarachchi, and Karin Leder (2012). "Comparison of response rates and cost-effectiveness for a community-based survey: Postal, internet and telephone modes with generic or personalised recruitment approaches". In: *BMC medical research methodology* 12.1, p. 132. ISSN: 1471-2288.

Statista (2018). In: URL: https://www.statista.com/topics/779/mobile-internet/.

*Al-Subaihi, Ali A (2008). "Comparison of web and telephone survey response rates in Saudi Arabia". In: *The Electronic Journal of Business Research Methods* 6.2, pp. 123–132.

Teo, Thompson SH, Vivien KG Lim, and Raye YC Lai (1999). "Intrinsic and extrinsic motivation in Internet usage". In: *Omega* 27.1, pp. 25–37.

Van Deursen, Alexander and Jan Van Dijk (2011). "Internet skills and the digital divide". In: *New media & society* 13.6, pp. 893–911.

Van Dijk, Jan AGM (2006). "Digital divide research, achievements and shortcomings". In: *Poetics* 34.4-5, pp. 221–235.

Vandenbroucke, Jan P (1998). "Observational research and evidence-based medicine: What should we teach young physicians?" In: *Journal of clinical epidemiology* 51.6, pp. 467–472. ISSN: 0895-4356.

*Vehovar, Vasja, Katja Lozar Manfreda, and Zenel Batagelj (2001). "Sensitivity of electronic commerce measurement to the survey instrument". In: *International Journal of Electronic Commerce* 6, pp. 31–51.

Viechtbauer, Wolfgang (2010). "Conducting meta-analyses in R with the metafor package". In: *Journal of Statistical Software* 36.3, pp. 1–48.

*Weible, Rick and John Wallace (1998). "Cyber research: The impact of the Internet on data collection". In: *Marketing Research* 10.3, pp. 18–31.

Williams, Douglas and J Michael Brick (2017). "Trends in US face-to-face household survey nonresponse and level of effort". In: *Journal of Survey Statistics and Methodology* 6.2, pp. 186–211. ISSN: 2325-0984.

*Wolfe, Edward W., Patrick D. Converse, and Frederick L. Oswald (2008). "Item-level nonresponse rates in an attitudinal survey of teachers delivered via mail and Web". In: *Journal of Computer-Mediated Communication* 14, pp. 35–66. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2008.01430.x.

*Woo, Youngje, Sunwoong Kim, and Mick P Couper (2015). "Comparing a cell phone survey and a web survey of university students". In: *Social Science Computer Review* 33.3, pp. 399–410. ISSN: 0894-4393. DOI: 10.1177/0894439314544876. URL: %3CGo%20to%20ISI%3E://WOS:000354306600007.

Wright, Kevin B (2005). "Researching Internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services". In: *Journal of computer-mediated communication* 10.3, JCMC1034.

*Yeager, David S, Jon A Krosnick, LinChiat Chang, Harold S Javitz, Matthew S Levendusky, Alberto Simpser, and Rui Wang (2011). "Comparing the accuracy of RDD telephone surveys and internet surveys conducted with probability and non-probability samples". In: *Public Opinion Quarterly* 75.4, pp. 709–747. DOI: 10.1093/poq/nfr020. URL: http://poq.oxfordjournals.org/content/75/4/709.full.pdf.

*Zuidgeest, Marloes, Michelle Hendriks, Laura Koopman, Peter Spreeuwenberg, and Jany Rademakers (2011). "A comparison of a postal survey and mixed-mode survey using a questionnaire on patients' experiences with breast care". In: *Journal of medical Internet research* 13.3, e68.

Chapter 5

How to Conduct Effective Interviewer Training: A Meta-Analysis

5.1 Abstract

Accepted for: Olson, Kristen, Jolene D. Smyth, Jennifer Dykema, Allyson Holbrook, Frauke Kreuter, and Brady T. West.(Eds.) Interviewer Effects from a Total Survey Error Perspective. CRC Press (Planned Publication Date: 2020)

Interviewer training can improve the performance of interviewers and thus also the quality of survey data. Yet, how effective interviewer training is in improving data quality and, more importantly, what factors drive its success are still largely unanswered questions. The present research uses meta-analytic methods to evaluate both the improvements in data quality due to interviewer training and the effectiveness of specific features of training in improving interviewer performance. The aspects of data quality considered are unit nonresponse, item nonresponse, correct administration of items, correct reading out of questions, and correct probing and recording of responses. Our meta-analysis of over 60 experimental studies revealed that comprehensive interviewer training improved all these factors by between five and 40 percentage points, and that using a broad variety of training methods and content, such as blended

learning, practice and feedback sessions, interviewer monitoring, and supplementary training materials, reinforced the positive effect of interviewer training on data quality.

5.2 Introduction

Interviewer training is an often overlooked factor in minimizing interviewer effects in interviewer-administered surveys (West and Blom 2017). In particular, the experimental variation of interviewer training and its content can provide information about the effectiveness of interviewer training and training methods. The present paper addresses these two aspects by using meta-analytic methods to summarize the results of interviewer training experiments.

Interviewers are one of the key parameters in the data collection process of interviewer-administered surveys (e.g. Singer, Frankel, and Glassman 1983; Groves et al. 2009). From a total survey error (TSE) perspective (e.g. O'Muircheartaigh and Campanelli 1998; West and Blom 2017), interviewers can influence four of the seven sources of survey error – namely, coverage, nonresponse, measurement, and processing error – and can even give rise to biased regression coefficients (e.g. Fischer et al. 2018). Furthermore, interviewer-administered surveys have often been found to produce over-reporting of socially desirable behavior and underreporting of socially undesirable behavior. This effect has been found to be even more pronounced when observable characteristics of the interviewer, such as gender or race, are related to the question content (e.g. Davis et al. 2010; Kreuter, Presser, and Tourangeau 2008, p.370). The literature has so far identified several other interviewer-related factors that have an impact on respondent reports, namely, interviewer experience (e.g. Hughes et al. 2002; Olson, Feng, and Witt 2008), interviewer expectations (e.g. Fowler and Mangione 1990; Olson, Kirchner, and Smyth 2016); and interviewer confidence and attitudes (e.g. Durrant et al. 2010; Mneimneh et al. 2018). The vast number of studies aimed at explaining, and thus reducing, interviewer effects through targeted study planning is therefore not surprising (O'Muircheartaigh and Campanelli 1998).

Kreuter, Presser, and Tourangeau (2008, pp. 371) identified four approaches that could be taken to reduce interviewer effects. The first approach relates to the choice of survey mode:

The largest interviewer effects are to be expected in face-to-face surveys compared to telephone surveys. However, the author warned against eliminating the interviewer altogether, as this "may introduce or increase other types of survey errors." Second, "if the biasing effects of an interaction among observable interviewer characteristics, question content, and respondent characteristics are well understood," interviewer effects could be reduced by deliberately matching interviewers and respondents. However, Kreuter, Presser, and Tourangeau (2008, pp. 371) pointed out that deliberate matching would not be feasible for most surveys, as respondent characteristics may not be known in advance. Hence, random assignment of interviewers to respondents was often recommended. Another approach to reducing interviewer effects proposed by Kreuter, Presser, and Tourangeau (2008, pp. 371) is supervising and monitoring interviewers in the field, reducing their workload, and altering the reward system to encourage them to focus on achieving not only a high number of completed cases but also high-quality data. And finally, Kreuter, Presser, and Tourangeau (2008, pp. 371) argued that interviewer training could reduce interviewer effects if interviewers learned to more systematically "explain the question-and-answer process to the respondent; motivate the respondent to provide high-quality answers; read questions exactly as worded; probe non-directively; and record answers without interpretation, paraphrasing, or additional inference about the respondent's opinion or behavior."

Ideal interviewer training should focus on two main areas of interviewer activity, namely, gaining respondents' cooperation and administering the question-and-answer process (Daikeler et al. 2017; Alcser et al. 2016). The importance of interviewer training for improving data quality has already been discussed in the literature (e.g. Lessler, Eyerman, and Wang 2008). Unfortunately, most survey programmes limit themselves to briefly describing their training concepts without questioning their effectiveness by means of experimental variation.

One reason why the effectiveness of general interviewer training is not questioned probably lies in the organization of fieldwork. Both large multinational survey programs, such as the Programme for the International Assessment of Adult Competencies (PIAAC; (OECD 2014)) and the European Social Survey (ESS; Loosveldt et al. (2014)), and small survey projects expect their fieldwork agencies to deploy interviewers who have undergone general interviewer

training. These interviewers are then given project-specific training to familiarize them with the features of the study in question, for example, the assessment of cognitive competencies in PIAAC OECD2014. Although the fieldwork agencies usually provide interviewers who have undergone general interviewer training, in many cases the type of general training they have had, and the effectiveness of this training, is a "black box".

Interviewer training has always been an integral part of the survey process, nevertheless the available literature on this subject is quite sparse. On the one hand, there is some research investigating the effect of interview training on specific data quality aspects such as unit non-response and correct probing (e.g. Fowler and Mangione 1990; Durand et al. 2006). On the other hand, there are suggestions and guidelines for interviewer training (e.g. Daikeler et al. 2017; Alcser et al. 2016). Yet, to my knowledge, only Lessler, Eyerman, and Wang (2008) have provided a comprehensive overview of the literature on interviewer training. However, as their overview was purely qualitative, it did not quantitatively evaluate the training concepts and results. The present paper aims to contribute to filling this gap in research by using meta-analytic methods to compare interview training experiments. The aim is to quantify the benefits of interviewer training and, more importantly, to determine what aspects of training (e.g., training length, use of blended learning, practice and feedback sessions) contribute to the reduction of interviewer effects.

The next section is devoted to the conceptual development of the research questions with reference to the literature. This is followed by a description of the meta-analytic methods used. The results section reports the impact of interviewer training on data quality and the training features that contributed most to these effects. The paper concludes with a discussion of the results and their implications for fieldwork.

5.3 Conceptual development of research questions

This section describes the theoretical background of interviewer training methods and formulates the questions to be answered by this meta-analysis. Classical interviewer training

consists of two parts – general and study-specific training (Daikeler et al. 2017; Loosveldt et al. 2014). The focus of the present paper is on general interviewer training, that is, the basic, cross-project part of interviewer training that aims to impart the knowledge and skills that a successful interviewer needs to achieve high data quality (see West and Blom 2017). However, as demonstrated by the total survey error (TSE) framework (Groves et al. 2009; Groves and Lyberg 2010), data quality can be compromised by several sources of error. The TSE framework has two sides—"measurement" and "representation"—both of which can be influenced by interviewers (see West and Blom 2017). "Measurement" comprises validity, measurement error and processing error; "representation" comprises coverage error, sampling error, nonresponse error, and adjustment error. On the measurement side, the literature shows that interviewers influence mainly measurement and processing error. On the representation side, interviewers have been found to have a particular impact on nonresponse error. However, depending on the survey design, they may also influence coverage error (West and Blom 2017).

Table 5.1 provides an overview of the different aspects of data quality that have been addressed in interviewer training experiments. It shows that, to date, the literature on experimental interviewer training has reported on the influence of interviewer training on measurement error, nonresponse error, and processing error.

The present study examines the impact of interviewer training on data quality. Specifically, it addresses six sources of error that compromise data quality: 1) unit nonresponse (nonresponse error); 2) item nonresponse (measurement error); 3) items that are incorrectly administered (measurement error and processing error); 4) items that are incorrectly read out (measurement error and processing error); 5) responses that are incorrectly probed (measurement error and processing error; and 6) responses that are incorrectly recorded (processing error). The aim is to determine whether these six sources of error are influenced by interviewer training and what training aspects contribute to the reduction of these errors, and thus to data quality. In the following, I first discuss nonresponse error and then address measurement error and processing error.

Table 5.1: Overview of the literature on interviewer tasks addressed in interviewer training experiments

Interviewer task	Survey error potentially introduced	Aspects already addressed with interviewer training experiments	References
Generate sampling frame	Coverage error	none	none
Make contact, gain cooperation, gain consent to additional parts of the survey	Unit nonresponse error	Response rate	(Basson and Chronister 2006; Dahlhamer et al. 2010; Cantor et al. 2004; Billiet and Loosveldt 1988; Mayer and O'Brien 2001; Schnell and Trappman 2006; Durand et al. 2006; Groves and McGonagle 2001)
Ask survey questions, conduct measurements and maintain motivation	Measurement error and item nonresponse error	Correctly administered, read and probed items, item nonresponse	(Guest 1954; Benson and Powell 2015; Dahlhamer et al. 2010; Billiet and Loosveldt 1988; Fowler Jr and Mangione 1986)
Record answers and measurements	Processing error	Correctly recorded items	(Fowler Jr and Mangione 1986)

5.3.1 Effect of refusal avoidance training on survey response rates

In their theoretical framework about survey participation, Groves and McGonagle (2001, pp.250-251) assert that two interviewer strategies – tailoring behavior to the perceived features of the sample person and maintaining interaction with the sample person – play a crucial role in gaining the cooperation of potential respondents. The authors posit that "maintaining interaction is the essential condition of tailoring, for the longer the conversation is in progress, the more cues the interviewer will be able to obtain from the householder" (p. 251). Moreover, they argue that the longer the interaction lasts, the harder it is for the sample unit to refuse to participate. Thus, the first research question to be answered by the present meta-analysis is:

Q1: Does general interviewer training that includes refusal avoidance training improve survey response rates compared with general interviewer training that does not include refusal avoidance training or with no interviewer training?

5.3.2 Effect of interviewer training on data quality

Especially in the case of measurement error and processing error, interviewers who are aware of interviewer effects and their consequences for data quality can react appropriately. Following Groves and Magilavy (1986, p.251) "interviewer variance or interviewer effects reflect the tendency for answers provided by the respondent and recorded in a questionnaire to vary depending on which interviewer is assigned to the respondent." A typical example is the tendency of respondents to report a higher income to a particular interviewer.

Reasons for interviewer effects include the activation of social norms by the interviewer's presence (Anderson, Silver, and Abramson 1988; Kane and Macaulay 1993; Bosnjak 2017) and systematic errors in administering the survey (e.g., failure to read questions as worded, directive probing, or failure to probe; (Fowler and Mangione 1990, pp. 265-266). Interviewer training alerts interviewers to the various causes of interviewer effects with the aim of preventing, or minimizing, them. Thus, the second question to be answered by the present meta-analysis is:

Q2: Are interviewer effects less pronounced if the interviewers undergo training beforehand?

5.3.3 Effect size heterogeneity

Unfortunately, interviewer training is not standardized or homogeneous in terms of duration, content, and training procedures, although initial efforts have been made in this direction in the following publications: the "General Interviewer Training Curriculum for Computer-Assisted Personal Interviews (GIT - CAPI)" (Daikeler et al. 2017); the "Guidelines for Best Practice in Cross-Cultural Surveys" (Survey Research Center 2016); and the brief interviewer training guidelines formulated by the American Association for Public Opinion Research (AAPOR 2016) and the International Organization for Standardization (ISO 2012). The low level of standardization of training content and methods gives rise to the following research question:

Q3: Are the effect size distributions heterogeneous?

Because of the lack of standardization, heterogeneous training outcomes, and thus effect size heterogeneity, can be expected. Heterogeneous distributions of effect size would imply that the success of interviewer training depends more on the content and methods of training than on the training itself. Accordingly, these factors must be examined more closely in order to be able to make statements on what constitutes successful training.

5.3.4 Training features that may improve data quality

In what follows, I first discuss optimal interviewer training duration and then address other interviewer training features that may improve data quality.

Interviewer training duration

Learning theory suggests that learning progress typically follows an S-shaped curve, starting slowly, accelerating, and then leveling off (Thorndike 1913). If the learning curve flattens out or becomes horizontal, learning progress stagnates. This phenomenon, which is referred to as a

learning plateau, occurs during the learning of complex skills (Thorndike 1913, pp. 99). Survey administration is an example of a complex skill. One aim of the present research is to determine the optimal duration of interviewer training to enable interviewers to learn the skills they need to avoid refusals and reduce interviewer effects. Hence, the fourth question to be answered by the meta-analysis is:

Q4: What is the optimal interviewer training duration to reduce (a) unit nonresponse and (b) the other error sources that affect data quality?

Interviewer training methods and determinants of effectiveness

According to Knowles, Holton, and Swanson (1984)' adult learning theory, one reason why adults learn differently than children is that they have accumulated a "reservoir of experience" that renders them "a rich resource for learning" (p. 45). Hence, "experiential techniques which tap the experience of the learners" are the most effective way of enabling adults to learn new skills (p. 46). Furthermore, individuals differ in their preferred learning style; some react more to visual information, some to auditory and others to kinesthetic information (Kelly 2010). Knowles posited that adults have "achieved a self-concept of essential self-direction" (p. 45) and engage in an educational activity because they are experiencing "some inadequacy in coping with current life problems" (p. 48). Therefore, they prefer self-directed, problem-centered learning[1].

Against this background, a flexible blended-learning approach to adult learning, which combines traditional face-to-face instruction with online learning, seems especially promising (Means et al. 2013). Blended learning combines the advantages of online learning, such as flexibility in terms of time and place, with those of face-to-face instruction, such as direct interaction with trainers and other trainees and live feedback.

As the literature on the effects of interviewer training on data quality shows considerable differences in interviewer performance with regard to nonresponse- and measurement-related data

[1]For a comprehensive overview of the literature on adult learning theory, see Tusting and Barton (2003).

quality (West, Kreuter, and Jaenichen 2013; West and Blom 2017), the effects of interviewer training on nonresponse error and on measurement error are also considered separately in the final research question:

Q5: Are cooperation rates and interviewers' survey administration skills improved by (a) practice and feedback sessions (vs. no practice and feedback sessions); (b) interviewer monitoring (vs. no interviewer monitoring); (c) supplementary written training material (vs. no supplementary training material); (d) listening to audio refusals (vs. not listening to audio refusals); (e) blended learning (vs. an unimodal approach), and (f) previous interviewing experience (vs. no previous interviewing experience)?

5.4 Data and Methods

This section describes the five steps of the meta-analytic procedure employed in the present study: 1) a comprehensive literature search; 2) checking of the eligibility of studies found; 3) coding of relevant data; 4) calculation of training effect sizes; 5) analysis of variables that moderate effect size (Lipsey and Wilson 2001; Borenstein et al. 2009).

5.4.1 Eligibility criteria and search strategy

One of the first steps in a meta-analysis is the definition of the criteria that studies must meet if they are to be included. Table 5.2 lists these eligibility criteria.

To ensure the quality of the meta-analysis, a comprehensive literature search was conducted. Because a meta-analysis that includes only published literature faces the problem of publication bias, grey literature was also eligible for inclusion (for further information, see the appendix section 5.7.1). During the search process, the most common reasons for the exclusion of studies were the lack of an experimental design and missing data quality indicators. Most of the studies rated the use of interviewer training as appropriate but did not evaluate how effective it was.

Table 5.2: Eligibility criteria

Eligibility Criterion	**Description**
Experimental design	Studies must employ an experimental design. We accepted both treatments versus control and pre-versus post group designs. In the first case, a group of trained interviewers is compared with a group of less trained or untrained interviewers, while in the pre-versus post-design group the experiment has up to four steps. First, the interviewers receive no or only elementary training, in the second step the data quality is measured, then the interviewers receive professional training, and in the fourth step, the data quality is measured again.
Downgraded training for control group	For both types of training, it was essential that the control group received either no or only an introductory briefing. This briefing should not have lasted longer than one hour.
Data quality measures	Data quality measures indicating the effectiveness of training are mandatory.
Training content on refusal avoidance and/ or measurement – related data quality	The interview tasks can be divided into two main areas. First, to encourage respondents to participate (nonresponse errors) and second, to achieve adequate data quality during the interview (measurement and processing errors). Therefore, the last selection criterion differs according to the interviewer's task and the measured data quality indicator. For the first task, the avoidance of refusals, we include studies with a classical refusal avoidance training (see Groves and McGonagle 2001). For the second task to improve data quality indicators in the survey process, data quality and interviewer behavior had to be an essential part of the training.

The PRISMA diagram (Moher et al. 2009) in figure 5.1 gives an overview of the search strategy. The search was limited to literature in English; over 2,000 results had to be excluded because the broad search terms led to literature related to job interviews, linguistic interviews, cognitive and clinical interviews of victims and witnesses, and studies without an experimental setting. Nineteen eligible publications were retrieved. Because many of the publications presented more than one experiment or effect size, the search yielded a total of 66 experimental comparisons. The most common indicator of data quality was the effect of interviewer training on the response rate (22), followed by the effect on correct recording of the response (14); on item nonresponse (12); on the reading of questions exactly as worded (12); on correct probing (6); and on correct item administration (4).

5.4.2 Coding procedure

Coding was performed by two independent coders (the coding scheme can be found in appendix Table 5.6). The lead coder coded all studies and instructed the second coder, who coded 30 percent of the studies. Intercoder reliability produced a Krippendorff's alpha (Krippendorff 2004) of .9 for the effect sizes and .95 for the moderator variables, indicating a match of at least 90 percent between the two coders. Reliability values of .8 and above indicate an almost perfect match (Hallgren 2012). Consequently, it can practically be ruled out that the effect sizes and moderator codings on which this meta-analysis is based were subjectively distorted by the coders.

5.4.3 Effect size metric and statistical method

During the search process, it became clear that interviewer training experiments report a variety of different data quality indicators as effect size metrics. From a methodological point of view, most of these data quality indicators are not substantively comparable, which is why it was decided to conduct a separate meta-analysis for each indicator (an overview provides Table 5.3). As the effect size metric was the same for all six data quality indicators, the effect sizes

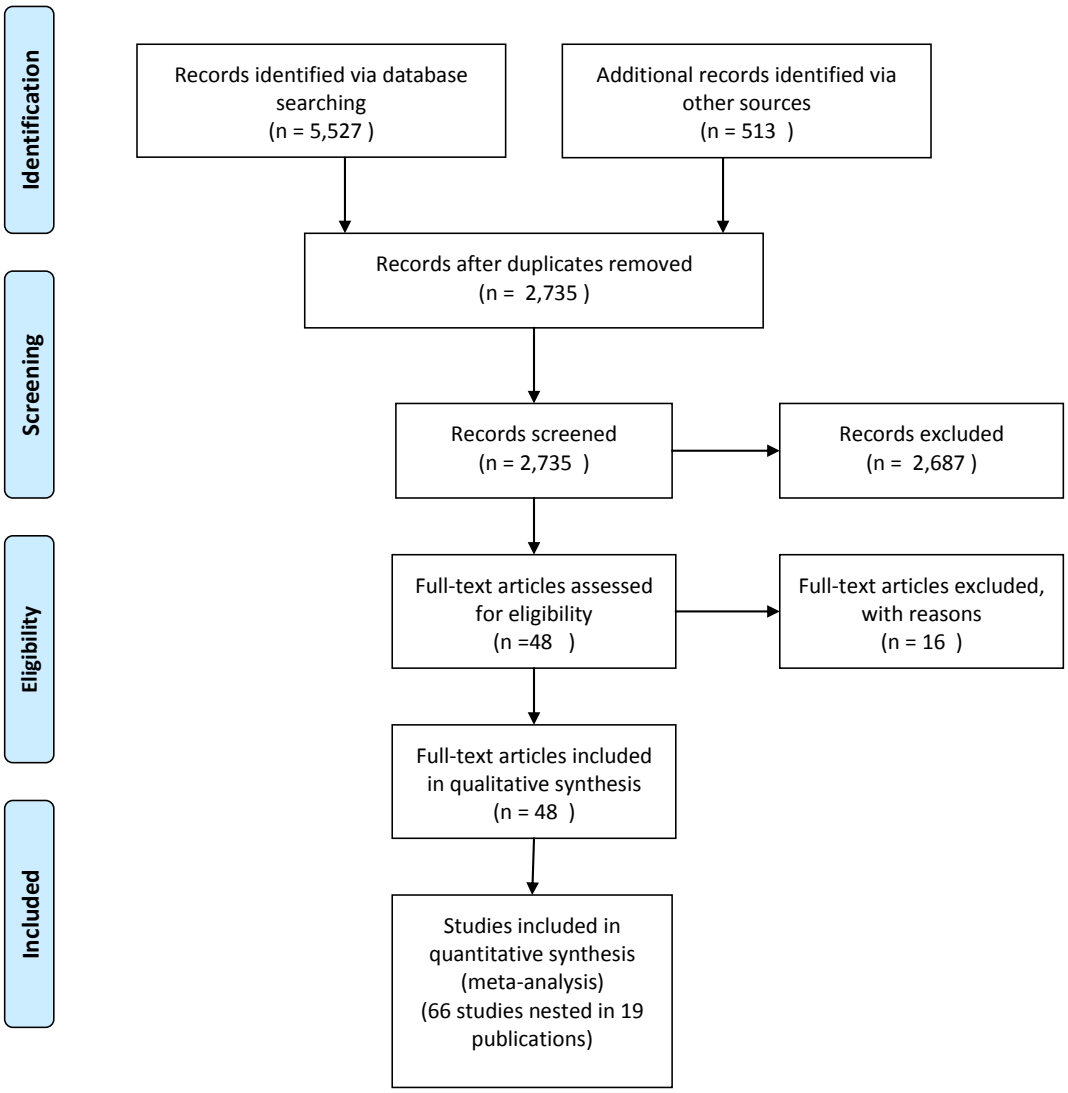


Figure 5.1: The literature search process

were calculated as follows (e.g., for correctly administered items):

$$RD = \frac{N_{cait}}{N_{ait}} - \frac{N_{caiu}}{N_{aiu}}$$

with RD = Rate Difference ,

N_{cait} *= Total Number Of Correctly Administered Items For Trained Group,*

N_{ait} *= Total Number Of Items For Trained Group,*

N_{caiu}*= Total Number Of Correctly Administered Items For Untrained Group,*

N_{aiu} *= Total Number Of Items For Untrained Group*

The statistical analysis for each of the six data quality indicators comprised five steps (Lipsey and Wilson 2001). First, the weighted mean response rate difference across all studies was computed. This variance component consisted of the study-level sampling error variance as well as an estimate of between-study variance (Borenstein et al. 2009). A random-effects analysis was used, as inference should be made for a population of studies larger than the set of observed studies (Hedges and Vevea 1998). In the next step, the confidence interval for the mean effect size was determined to indicate the degree of precision of the estimate and whether the mean effect size was statistically significant. In the third step, a homogeneity analysis was performed to assess whether the effect sizes came from the same population (random effects assumption). In the fourth step, the robustness and quality of the findings were checked with an outlier analysis and publication bias checks. In the final step, a mixed-effect model analysis was performed for each moderator variable to determine which variables had a significant influence on the response rate differences. Studies that did not provide information on moderator variables were excluded from the respective analyses. The R package "metafor" (version 1.9-9) was used for the analyses (Viechtbauer 2010).

5.4.4 Publication bias and sensitivity analyses

In the next step, we examined whether a publication bias might have affected the estimates of the mean effect size. To this end, we checked both the funnel plots and the Egger's regression

Table 5.3: Description of effect sizes

Response Rate	Experimental interviewer group received Refusal-Avoidance-Training (RAT) and control group did not, invited vs. participated respondents in each group
Item Nonresponse	Experimental interviewer group received advanced interviewer training and control group not, counting item nonresponse in each group
Administering	Experimental interviewer group received advanced interviewer training and control group not, counting correctly administered questions per interview (audio tape error index)
Probing	Experimental interviewer group received advanced interviewer training and control group not, counting correctly probed questions per interview (audio tape)
Reading	Experimental interviewer group received advanced interviewer training and control group not, counting correctly read questions per interview (audio tape)
Recording	Experimental interviewer group received advanced interviewer training and control group not, counting correctly recorded questions per interview (audio tape)

tests (see appendix Figure 5.4 and Table 5.5) and found that a publication bias problem existed, as a disproportionate number of significant results had been included in the meta-analyses. One reason for this may have been the generally insufficient number of studies in this area. Outlier tests were conducted in the sensitivity analysis. For each model, 10 percent of outlier studies were excluded, and no significant difference between the original and outlier-adjusted effect sizes was found.

5.5 Results

In this section, the overall effect of interviewer training on response rates (Q1) and on the remaining data quality indicators (Q2) is reported. Then, the question of whether the effects size distributions were heterogeneous (Q3) is addressed. And finally, the impact of each training feature on interviewer training success (Q4 and Q5) is reported.

5.5.1 What is the effect of interviewer training on data quality? (Q1– Q3)

In the following section, the various data quality indicators are discussed in detail, and the extent to which they were improved through interviewer training is reported.

Q1: Effect of refusal avoidance training on response rates

Figure 5.2 shows a forest plot summarizing the study-level differences in response rates between trained and untrained interviewers. On the x-axis the differences in data quality between trained and untrained interviewers are presented. Positive values mean better data quality for trained interviewers, and all effect sizes that do not cross the zero line are significantly different from zero. On the y-axis, all included studies, their effect sizes, and confidence intervals (CIs) are listed one by one. The last line of each quality measure shows the sampling error weighted mean effect size under the random effects assumption. The effect size distribution in the forest plot indicates that most response rate comparisons show that trained interviewers achieved higher response rates than untrained interviewers. Surprisingly, there were quite a few zero findings. The sampling error weighted mean effect size estimate, calculated across all 22 effect sizes assuming random effects, was 0.05 (95% CI = 0.00/0.11). This result shows that the response rates achieved by trained interviewers were, on average, five percentage points higher than those achieved by untrained interviewers. Our first research question (Q1) can therefore be answered in the affirmative. However, the small magnitude of improvement is surprising and indicates that interviewer training has quite a minor impact on respondent participation rates.

Q2: Influence of interviewer training on data quality

Taking into account the other data quality indicators (item nonresponse, correct administration of the items; reading out questions correctly; probing responses correctly; recording responses

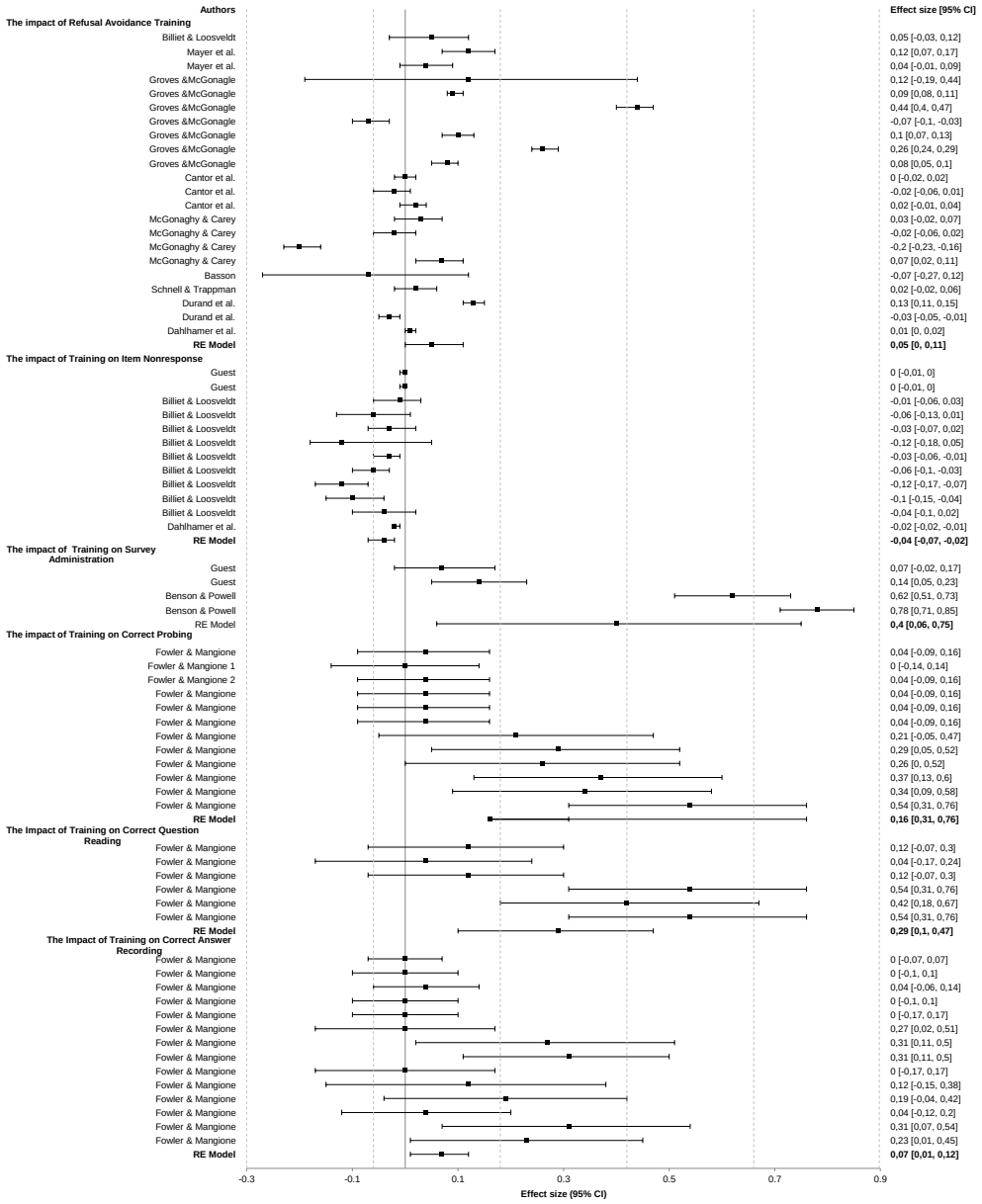


Figure 5.2: Forest plots for data quality indicators: Trained vs. untrained interviewers

correctly), our results confirm that trained interviewers achieved significantly higher data quality than untrained interviewers (see Figure 5.2). The average effect sizes were always in the expected direction – on the right-hand positive side of the zero line. In particular, we found that, in the case of trained interviewers compared to untrained interviewers, the item nonresponse rates were significantly lower (4%; 95% CI = 0.07/0.02); 40 percentage points more items were correctly administered (95% CI = 0.06/ 0.75); 29 percentage points more questions were read out correctly (95% CI = 0.10/0.47); 16 percentage points more responses were probed correctly (95% CI = 0.06/0.25); and seven percentage points more responses were recorded correctly (95% CI = 0.01/0.12).

It should be pointed out that, due to the small number of studies and to possible distortions by the authors, the overall picture conveyed by these results should be considered rather than looking at the data quality indicators individually. However, this overall picture is quite clear: It shows that interviewer training significantly improves both unit nonresponse and the other data quality indicators. Thus, Q1 and Q2 can be answered in the affirmative.

Q3: Effect Size Heterogeneity

The continued absence of standardized interviewer training, and the resulting heterogeneity of training approaches, prompted us to ask whether effect size distributions were heterogeneous (Q3), which would result in further moderator analysis. This question can be answered in the affirmative: Our analyses revealed a heterogeneous effect size distribution ($p \leq .05$) assuming random effects for all six data quality indicators (see appendix Table 5.7). To examine whether – and, if so, which – interviewer training features influenced the effect of interviewer training, we conducted a moderator analysis.

5.5.2 Moderator analysis: Which features render interviewer training successful? (Q4 and Q5)

In this section, we present the results for the moderator variables. We report these results for three of the six data quality indicators (response rates, item nonresponse and correct item administration), as eligible studies with a variation on the moderator variables could be identified only for these three quality indicators. Specifically, we were interested in whether duration of interviewer training (Q4), practice and feedback sessions, additional supplementary training material, interviewer monitoring, blended-learning-based training, and previous interviewing experience (Q5) had an impact on the training outcomes. In what follows, we discuss the results for each of the aforementioned data quality indicators.

Q4: What is the optimal interviewer training duration?

a. Reduction of unit nonresponse. The duration of interviewer training was found to have only a small impact on response rates. On average, the response rates achieved by interviewers with an average training duration of five to 10 hours were seven percentage points higher than those achieved by untrained interviewers (see Figure 5.3).

The response rates achieved by interviewers who received only one to five hours of training were, on average, four percentage points higher than those achieved by untrained interviewers, and the response rates achieved by interviewers who received 10 hours of training or more were, on average, six percentage points higher than those achieved by untrained interviewers. On average, the response-rate gap between a three-hour refusal avoidance training and a 7.5-hour refusal avoidance training was only three percentage points.

b. Item nonresponse and correct survey administration. For measurement-error-related interviewer tasks, such as preventing item nonresponse, our data suggest a minimum training duration of 11 hours (see Figure 5.3). As our data lacked studies that focused on the administration of items, testing of the effect of training duration on this parameter was not possible. The recommended training duration can be regarded only as an estimate of the importance of

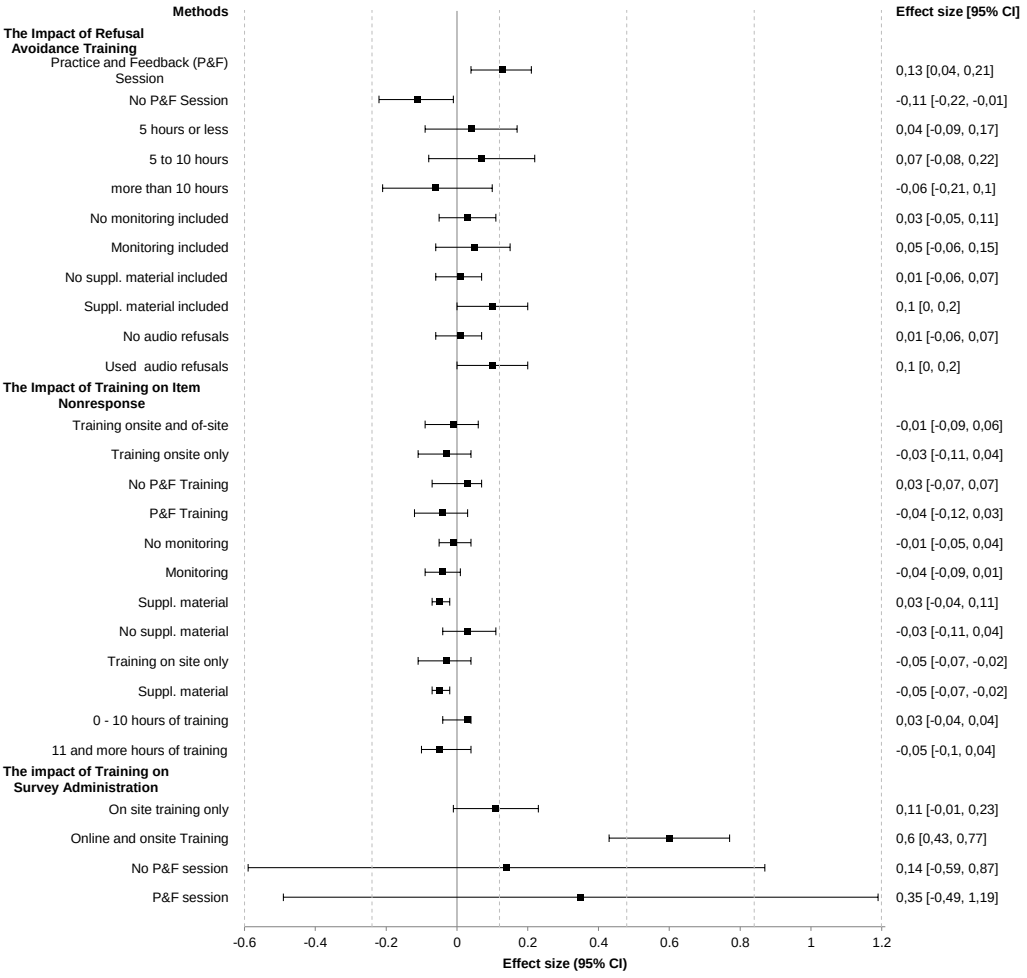


Figure 5.3: Forest plots for data quality indicators: Trained vs. untrained interviewers

the focus of the training content. Unfortunately, only limited practical recommendations, such as coverage of new content, more detailed treatment of existing content, and more practice and feedback sessions can be inferred from our moderator analysis. There is no doubt that there is a lack of primary studies furnishing empirical evidence on training content from which recommendations for practitioners could be derived. Nevertheless, with regard to our fourth research question (Q4), we can conclude from these findings that between five and 10 hours should be devoted to to refusal avoidance training and 11 hours or more to training aimed at reducing other error sources that compromise data quality.

Q5: Which training methods work best?

Not only training duration but also, and especially, training methods are determining factors for the success of training. Our analysis revealed that, when interviewer training included practice and feedback sessions, the response rates achieved by trained interviewers were, on average, 13 percentage points higher than those achieved by the control group (see forest plot in Figure 5.3). Our analysis also revealed better response rates for interviewer training that used interviewer monitoring versus training that did not. The use of supplementary training manuals improved response rates by 10 percentage points compared with training that did not provide supplementary material. Furthermore, training was more effective for interviewers who had no previous interviewing experience. Training courses that included blended learning (combined online and onsite training) resulted in more correctly administered items than did purely onsite training. Besides blended learning, previous interviewing experience was the only variable that had an impact on correct item administration.

Table 5.4 provides an overview of the training features that influenced specific data quality indicators. We could not identify one specific training feature that affected all data quality indicators. It is noteworthy that each data quality indicator seems to have been influenced by different training features. Unit nonresponse was influenced by training duration and by practice and feedback sessions; for item nonresponse, the duration of training and interviewer monitoring were salient; correct item administration was influenced by blended learning and

Table 5.4: Moderator overview

+ = tested & significant ($p \leq .005$); x = tested & not significant but expected direction; - = tested & not significant and not in expected direction; o = not tested - no variation

Moderator/ Indicator	Training Lengths	Practice & Feedback	Monitoring	Suppl. Material	Blended Learning	Prior Interviewer Experience
Unit-Nonresponse	+	+	x	x	o	x
Item Administration	o	x	x	o	+	+
Item Nonresponse	+	x	+	x	x	x
Probing	x	-	o	-	o	o
Recording	x	-	o	-	o	o

previous interviewing experience. Therefore, to achieve sufficient data quality with several data quality measures, a mix of interactive training methods such as practice and feedback sessions, interviewer monitoring, blended learning, and supplementary material is recommended.

5.6 Conclusion and discussion

Although the training of interviewers makes an essential contribution to data quality, it has too often been an overlooked parameter in survey research. The aim of the present study was to answer the question as to the impact of interviewer training on data quality and the training features that are most promising in this regard. The results of my meta-analysis of 66 experimental studies provide empirical evidence that interviewer training improves data quality by up to 40 percentage points. As the moderator analyses show, I could not identify one specific training feature that affected all data quality indicators. Moreover, I found that different training features, for example, practice and feedback sessions and blended learning approaches, significantly improved data quality in terms of better response rates and more correct item administration. This shows that not only strongly application-oriented learning content, such as practice and feedback sessions (Knowles, Holton, and Swanson 1984), but also a

diverse training strategy consisting of interviewer monitoring, blended learning, supplementary materials, and audio examples, is most effective. With regard to optimal training duration, my findings suggest that five to 10 hours would be the optimal duration for refusal avoidance training, and that at least 11 hours should be devoted to the remaining training content.

At least four implications for fieldwork can be concluded from my results. First, training should not focus primarily on persuading reluctant respondents. The evaluation of the training duration in this paper can be seen as an estimate of the importance of the focus of the training content. What was especially surprising was the albeit significant but low improvement in the response rate as a result of interviewer training. This finding suggests that there are only a few trainable skills that influence the recruitment of respondents. It would appear that it is not such much a particular skill on the part of the interviewer that influences the respondent's decision to participate but rather the interaction between the characteristics of the interviewer and those of the potential interviewee (Groves, Cialdini, and Couper 1992; Jäckle et al. 2013; Olson, Kirchner, and Smyth 2016). In addition, gaining the cooperation of nonrespondents does not necessarily lead to lower nonresponse bias, but rather it may even result in increased measurement bias (West and Olson 2010; West et al. 2018; Fischer et al. 2018). In conclusion, recruitment strategies should constitute a substantial – but not the main – focus of interviewer training.

Second, interviewer training should continue to focus on the correct administration of the question-and-answer process, as my findings suggest that considerable data quality improvements can be achieved through training in this task. This finding is also in line with studies that have found that interviewers have a substantial impact on measurement bias (Fischer et al. 2018; West et al. 2018). Third, my results show that training content can best be conveyed by using a wide variety of methods. In particular, practice and feedback sessions should be included in the training program, as adults learn primarily from experience. Finally, another finding of this paper is that already experienced interviewers should participate in regular training, as the quality of the data they collect also benefits from re-training.

The present study has a number of limitations. The first relates to its scope. Researchers

may also be interested in the effects of interviewer training on data quality indicators other than the six tested here. The impact of specific training methods and content on interviewer class correlation coefficients, the bias of estimators, the collection of sensitive information, the collection of biomarkers, and the achievement of high consent rates are questions that remain largely unanswered. However, they must first be addressed by primary research before evidence-based meta-analytic work is possible.

Another limitation of this study is that I could not include in the moderator analysis the effect sizes of three determinants of data quality, namely correctly reading out questions, probing responses, and recording responses. As these effect sizes were either not reported at all or were only occasionally reported in the studies included in my meta-analysis, I could not empirically assess the influence of interviewer training on these data quality indicators. However, mine is the first meta-analysis in this field to provide recommendations for the remaining three effect sizes, unit nonresponse, item nonresponse, and correct item administration.

A third limitation of my study is that the gaps in primary research rendered it necessary to conduct six separate meta-analyses, each with a limited number of studies, which gave rise to statistical performance problems. Nevertheless, all my results point in the same direction and contribute to a consistent overall picture—namely, that proper interviewer training consists of a combination of different training methods and should address nonresponse- and measurement-related data quality aspects.

Both the lack of statistical performance and the lack of variation on some moderators were caused by the small number of primary research studies. Therefore, in order to increase transparency in interviewer training, I strongly encourage researchers to conduct further experimental primary research on training methods, and especially on training content. The aim should be to develop a generally accepted gold standard for evidence-based interviewer training that offers further implications for fieldwork.

Such an interviewer training gold standard should address both measurement- and representation-related content. However, the focus should be on the prevention of measurement errors. Training content should include item nonresponse, questionnaire administration, correct probing,

and correct recording of responses. This content should be taught using a broad mix of methods that address different types of learners. What is particularly important for adult education is learning that is based on practical experience in this field and that taps the general life experience of the learner. Therefore, practice and feedback sessions are especially appropriate. Free time management using blended learning approaches also has a positive effect on training success. In their interviewer training manuals, Daikeler et al. (2017) and Alcser et al. (2016) propose a module-based training structure and special modules for already experienced interviewers, which is in line with the findings of this meta-analysis.

Following West and Blom (2017), who emphasized the importance of interviewer training, behavior, and skills, the present work has demonstrated how training can effectively enhance interviewer skills. In practice, interviewer training and monitoring is often outsourced to field institutes and is therefore difficult to influence. Nevertheless, there are ways and means of doing so, for instance, by including the type and scope of interviewer training in calls for tenders, interviewer certification systems, and in-house training guidelines. The use of training methods based on blended learning opens up new possibilities to create professionally developed training materials at lower costs. Further potential for better data quality undoubtedly lies in (mobile) interviewer monitoring and dashboard systems with the option of (re)training specific skills. On a final note, this paper hopes to encourage researchers to critically question interviewer training and, if necessary, adapt it to current research.

5.7 Appendix

5.7.1 Publication bias

Publication bias exists if the preparation, submission or publication of research findings depend on characteristics of just these research results, e. g. their direction or statistical significance. Publishing only results that show a significant finding disturbs the balance of findings (Weiss and Wagner 2011). We used three techniques to overcome this problem. First, we examined conference abstracts (American Association for Public Opinion Research (AAPOR), European Survey Research Association (ESRA), Joint Statistical Meeting (JSM)), second we used the reference lists of the already located manuscripts and applied a snowballing technique and the last strategy was to ask for appropriate research via mailing lists and email. We followed conference presentations and papers with restricted access by email and asked in this regard for similar research.

The funnel plots in Figure 5.4 are a visual method used to inspect publication biases (Egger et al. 1997). It shows the individual observed effect sizes on the x-axis against the corresponding standard errors. It is important that the point cloud on both sides of the line is approximately equal in number and distribution, which is not for all of our effect sizes the case. These results are emphasized by the Egger's regression test, which tests the asymmetry of the funnel plot (see Appendix table 5.5).

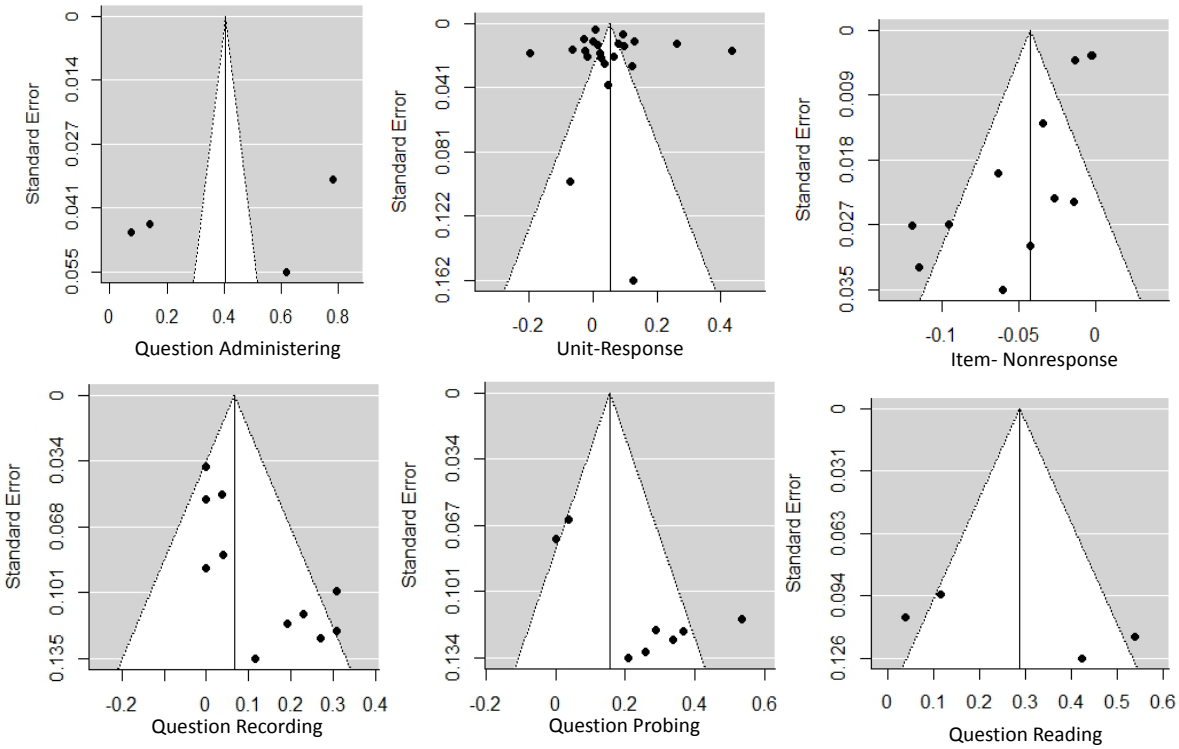


Figure 5.4: Publication bias: Funnel plots for data quality indicators

Table 5.5: Publication bias check: Egger's regression test

Effect Size Measure	**Regression Test for Funnel Plot Asymmetry**
Response Rate	0.5113
Administration	0.5111
Item Nonresponse	0.0005
Question Reading	0.0591
Probing	0.0003
Recording	0.0012

5.7.2 Coding

Table 5.6: Coding scheme

Variable	Scale/Categories
Case Number	string
Authors	string
Reference	string
Title	string
Year	continous
Published	2 - Yes/ 1 - No
Experiment Number (If study has more than one)	continous
Identifier	string
Invited in treated Group	continous
Participated in treated Group	continous
Number of Interviewers in treated Group	continous
Number of Interviews in treated Group	continous
Invited in untreated Group	continous
Participated in untreated Group	continous
Number of Interviewers in untreated Group	continous
Number of Interviews in untreated Group	continous
Pre/ Post or Control/Trearment	2 - Control/ Treatment 1 - Pre/ Post
Control group had also a basic training	2 - Yes/ 1 - No
Listened to audio refusals	2 - Yes/ 1 - No
Prior Experienced interviewers	2 - Yes/ 1 - No
Lenght of Training in hours	continous
Using supplementary Training material	2 - Yes/ 1 - No
Monitoring	2 - Yes/ 1 - No
Practice & Feedback Sessions included	2 - Yes/ 1 - No
Training for Telphone Interviewers only	2 - Yes/ 1 - No
Training for Face to Face Interviewers only	2 - Yes/ 1 - No
Includes Blended Learning	2 - Yes/ 1 - No
Training for all modes	2 - Yes/ 1 - No
Refusal Avoidance Training Only	2 - Yes/ 1 - No

5.7.3 Random effects model and meta regression summary statistics

Table 5.7: Sampling error weighted mean effect sizes and heterogeneity

Meta-analytic Summary Statistics (random effect model)				**Heterogeneity Estimators**		
Data Quality Indicator	**k**	**Mean Response Difference (95 % CI)**	**T (se)**	**Q_e total (df/ p)**	**I**	**H**
Response Rate	22	0.053 (-0.008/ 0.1069)	0.0155 (0.0051)	1355.9482 (21/0.0001)	98.96%	96.49%
Item Nonresponse	12	-0.0427 (-0.0658/-0.0196)	0.0012 (0.0007)	63.1317 (11/0.0001)	95.20%	20.82%
Correct Question Recording	14	0.0658 (0.0138/0.1181)	0.0039 (0.0036)	23.5360 (13/0.0357)	43.89%	1.78%
Correct Question Probing	12	0.1557 (0.0604/0.2510)	0.0195 (0.0119)	33.7251 (11/0.0004)	73.69%	3.80%
Correct Question Reading	6	0.287 (0.1029/0.4711)	0.0413 (0.0335)	22.7093 (5/0.0004)	78.51%	4.65%
Correct Question Administration	4	0.4047 (0.0614/0.748)	0.1206 (0.1002)	212.4658 (3/0.0001)	98.40%	62.35%

References

AAPOR (2016). *Standard definitions: Final dispositions of case codes and outcome rates for surveys*. 9th. Oakbrook Terrace, IL: American Association for Public Opinion Research.

Alcser, Kirsten, Judi Clemens, Lisa Holland, Heidi Guyer, and Mengyao Hu (2016). "Interviewer recruitment, selection, and training". In: *Guidelines for best practice in cross-cultural surveys. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor*.

Anderson, Barbara A, Brian D Silver, and Paul R Abramson (1988). "The effects of the race of the interviewer on race-related attitudes of black respondents in SRC/CPS national election studies". In: *Public Opinion Quarterly* 52.3, pp. 289–324.

*Basson, Danna and Michael Chronister (2006). "Recordings of prior refusals: Do they improve later conversion attempts?" In: *Methodology of Longitudinal Surveys conference, Essex, England*.

*Benson, Mairi S. and Martine B. Powell (2015). "Evaluation of a comprehensive interactive training system for investigative interviewers of children". In: *Psychology, Public Policy, and Law* 21.3, pp. 309–322. DOI: 10.1037/law0000052.

*Billiet, Jacques and Geert Loosveldt (1988). "Improvement of the quality of responses to factual survey questions by interviewer training". In: *Public Opinion Quarterly* 52.2, pp. 190–211. ISSN: 0033-362X.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein (2009). *Introduction to meta-analysis*. John Wiley and Sons. 457 pp. ISBN: 9780470057247.

Bosnjak, Michael (2017). "Mixed-mode surveys and data quality". In: *Methodische Probleme von Mixed-Mode-Ansätzen in der Umfrageforschung*. Ed. by Stefanie Eifler and Frank Faulbaum. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 11–25. ISBN: 978-3-658-15834-7. DOI: 10.1007/978-3-658-15834-7_1.

*Cantor, David, Bruce Allen, Sid J Schneider, Tracey Hagerty-Heller, and Angela Yuan (2004). "Testing an automated refusal avoidance training methodology". In: *annual meeting of the American Association for Public Opinion Research, Phoenix, AZ*.

*Dahlhamer, James M, Marcie L Cynamon, Jane F Gentleman, Andrea L Piani, and Michael J Weiler (2010). *Minimizing survey rrror through interviewer training: New procedures applied to the National Health Interview Survey (NHIS).*

Daikeler, Jessica, Henning Silber, Michael Bosnjak, Anouk Zabal, and Silke Martin (2017). "A general interviewer training curriculum for computer-assisted personal interviews". In: *GESIS Survey Guidelines* Version 1. DOI: 10.15465/gesis-sg_en_022.

Davis, Rachel E, Mick P Couper, Nancy K Janz, Cleopatra H Caldwell, and Ken Resnicow (2010). "Interviewer effects in public health surveys". In: *Health Education Research* 25.1, pp. 14–26. ISSN: 0268-1153. DOI: her/cyp046.

*Durand, Claire, Marie-Eve Gagnon, Christine Doucet, and Eric Lacourse (2006). "An inquiry into the efficacy of a complementary training session for telephone survey interviewers". In: *Bulletin de Méthodologie Sociologique* 92.1, pp. 5–27. ISSN: 0759-1063.

Durrant, Gabriele B, Robert M Groves, Laura Staetsky, and Fiona Steele (2010). "Effects of interviewer attitudes and behaviors on refusal in household surveys". In: *Public Opinion Quarterly* 74.1, pp. 1–36. ISSN: 0033-362X. DOI: 10.1093/poq/nfp098. URL: %5Curl%7Bhttp://poq.oxfordjournals.org/content/74/1/1.full.pdf+html%7D.

Egger, Matthias, George Davey Smith, Martin Schneider, and Christoph Minder (1997). "Bias in meta-analysis detected by a simple, graphical test". In: *Bmj* 315.7109, pp. 629–634. ISSN: 0959-8138.

Fischer, Micha, Brady T West, Michael R Elliott, and Frauke Kreuter (2018). "The impact of interviewer effects on regression coefficients". In: DOI: 10.1093/jssam/smy007. eprint: http://oup.prod.sis.lan/jssam/advance-article-pdf/doi/10.1093/jssam/smy007/24801517/smy007.pdf. URL: https://doi.org/10.1093/jssam/smy007.

*Fowler Jr, Floyd J. and Thomas W. Mangione (1986). *Reducing interviewer effects on health survey data. Executive summary.* Report. National Center for Health Services Research and Health Care Technology.

Fowler, Floyd J. Jr. and Thomas W. Mangione (1990). *Standardized survey interviewing. Minimizing interviewer-related error.* Vol. 18. Applied Social Research Methods Series. Newbury Park, CA: Sage Publications.

Groves, Robert M, Robert B Cialdini, and Mick P Couper (1992). "Understanding the decision to participate in a survey". In: *Public Opinion Quarterly* 56.4, pp. 475–495. ISSN: 0033-362X.

Groves, Robert M., Floyd J. Jr. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau (2009). *Survey methodology*. Vol. 2. Hoboken, NJ: John Wiley & Sons.

Groves, Robert M. and Lars Lyberg (2010). "Total survey error: Past, present, and future". In: *Public Opinion Quarterly* 74.5, pp. 849–879. DOI: 10.1093/poq/nfq065.

Groves, Robert M and Lou J Magilavy (1986). "Measuring and explaining interviewer effects in centralized telephone surveys". In: *Public opinion quarterly* 50.2, pp. 251–266.

*Groves, Robert M and Katherine A McGonagle (2001). "A theory-guided interviewer training protocol regarding survey participation". In: *Journal of Official Statistics* 17.2, pp. 249–265. ISSN: 0282-423X.

*Guest (1954). "A new training method for opinion interviewers". In: *Public Opinion Quarterly* 18.3, pp. 287–299.

Hallgren, Kevin A (2012). "Computing inter-rater reliability for observational data: an overview and tutorial". In: *Tutorials in quantitative methods for psychology* 8.1, p. 23.

Hedges, Larry V and Jack L Vevea (1998). "Fixed-and random-effects models in meta-analysis". In: *Psychological methods* 3.4, p. 486. ISSN: 1939-1463.

Hughes, A., James Chromy, K. Giacoletti, and D. Odom (2002). "Impact of interviewer experience on respondent reports of substance use". In: *Redesigning an ongoing national household survey*. Ed. by Joseph Gfroerer, Joe Eyerman, and James Chromy. Washington: Substance Abuse and Mental Health Services Administration, pp. 161–184.

ISO, norm (2012). "Market, opinion and social research — Vocabulary and service requirements". In: *ISO standards 20252:2012(E)*.

Jäckle, Annette, Peter Lynn, Jennifer Sinibaldi, and Sarah Tipping (2013). "The effect of interviewer experience, attitudes, personality and skills on respondent co-operation with face-to-face surveys". In: *Survey Research Methods* 7.1, pp. 1–15. ISSN: 1864-3361.

Kane, Emily W. and Laura J. Macaulay (1993). "Interviewer gender and gender attitudes". In: *Public Opinion Quarterly* 57.1, pp. 1–28. ISSN: 0033362X.

Kelly, M (2010). "Learning styles-Understanding and using learning styles". In: *Retrieved February* 25.

Knowles, Malcolm S, Elwood F Holton, and Richard A Swanson (1984). *The adult learner*. Routledge.

Kreuter, Frauke, Stanley Presser, and Roger Tourangeau (2008). "Social desirability bias in CATI, IVR, and web surveys". In: *Public Opinion Quarterly* 72, pp. 847–865.

Krippendorff, Klaus (2004). "Reliability in content analysis: Some common misconceptions and recommendations". In: *Human Communication Research* 30.3, pp. 411–433. URL: http://d%20x.doi.org/10.1111/j.1468-2958.2004.tb00738.x.

Lessler, Judith T., Joe Eyerman, and Kevin Wang (2008). "Interviewer training". In: *International handbook of survey methodology*. Ed. by Edith D. De Leeuw, Joop J. Hox, and Don A. Dillman. New York, NY: Taylor & Francis Group, pp. 442–460. ISBN: 978-0-8058-5753-5 978-0-8058-5752-8.

Lipsey, Mark W and David B Wilson (2001). "Analysis issues and strategies". In: *Practical Meta-Analysis*. Ed. by MW Lipsey and DB Wilson. Thousand Oaks, CA: SAGE Publications, Inc, pp. 105–128.

Loosveldt, Geert, Koen Beullens, Caroline Vandenplas, Hideko Matsuo, Lizzy Winstone, Ana Villar, and Verena Halbherr (2014). *ESS interviewer briefing: Note for national coordinators*. Report.

*Mayer, Thomas S and Eileen O'Brien (2001). "Interviewer refusal aversion training to increase survey participation". In: *Proceedings of the annual meeting of the American Statistical Association*.

Means, Barbara, Yukie Toyama, Robert Murphy, and Marianne Baki (2013). "The effectiveness of online and blended learning: A meta-analysis of the empirical literature". In: *Teachers College Record* 115.3, pp. 1–47.

Mneimneh, Zeina N, Michael R Elliott, Roger Tourangeau, and Steven G Heeringa (2018). "Cultural and interviewer effects on interview privacy: Individualism and national wealth". In: *Cross-Cultural Research*, p. 1069397118759014. ISSN: 1069-3971.

Moher, David, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement". In: *Annals of internal medicine* 151.4, pp. 264–269. ISSN: 0003-4819.

OECD (2014). *PIAAC Technical standards and guidelines*.

Olson, Kristen, Chun Feng, and Lindsey Witt (2008). "When do nonresponse follow-ups improve or reduce data quality? A meta-analysis and review of the existing literature". In: *International Workshop on Total Survey Error. Research Triangle Park, NC. http://www. niss. org/sites/default/files/OlsonTSEWorkshopNRMEReview080108. pdf (accessed 9/24/10)*.

Olson, Kristen, Antje Kirchner, and Jolene Smyth (2016). "Do interviewers with high cooperation rates behave differently? Interviewer cooperation rates and interview behaviors". In: *Survey Practice* 9.2.

O'Muircheartaigh, Colm and Pamela Campanelli (1998). "The relative impact of interviewer effects and sample design effects on survey precision". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 161.1, pp. 63–77. ISSN: 1467-985X.

*Schnell, Rainer and Mark Trappman (2006). *The effect of the refusal avoidance training experiment on final disposition codes in the German ESS-2*. Report. Working Paper 3/2006, Germany: Center for Quantitative Methods and Survey Research, University of Konstanz.

Singer, Eleanor, Martin Frankel, and Marc B. Glassman (1983). "The effect of interviewer characteristics and expactations on response". In: *Public Opinion Quarterly* 47.1, pp. 68–83.

Survey Research Center, Michigan (2016). *Guidelines for best practice in cross-cultural Surveys*. Report. Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

Thorndike, Edward Lee (1913). *The psychology of learning*. Vol. 2. New York, NY: Teachers College, Columbia University.

Tusting, Karin and David Barton (2003). *Models of adult learning: A literature review*. NIACE. ISBN: 1862012806.

Viechtbauer, Wolfgang (2010). "Conducting meta-analyses in R with the metafor package". In: *Journal of Statistical Software* 36.3, pp. 1–48.

Weiss, Bernd and Michael Wagner (2011). "The Identification and prevention of publication bias in the social sciences and economics". In: *Journal of Economics and Statistics* 231.5/6, pp. 661–684. ISSN: 00214027. URL: http://www.jstor.org/stable/23813343.

West, Brady T, Frederick G Conrad, Frauke Kreuter, and Felicitas Mittereder (2018). "Can conversational interviewing improve survey response quality without increasing interviewer effects?" In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181.1, pp. 181–203. ISSN: 1467-985X.

West, Brady T, Frauke Kreuter, and Ursula Jaenichen (2013). ""Interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse?" In: *Journal of Official Statistics* 29.2, pp. 277–297. ISSN: 2001-7367.

West, Brady T and Kristen Olson (2010). "How much of interviewer variance is really nonresponse error variance?" In: *Public Opinion Quarterly* 74.5, pp. 1004–1026. ISSN: 0033-362X.

West, Brady and Annelies G Blom (2017). "Explaining interviewer effects: A research synthesis". In: *Journal of Survey Statistics and Methodology* 5.2, pp. 175–211. DOI: 10.1093/jssam/smw024.

Chapter 6

Conclusion and Discussion

This dissertation had two objectives, first, to derive recommendations for survey implementation from evidence-based practice with the use of meta-analyses and randomized controlled trials, and second to demonstrate the applicability of meta-analyses in survey methodology research. In the course of this dissertation I demonstrated how randomized controlled trials and meta-analyses based on randomized controlled trials could be performed in the field of survey methodology and how beneficial implications for conducting surveys can be inferred from evidence-based research. Since the conclusions of the research were already drawn in the individual chapters, I will briefly summarize the findings and discuss the four studies with regard to their role for evidence-based methods in survey methodology.

My first study (chapter 2) was a randomized controlled trial in the field of device effects in web surveys. Specifically, I focused on the usage of filter and follow-up questions. I randomly assigned the respondents to one of two question formats - *interleafed* or *grouped* format - as well as to PC or smartphone device. I showed that mobile respondents do not trigger fewer filter questions than PC respondents. However, I found that mobile respondents provide lower data quality in terms of more item-nonresponse, heaping, and middle category ticks in the follow-ups. Although this study was designed as a randomized controlled trial, some methodological difficulties emerged during its implementation. First of all, the fieldwork institute could not meet the quotas for the assignment to smartphone mode for some of the socio-demographic variables

since many more subjects have refused to participate via smartphone in the first place and those differed significantly from the participants. As a result, I observe significant differences between smartphone and PC respondents. This means that the objective of randomization was not met. The threat to external validity of this study exists not only because of the nature of the sample, but also because respondents were "forced" into the smartphone mode. They may have been less motivated to comply in the follow-up questions, thus the results should be read with caution. It can be concluded from this study that even randomized controlled trials do not necessarily lead to causal conclusions and that an optimal study design cannot always be implemented accordingly. Admittedly the chosen experimental design is to be preferred to an observational design, since this would be affected by the self-selection of the respondents into the respective devices and any differences in data quality would not be clearly attributable to the device only but also to the self-selection.

The second study (chapter 3) of this dissertation was a meta-analysis of randomized controlled trials. The focus of this study was on web response rates in comparison to ther surveys' response rates. In this study I replicated and extended a previous meta-analysis (Lozar Manfreda et al. 2008) and found that web surveys still have on average 12 percentage points lower response rates than other modes. Furthermore, I was able to show that a number of survey characteristics (prenotifications, the sample recruitment strategy, the survey's solicitation mode, the type of target population, the number of contact attempts, and the country in which the survey was conducted) moderated the level of response rate difference. In the primary studies included in this meta-analysis, the respondents were again randomly assigned to a mode, similar to the procedure described above in the first study (chapter 2), resulting in similar challenges with external validity as explained in the previous section. While the mode of assignment was randomized in the primary studies, this did not apply to the survey characteristics such as incentive usage of the primary studies that moderated the response rate difference. Therefore, no causal conclusions can be drawn for the survey characteristics, because the absence of an experimental design of the moderators could lead to significant effects of (unobserved) third variables or pseudo correlations. A remedy would be an experimental variation of the moderators, which is, however, difficult to implement in practice and requires large samples.

The third study of this dissertation (chapter 4) used almost the same dataset as chapter 3 focusing on which countries obtain high response rates in web surveys and the factors (social, economic and technological factors as well as the survey participation propensity) that determine high web response rates. I found that web surveys achieve high response rates in countries with a high population growth, high internet coverage, and a high survey participation propensity, whereas they are at a disadvantage in countries with a high population age and mobile phone coverage. Due to the small number of countries and the strong presence of US studies, no multilevel meta-analysis could be applied to this question, which would have disentangled the variance on the country level. The advantage of a multilevel approach would have been the avoidance of pseudo correlations, since with my approach interaction effects between study designs and countries can be present. For instance, a particular country uses repeatedly a particular solicitation mode and the effect found is not due to the country but to the solicitation mode. Furthermore, there are certainly other factors that influence web response behavior at the country level, such as data security attitudes, but the databases that provide such macro variables over years and countries are missing.

The fourth study of this dissertation (chapter 5) was a meta-analysis that tested the effectiveness of interviewer training and aimed to find which training methods are effective. This study had to deal with the heterogeneity problem of the primary studies. Although, the effectiveness of interviewer training on data quality was measured in all primary studies included in this meta-analysis within the framework of randomized controlled trials, various concepts of data quality were used as a basis, such as item nonrepsonse, unit nonresponse and correctly administered items. This resulted in six different subsets of meta-analyses with only a limited number of studies. Conceptually, the comparison of different data quality measurements would not have made sense, but the chosen approach questioned the robustness of the findings and the moderators had little variation and allowed hardly any conclusions. Thus, this meta-analysis is the first to address the effects of interviewer training but its realization might have been a a bit premature. Therefore, a replication after more primary studies are conducted is recommended as an extension of this work.

This last study illustrates very clearly a basic problem of meta-analyses in survey methodology

- the incomparability of effect sizes. The measurement concepts for data quality are often heterogeneous, so that even a monitoring of these underlying concepts via a moderator analysis will not account for the incomparability of concepts.

Overall, this dissertation provides a few important take home messages. This research has shown that the usage of experiments often leads to challenges in practice. On the one hand, the experimental assignment, to for instance a certain mode or device, as shown in chapter 2, 3 and 4, is not always possible, since the respondents have the opportunity to refuse participation at any time, and on the other hand only one or a few factors can be varied experimentally at a given time for reasons of capacity.

Furthermore, when coding some of the studies in meta-analyses, I had to make assumptions about unreported facts (such as the final number of invited subjects in chapter 3 and 4). At this point the field of survey methodology lacks uniform reporting methods and standards. A possible long-term solution would be to provide standardized study information on reporting and to disclose the analysis strategy via log files (open methodology) enacted by, for instance, the journals.

Another challenge I encountered in my work on the application of evidence-based methods is the publication bias problem in chapter 5. The field of the survey methodology can learn from other disciplines, such as pre-registration in psychology in terms of to prevent that significant results are more likely to be published. In some psychological journals and databases it is possible to pre-register journal articles and research projects prior to their implementation and thus commit to publishing the results regardless of whether they are significant or consistent with existing theories.

Moreover, this dissertation has shown that survey research, like other disciplines, finds heterogeneous insights for the same research questions across primary studies. In this context, systematic accumulation within the framework of meta-analyses can provide added value. However, meta-analyses consist of many individual steps from the definition of eligibility criteria, literature search, screening, coding, to analysis. In particular, the coding of results is the most time-consuming and complex step. However, this step could have been carried out by the au-

thors of primary studies themselves by collecting and registering their results in a standardized form in databases. This would result in large databases with results from primary studies and would considerably facilitate the systematization of findings and conclusions. Other disciplines, such as medicine (e.g. *clinical evidence database: https:bestpractice.bmj.com/info/evidence-information/*; physio therapy (e.g. *physiotherapy evidence database: https://www.pedro.org.au*); and educational research (e.g. *professional learning and student achievement database: https://learningforward.org/publications/evidence-database*), are pioneers in this regard.

Finally, the role of replications in evidence-based methods should be addressed. To start with, the replication and update of systematic reviews and meta-analyses as shown in chapter 3 is essential, as the accumulation of new and old findings can help identify new research directions and challenges. Accumulation also allows statements on whether additional primary studies are needed or whether the effects are time constant (Bosnjak 2018; Borenstein et al. 2009). Moreover, primary studies should also be replicated to verify their consistency and to accumulate results. Auspurg and Brüderl (2019) suggest that as a first step, studies could replicate an effect already found by another study, while in a second step, they could add a previously unexplored finding.

To finish, this dissertation was able to draw conclusions from four studies on evidence by randomized controlled trials and meta-analyses and present the resulting challenges. In the training of survey methodologists, the teaching of evidence-based methods should always be enhanced with teachings on observational studies, since there are instances in which randomization is impossible in the field of survey methodology. The goal is always to chose the most feasible research design that is less prone to error for each research question so I would like to conclude with the quote from Sackett and Wennberg (1997, p.1636) “Each method should flourish, because each has features that overcome the limitations of the others when confronted with questions they cannot reliably answer”.

References

Auspurg, Katrin and Josef Brüderl (2019). *Is there a credibility crisis in sociology? And if yes, what can be done?* MZES Open Social Science Conference 2019. Conference Presentation. URL: https://www.mzes.uni-mannheim.de/openscience/wp-content/uploads/2019/01/Auspurg_Br%5C%C3%5C%BCderl-Credible-Sociology-Mannheim.pptx (visited on 03/26/2019).

Borenstein, Michael, Larry V Hedges, Julian Higgins, and Hannah R Rothstein (2009). *Introduction to meta-analysis*. Hoboken, CA: Wiley Online Library. ISBN: 0470743387.

Bosnjak, Michael (2018). "Evidence-based survey operations: Choosing and mixing modes". In: *The Palgrave Handbook of Survey Research*. Springer, pp. 319–330.

Lozar Manfreda, Katja, Michael Bosnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar (2008). "Web surveys versus other survey modes: A meta-analysis comparing response rates". In: *Journal of the Market Research Society* 50.1, p. 79. ISSN: 0025-3618.

Sackett, D. L. and J. E. Wennberg (1997). *Choosing the best research design for each question*. Journal Article. URL: https://www.ncbi.nlm.nih.gov/pubmed/9448521%20https://www.ncbi.nlm.nih.gov/pmc/PMC2128012/.